Title
Evolution of the energy-coupling factor (ECF) transporters and comparative riboswitch analysis

Author
Sun, Eric I-Chung

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Evolution of the Energy-Coupling Factor (ECF) Transporters and Comparative Riboswitch Analysis

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Biology

in

Biology

by

Eric I-Chung Sun

Committee in charge:

Professor Milton Saier, Chair
Professor Eric Allen
Professor Randy Hampton
Professor Joseph Pogliano
Professor Wei Wang

2012

The Dissertation of Eric I-Chung Sun is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

TABLE OF CONTENTS

Signature Page iii

Table of Contents iv

List of Figures v

List of Tables viii

Acknowledgements ix

Vita x

Abstract of the Dissertation xii

I. Introduction 1

II. Materials and Methods 11

III. Results 16

IV. Discussion 39

V. Figures and Tables 51

References 108

LIST OF FIGURES

Figure 1A: Binary alignments of S subunit homologues of putative members of the ECF transporter family (ThiW/BioY) 78
Figure 1B: Binary alignments of S subunit homologues of putative members of the ECF transporter family (ThiW/YhaG) 78
Figure 1C: Binary alignments of S subunit homologues of putative members of the ECF transporter family (ThiW/YjbB) 79
Figure 1D: Binary alignments of S subunit homologues of putative members of the ECF transporter family (YhaG/YjbB) 79
Figure 1E: Binary alignments of S subunit homologues of putative members of the ECF transporter family (BioY/YhaG) 80
Figure 1F: Binary alignments of S subunit homologues of putative members of the ECF transporter family (BioY/YjbB) 80
Figure 2A: Average hydropathy, amphipathicity and similarity plot (ThiW)..... 81
Figure 2B: Average hydropathy, amphipathicity and similarity plot (YhaG)..... 82
Figure 2C: Average hydropathy, amphipathicity and similarity plot (YjbB) 83
Figure 3A: The AveHAS plot for the S subunits of type-I ECF transporters (BioY) .. 84
Figure 3B: The AveHAS plot for the S subunits of type-I ECF transporters (CbiM).. 85
Figure 3C: The AveHAS plot for the S subunits of type-I ECF transporters (YkoE).. 86
Figure 4A: Binary alignments of suspected intragenic duplications in S subunits (BioY homologues) 87
Figure 4B: Binary alignments of suspected intragenic duplications in S subunits (CbiM homologues)..... 87
Figure 4C: Binary alignments of suspected intragenic duplications in S subunits (YkoE homologues)..... 88
Figure 5: Binary alignment of the N-terminal and C-terminal halves of YjbB homologues..... 88
Figure 6A: Binary alignments of S and T subunits (CbiM/CbiO) 89
Figure 6B: Binary alignments of S and T subunits (NikM/NikO) 89

Figure 7A: AveHAS plots for the S (left) and T (right) subunits (CbiM/CbiQ) 90
Figure 7B: AveHAS plots for the S (left) and T (right) subunits (NikM/NikQ) 91
Figure 8: Venn diagram of the organismal distributions of gene clusters of the BioY and ThiW transporter components 92
Figure 9: Phylogenetic trees for the three components of type-I ECF transporter homologues 93
Figure 10A: Phylogenetic trees for the A subunit homolgoues of type-II ECF transporters 94
Figure 10B: Phylogenetic trees for the T subunit homolgoues of type-II ECF transporters 95
Figure 11: Superfamily trees for the S subunit homologues belonging to type-I/II ECF transporters as well as the membrane constituents of representative ABC2 members 96
Figure 12: Growth analysis of thiamine synthesis/transport-null strains of E. coli expressing various transporter constituents of ThiW from *Mycobacterium smegmatis* str. MC2 155 97
Figure 13: Proposed pathway of ECF and ABC2 evolution 98
Figure 14: Total riboswitch distribution imposed on the phylogenetic tree of the major microbial phyla 99
Figure 15: Proportion of riboswitches with different metabolic roles within individual phylum 100
Figure 16: Phylogenetic tree of homologous MgtA proteins in Enterobacteriales 101
Figure 17: Phylogenetic tree of homologous MetX proteins in Rhizobiales 102
Figure 18: Phylogenetic tree of homologous SpeF proteins in Rhizobiales 103
Figure 19: Phylogenetic tree of homologous YbhL proteins in Rhizobiales 104
Figure 20: Phylogenetic tree of orthologous MetX proteins regulated by functionally equivalent SAM riboswitches 105
Figure 21: Phylogenetic tree of orthologous MetY proteins regulated by functionally equivalent SAM riboswitches 106

Figure 22: Phylogenetic tree of orthologous YkkC proteins regulated by functionally equivalent riboswitches 107

LIST OF TABLES

Table 1: Fully sequenced microbial genomes used for riboswitch analyses 52
Table 2: Comparison scores for the transmembrane porters derived from four permease families 60
Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues 61
Table 3B: Genomic cluster analysis of BioY and ThiW transporter homologues 67
Table 4: Comparisons between transporters of different specificities in the ECF Family 68
Table 5: Collection of ECF energizing components 69
Table 6: Collection of S subunit homologues from the type-I/II ECF transport and ABC transport systems 71
Table 7: Distribution of the coenzyme riboswitch class in major bacterial phyla 72
Table 8: Distribution of riboswitch classes responsible for amino acid metabolism in major bacterial phyla 73
Table 9: Distribution of riboswitch classes responsible for the biogenesis of ribosomal subunits in major bacterial phyla 73
Table 10: Distribution of nucleotide derivative riboswitch classes in major bacterial phyla 74
Table 11: Distribution of ion/sugar riboswitches in major bacterial phyla 74
Table 12: Distribution of functionally uncharacterized riboswitches in major bacterial phyla 75
Table 13: Combined riboswitch distribution across fully sequenced microbial genomes 76
Table 14: Proportion of riboswitches for individual effector classes in each bacterial phylum 77

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Milton Saier for his belief that his students can always achieve what they believe in. His encouragement and advice was critical for the formulation of my dissertation, but mostly I would like to thank him for fostering a spirit of independent research that contributed to the pursuit of my research interests, not to be confined by the limit of my own knowledge.

I would also want to thank Vamsee Reddy, Ming Ren Yen and Dorjee G. Tamang for all their help in providing technical help on running our analysis programs in the Saier lab. I am especially grateful for their helpful advice and assistance.

A special thanks goes to Dr. Dmitry Rodionov in the lab of Andrei Osterman from the Sanford-Burnham Medical Research Institue. His mentorship helped me immensely with the analysis of riboswitch sequences and comparative genomics.

Chapter III, in full, is in press in the International Journal of Bioinformatics, 2012. Sun, Eric; Saier, Milton. The dissertation author was the primary investigator and author of this paper.

Chapter III, in part, is currently being prepared for publication. Sun, Eric; Leyn, Semen; Kazanov, Marat; Novichkov, Pavel; Saier, Milton; Rodionov, Dmitry. The dissertation author was the primary investigator and author of this material.

VITA

2005	Bachelor of Arts, University of California, Berkeley
2006-2007	Research assistant, University of California, Irvine
2012	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Sun, E.I., Leyn, S.A., Kazanov, M.D., Novichkov, P.S., Saier, M.H. Jr., Rodionov DA (2013). Functions and evolution of regulons controlled by RNA regulatory elements in complete bacterial genomes. Manuscript in preparation.

Sun, E.I., Saier, M.H. Jr. Evolution of the Energy-Coupling Factor (ECF) Sub-superfamily of the ABC Supferfamily. *Int. J. of Bioinformatics*, in press.

Yee, D.C., Shlykov, M.A., Aurora, S., Chen, J.S., Reddy, V.S., Sun, E.I., Saier, M.H. Jr. (2012) The Microbial Rhodopsin (MR) superfamily. *Nature Communications*, submitted.

Reddy, V.S., Shlykov, M.A., Castillo, R., Sun, E.I., Saier, M.H. Jr. (2012) The major facilitator superfamily (MFS) revisited. FEBS J. 279:2022-35.

Lam, V.H., Lee, J.H., Silverio, A., Chan, H., Gomolplitinant, K.M., Povolotsky, T.L., Orlova, E., Sun, E.I., Welliver, C.H., Saier, M.H. Jr. (2011) Pathways of transport protein evolution: recent advances. *Biol. Chem*. 392:5-12.

Saier, M.H. Jr., Wang, B., Sun, E.I., Matias, M., Yen M.R. (2010) Molecular archeological studies of transmembrane transport systems. *Structural Bioinformatics of Membrane Proteins*. p29-42. Ed. D. Frishman. ISBN: 978-3-7091-0044-8.

Yen, M.R., Chen, J.S., Marquez, J.L., Sun, E.I., Saier, M.H.. (2010) Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. *Methods Mol. Biol*. 637:47-64.

Wang, B., Dukarevich, M., Sun, E.I., Yen, M.R., Saier, M.H. Jr. (2009) Membrane Porters of ATP-Binding Cassette Transport Systems Are Polyphyletic. *J. Membr. Biol*. 231:1-10.

Li, C., Li, H., Zhou, S., Sun, E., Yoshizawa, J., Poulos, T.L., Gershon, P.D. (2009) Polymerase translocation with respect to single-stranded nucleic acid: looping or wrapping of primer around a poly(A) polymerase. *Structure*. 17:680-9.

ABSTRACT OF THE DISSERTATION

Evolution of the Energy-Coupling Factor (ECF) Transporters and Comparative Riboswitch Analysis

by

Eric I-Chung Sun

Doctor of Philosophy in Biology

University of California, San Diego, 2012

Professor Milton Saier, Chair

Energy-coupling factor (ECF) porters catalyze uptake of vitamins and trace minerals with high affinity. They consist of two membrane constituents called S (substrate recognition) and T (energy transducing) subunits as well as one or two energizing ATPase(s), the A subunit(s). The S subunit is thought to recognize the substrate while the T subunit interacts with the ATPase. We here show that each of these three subunits is monophyletic, that the S and T subunits are homologous but distantly related to each other, and that these subunits are homologous to the integral membrane subunits of conventional ATP-binding cassette (ABC) porters. ECF porters therefore comprise an offshoot (a "sub-superfamily") of the ABC superfamily in spite of some distinctive features. We propose a pathway by which all of these transport systems may have evolved. An intragenic duplication of a genetic element encoding a

3 transmembrane segment (TMS) peptide gave rise to 6 TMS proteins, and these sometimes lost the C-terminal TMS to give 5 TMS proteins. The transmembrane subunits of the ECF porters are also homologous to certain secondary carriers, and we provide preliminary evidence that the S subunit of a thiamine porter can function by itself (secondary active transport) or when complexed with both T and A subunits (primary active transport). Phylogenetic analyses of the three ECF subunits revealed that extensive shuffling of these constituents occurred over evolutionary time although the T and A subunits, when encoded separately from the S subunits, frequently coevolved. Additionally, our genomic analyses using riboswitch regulatory sequences regulating expression of ECF transporter genes promises to reveal potential substrates of these diverse transporters.

In the second part of this dissertation, I analyzed the distribution of riboswitches and used the results for comparative genome analysis. Riboswitch analysis offers a unique advantage in that the substrate binding (aptamer) domain of a riboswitch are highly conserved, allowing more accurate functional predictions. With the rapid accumulation of complete prokaryotic genomes and experimental validation of riboswitch sequences, such phylogenetic/functional analyses allow more extensive annotation of previously uncharacterized genes and comparison of metabolic pathways between different organisms. Our study is focused on the discovery of candidate riboswitches in fully sequenced bacterial genomes and regulon reconstruction of riboswitch-regulated genes, which are collected and manually curated using a RegPredict web-based tool. From these annotations we can thoroughly analyze the conservation of orthologous metabolic pathway and the distribution of different types of riboswitches and establish their probable evolutionary pathways in the Domain Bacteria. My results indicated that some riboswitches, especially those that correspond to coenzymes TPP and cobalamin, are widespread and possibly originated

very early. However, many more riboswitches in our analyses are restricted to just a single Order/Family of bacteria and likely represent more recent evolutionary innovations.

I.

Introduction

For the past several years, our laboratory has designed methods and software to identify increasingly distant phylogenetic relationships between membrane transport proteins (Chang et al., 2004; Lee et al., 2007; Mansour et al., 2007; Tamang & Saier, 2006; Debut et al., 2006; Prakash et al., 2003; Hvorup et al., 2003; Zhai & Saier, 2002; Reddy et al., 2012). Our studies and those of many other laboratories have revealed that common ancestry often implies common structure, mechanism and function, with degrees of structural and functional divergence correlating with degrees of sequence divergence. Out of more than eighty thousand structures reported in the Protein Data Bank (PDB, see http://www.pdb.org/), membrane protein structures comprise less than three percent. Given the difficulty of membrane protein crystallization, the identification of superfamily relationships is extremely important for structural and functional prediction and the establishment of the evolutionary pathways taken for the appearance of transmembrane proteins.

Energy-coupling Factor (ECF) transporters are particularly well suited for phylogenetic and structural analyses given their high degrees of sequence divergence. They were originally identified as a group of novel transmembrane substrate binding proteins by Henderson et al. (1979). These transporters had been included in the ABC superfamily in the Transporter Classification Database (TCDB, www.tcdb.org) due to apparent sequence similarity among these systems, but quantitative sequence analyses have not been conducted. Recent genomic analyses have uncovered additional ECF transporters in multiple archaeal and bacterial genomes (Rodionov et al., 2006, 2009; Hebbeln et al., 2007), and these studies greatly expand our knowledge of their distribution, functions, and potential modes of energy coupling.

Like ECF systems, ABC uptake systems exhibit two transmembrane components and two cytoplasmic ATPases. The ATPases of ECF transporters are clearly homologous to those of all ABC systems, reflecting potential phylogenetic relationships between these two groups of transporters. However, members of the ECF porters are distinguishable by the absence of soluble periplasmic binding proteins and by the presence of two highly divergent transmembrane subunits that may serve distinct roles in transport (Hebbeln et al., 2007; Neubauer et al., 2009; Rodionov et al., 2006, 2009; Eitinger et al., 2011). It has been noted that the sequence similarity among S subunits of the ECF systems is fairly low when compared to the moderately well conserved T subunits, but little is known about the origin of the S subunits and their evolutionary histories. Within "traditional" ABC uptake porters such as the maltose uptake porter of *E. coli*, both subunits are believed to have overlapping but quantitatively divergent functions (Oldham et al., 2007), and the ECF systems could be evolutionarily related to the ABC systems.

ECF transporters may be capable of functioning independently of ATP hydrolysis. Many organisms possess only the S subunits (especially well documented for biotin transporters) but not the other ECF components (Rodionov et al., 2009). *In vivo* expression systems for some of these ECF homologues indicate possible secondary active transport mechanisms when the S subunit is expressed alone, although the physiological relevance is unclear (Rodionov et al., 2006; Hebbeln et al., 2007; Finkenwirth et al., 2010; see also this study regarding thiamine uptake by a ThiW homologue from *Mycobacterium smegmatis str. MC2 155*). With the existence of other transport systems that potentially possess a dual mode of energy coupling *in*

vivo (Kuroda et al., 1997; Hvorup et al., 2003), the evolution of transporters, especially with regards to the development of energy coupling mechanisms, appears to be more circuitous than was originally thought.

In this dissertation, I identify four previously recognized families of primary and secondary active transport systems in which the permease constituents prove to be members of different families: BioY (TC#3.A.1.25.1; see the Transporter Classification Database [TCDB]; www.tcdb.org), ThiW (TC#3.A.1.26.1), YhaG (TC#2.A.88.4.1), and YjbB (TC#2.A.58.2.1). These systems are responsible for the uptake of biotin, thiamine, tryptophan, and phosphate, respectively, with the latter two transporters functionally assigned as secondary active transporters (Sarsero et al., 2000; Kohler et al., 2001; Lebens et al., 2002; Murer et al., 2000; Miyamoto et al., 2007; Tenenhouse, 2005). Clues concerning the evolutionary relationships between members of the divergent ECF family can be gleaned through the use of topological and sequence analyses. In addition, my study addresses the similarities and differences between the S and the T subunits, their functional distinctions, and their evolutionary relationships with ABC porters.

Based on earlier genome context analyses (Rodionov et al., 2009), I elaborate on the consequences of genomic arrangement of the ECF family on its evolution. Specifically, several ECF transporters within the same organism may share a common energizing AAT module (ter Beek et al., 2011), a property not uncommon among members of the ABC superfamily. Such systems have been categorized as type-II ECF porters, whereas others that have dedicated energizing modules (as indicated by genome context analyses) were grouped into type-I ECF porters (Rodionov et al.,

2009). In addition to establishing the evolutionary pathways of ECF porters, such genomic organization may offer clues regarding transporter regulation on the transcriptional level and potential for secondary energization in isolated S subunits.

As ECF transporters participate in the uptake of many metabolically important micronutrients and are indispensable for the growth of many pathogenic microorganisms that lack corresponding *de novo* synthetic pathways, the study of ECF transporters may allow additional strategies for the treatment of diseases associated with these organisms. In the final part of the ECF transporter analyses, I propose to search for novel ECF transporters and their potential substrates using operonic riboswitch elements as a guide.

As an extension of my work with ECF transport systems, I analyzed riboswitch sequence across fully sequenced microbial genomes. Riboswitches are genetic elements found mostly in bacteria and frequently reside at the 5' UTR of mRNAs. Riboswitches achieve gene regulation by possessing two alternative structural states and are likened to an "on/off" switch (Nudler & Mironov, 2004). Direct binding of substrate to the aptamer domain and altering of secondary/tertiary structures in the expression platform (which overlaps with antiterminators and/or anti-Shine-Dalgarno sequences) confer gene regulation, usually by ways of transcription termination, inhibition of translation initiation, and/or attenuation of mRNA stability (Nudler & Mironov, 2004; Breaker, 2012). Additionally, there are also riboswitches that behave like ribozymes and self-cleave when a specific conformation is achieved, and many eukaryotic riboswitches that situate in the intron regions have been found to influence splicing patterns of mRNAs (Bastet et al., 2011; Breaker, 2012). As there is usually

no requirement for a protein factor to mediate substrate recognition, it has been proposed that riboswitch elements represent an ancestral gene regulatory mechanism that had arisen prior to the rise of protein/peptides (Vitreschak et al., 2004; Garst et al., 2011; Breaker, 2012). Their proposed ancient origin is also evident in that many of the substrates they recognize have central roles in cellular metabolism across different organisms (Nudler & Mironov, 2004; Breaker, 2012). Their proposed ancient origin may have contributed to the wide distribution of functionally similar riboswitches that recognize the same ligand. As an example, there have been seven riboswitch classes with four distinct binding pockets that are specific for the S-Adenosyl methionine (SAM) coenzyme although it is not clear whether members share common ancestries (Breaker, 2012). This variation demonstrates great flexibility inherent in an RNA molecule for recognition of small ligands through the use of only four nucleotides. However, this inherent flexibility does make the analysis of riboswitch difficult, especially considering the small sizes of many of these cis-regulatory elements. Nonetheless, efforts have been made both in bioinformatic and *in vitro* analyses to confirm the diversity and distribution of riboswitch sequences.

As riboswitches specific to a particular ligand are highly conserved, many riboswitch-regulated genes with previously unknown functions or ligand specificities can be reliably annotated from environmental sampling. Previous environmental metagenomic efforts from three diverse environmental sources (Sargasso Sea, Minnesota soil, and whale falls) revealed a high abundance of TPP, cobalamin, and glycine riboswitches, followed by SAM, FMN, ykkC-yxkD, and yybP-ykoY riboswitches. On the opposite end of the spectrum, lysine, purine, glmS, and ykoK

riboswitches were rare in their taxonomic distribution (Kazanov et al., 2007). The study had established relative abundances of riboswitches in alpha/beta/gamma-proteobacteria, Firmicutes, and Bacteroidetes/Chlorobi (Kazanov et al., 2007). Although environmental sequencing can give us clues about the relative abundance of known riboswitches as well as discovering novel putative riboswitches, it is only through careful integration of complete genomes with riboswitch analyses that a meaningful picture of metabolic networks and their probable evolutionary pathways will emerge. Additionally, environmental metagenomics has inherent bias in estimating the relative abundances of riboswitches across different samples due to bias in the types of organisms that are typically found in a certain locale; this is demonstrated by the abundance of glycine riboswitches detected from the Sargasso Sea due to the abundance of *Candidatus Pelagibacter ubique* (Giovannoni et al., 1990; Rappé & Giovannoni, 2003; Kazanov et al., 2007).

Many of these widespread riboswitches have also been analyzed in fully sequenced genomes as a part of the overall regulatory network analysis in previous publications (Rodionov et al., 2002a; Rodionov et al., 2003a/b), and they provide a better regulatory and evolutionary profile for individual riboswitch classes, which are usually difficult to accomplish through purely metagenomic efforts. For instance, for TPP riboswitches gene encoding a biosynthetic enzyme (ThiC) that is part of the hydroxymethylpyrimidine (HMP) pathway is closely associated with the riboswitch in bacteria as are many genes in the parallel hydroxyethylthiazole (HET) pathway (for instance ThiOSG); the pervasiveness of such riboswitch-controlled biosynthetic regulons highly suggests its origin with TPP riboswitches, which can then duplicate

and transfer to various transporter genes (Rodionov et al., 2002a). Cobalamin riboswitches primarily regulate enzymes responsible for the synthesis of the corrin ring from uroporphyrinogen-III as well as its subsequent modification into adenosylcobalamin. Both aerobic and anaerobic pathway enzymes are regulated by the riboswitch. In addition, a large number of known or predicted cobalt transporters, which are widespread in both bacteria and archaea as well as B12-dependent enzymes, were found to be regulated by cobalamin riboswitches (Rodionov et al., 2003a). FMN riboswitches are closely associated with *ribDE(B/A)H* operons in Gram-positive bacteria and either *ribB* or *ribH2* genes in other bacterial taxa (especially in proteobacteria) and may have their origins there. The same study also pointed to differences in the expression platform between two riboswitch orthologues that can lead to two distinct modes of regulation (Vitreschak et al., 2002). For lysine riboswitches, both the biosynthetic genes (*lysC*, *lysA*, and *dapA*) and the putative/known transporters (*yvsH*, *lysW*, *lysP* and *lysXY*) genes are frequently regulated by the riboswitch in firmicutes whereas *lysC* and *lysW* are found in other bacterial groups, particularly in gamma-proteobacteria (Rodionov et al., 2003b). The assignment of gene function through positional cluster and regulatory sequence analyses is especially applicable to T-box regulatory systems, which are found primarily in various Gram-positive bacteria, as their substrate specificities are determined solely from their anti-anticodons; the pervasiveness of the T-boxes in Gram-positive organisms suggests their evolutionary origin in those taxa (Vitreschak et al., 2008). The comprehensive analysis of riboswitch distribution and regulon

analysis might offer a glimpse into the putative ancestral RNA world and the evolution of gene regulatory mechanisms.

In this section of the dissertation, I propose to incorporate operon analysis coupled with riboswitch detection to elucidate the phylogeny of many annotated riboswitches. I combine inference of riboswitch sequences using positional weight matrices constructed from the Rfam database and operon analyses in completely sequenced bacterial genome to carry out evolutionary prediction of regulatory networks for different riboswitch functional classes. We separate diverse types of riboswitches roughly into six groups based on the types of ligands they bind and the cellular processes they regulate.

In group one, I include all the riboswitches that recognize coenzymes. This group of riboswitches is the most widely distributed and includes TPP, cobalamin, SAM, SAM_alpha, SAH_riboswitch, SAM-IV, SAM-SAH, SMK box, SAM-Chlorobi, FMN, MOCO_RNA_motif glycine, and THF riboswitches (SAM_alpha, SAH_riboswitch, SAM-IV, SAM-SAH, SMK box, SAM-Chlorobi likely have varying degree of affinity for SAM coenzyme). In group two, I include riboswitches that recognize amino acids or uncharged tRNAs, including all T-boxes specific for 19 amino acids. Riboswitches that bind directly to amino acids include lysine, glycine, and *glnA* riboswitches. In addition, I also include regulatory sequences (attenuators) that kinetically regulate expression by directly encoding regulatory sequences enriched in the codons for the corresponding amino acids they regulate; these are His-, Leu-, Thr-, and Trp-leaders.

In group three, I include riboswitches that respond to ribosomal subunits (S15, L10_leader, L13_leader, L19_leader, L20_leader, and L21_leader). In group four, I include riboswitches that bind to nucleotide derivatives; these include *pyrR*, purine, GEMM_RNA_motif, preQ1 and preQ1-II. In the case of *pyrR*, the regulatory sequence binds to a PyrR regulatory protein; however, this riboswitch sequence is highly structured and contains part of an anti-terminator sequence. As a result, the *pyrR* regulator likely carries out regulation via a mechanism similar to that of a traditional riboswitch. In group five, we included riboswitches that recognize ions (Mg_sensor and ykoK) and sugar (glmS).

In the final group, I include riboswitches with unknown effectors or metabolic roles; these riboswitches include *ykkC-yxkD*, mini-*ykkC*, *ydaO-yuaA*, *serC*, *speF*, *sucA*, *ybhL*, *ylbH* and *yybP-ykoY*. Some of these putative riboswitches tend to be small in sizes, and many were originally discovered through bioinformatic analyses. I hope to shed light on these putative riboswitches and the genes they regulate in order to provide guidelines for future experiments.

The integration of riboswitch prediction with comparative genome analysis will provide a complementary approach to existing methods that relies on the analysis of riboswitch sequences in partial genomes.

II.

Materials and Methods

Bioinformatic procedure for establishing distant protein homolgoues

To search for homologues of proteins in question, sequences (TC#3.A.1.25.1 for BioY; TC#3.A.1.26.1 for ThiW; TC#2.A.88.4.1 for YhaG, and TC#2.A.58.2.1 for YjbB) were used as query sequences in Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) searches with one or two iterations and a cutoff value of e^{-4} (Altschul et al., 1997). To facilitate analyses, all sequences with greater than 80% identity were eliminated using a modified CD-HIT program (Li & Godzik, 2006; Yen et al., 2009). The remaining sequences were aligned using the ClustalX program with default parameters (Thompson et al., 1997). The sequences in the resultant multiple alignments were compared to the BLAST hits obtained when a member of another family was used as the query sequence using the IC program, which looks for homologues with similar sequences obtained with both BLAST searches (Zhai & Saier, 2002). Comparison scores for the best binary alignments were then further analyzed using the GAP program (Devereux et al., 1984), which reveals the precise location of the alignment. Comparison scores of ≥10 SD, corresponding to a probability of 10^{-24} that the observed degree of similarity arose by chance in a continuous alignment of 60 or more amino acyl residues (Dayhoff et al., 1983), is considered to be sufficient to establish homology for the two proteins (Saier, 1994; Yen et al., 2009).

Several programs used to approximate the locations of transmembrane segments (TMSs) were the HMMTOP (Tusnady & Simon 2001), TMHMM (Krogh et al. 2001), PredictProtein (Rost et al., 2003), SOSUI (Hirokawa et al., 1998), and TMpred (Hofmann & Stoffel, 1993) algorithms. Hydropathy, amphipathicity and

topology predictions for individual proteins were made using a modified WHAT program (Zhai & Saier, 2001a; Yen et al., 2009), while average hydropathy, amphipathicity and similarity plots for a multiply aligned group of homologues were generated using a modified AveHAS program (Zhai & Saier, 2001b; Yen et al., 2009).

The ClustalX and the SuperfamilyTree (SFT) programs were used to independently generate phylogenetic trees. The latter program uses large scale comparisons between transporters of different specificities using tens of thousands of bit scores from BLAST searches to generate phylogenetic trees (Zhai et al., 2002; Yen et al., 2009, 2010; Chen et al., 2011). The MEME suite motif analysis program was used to detect subunit-specific motifs (Bailey & Elkan, 1994).

Strain construction

A thiamine synthesis/transport-null strain of *E. coli (*Δ*thiH/*Δ*thiBPQ*, herein referred to as the double knock-out [DK] strain) was constructed from BW25113 (*lacIq rrnBT14* Δ*lacZWJ16 hsdR514* Δ$araBAD_{AH33}$ Δ$rhaBAD_{LD78}$) using PCR recombination as described by Datsenko and Wanner (2000). As a positive control for thiamine transport, another strain unable to synthesize thiamine *(*Δ*thiH*, herein referred to as the single knock-out [SK] strain) was constructed from BW25113A. A ThiW homologue (gi#118469060, which encodes 5'-SAA'-3' of a thiamine transport complex and gi#118472300, which encodes the T subunit of this complex) from *Mycobacterium smegmatis* str. MC2 155 were inserted into the pZE12 vector in various combinations (SAA'T denotes insertion of both ORFs; S denotes insertion of shortened gi#118469060 with both ATPase domains removed, and SAA' or T alone

indicates expression of either ORF in isolation) using EcoRI and XbaI restriction sites before being transformed into the thiamine synthesis/transport-null strain using the standard heat-shock transformation protocol. For a negative control, an empty pZE12 vector was transformed into the same strain. The mutants were propagated in LB medium and were stored in 40% glycerol at -80□C prior to growth and uptake analyses.

***In vivo* uptake analyses**

20ul frozen stock of transformants was inoculated into 5ml M9 minimal medium with 0.2% glucose + 100 μg/ml ampicillin and grown for 14-16 hours at 37□C in a shaker at 250 rpm to eliminate residual thiamine. Before the experiment, the OD_{600} values of the starter cultures were measured, and equal numbers of Colony Forming Units (CFU, with 1 OD_{600} = $5x10^8$ CFU/ml) was inoculated into growth medium. This medium was the same as the minimal medium used for the overnight culture except for the addition of 50 μM IPTG and a filter-sterilized thiamine solution. For each experimental condition, every 5 ml of culture medium was inoculated with $5x10^6$ CFU and incubated at 37□C with shaking. The OD_{600} measurements were then taken at regular intervals.

Discovery of novel ECF transporters and comparative genome analyses using riboswitch elements

A covariance model that combines both probabilistic models of RNA secondary structure and sequence consensus (Eddy & Durbin, 1994) was used for

riboswitch discovery in complete bacterial genomes. The complete list of 255 microbial genomes can be found in table 1, and it includes 68 Firmicutes, 17 Actinobacteria, 113 proteobacteria, 14 cyanobacteria, and 43 miscellaneous phyla. The covariance models of riboswitches were taken from the Rfam database (http://rfam.sanger.ac.uk/) (Griffiths-Jones et al., 2003), and models were constructed for 42 orthologous riboswitches plus 19 T-boxes. The scanning of genomes was done using Infernal (Nawrocki et al., 2009). The obtained candidate regulatory elements were loaded into the RegPredict platform (Novichkov et al., 2010) for subsequent manual curation and analysis (http://regpredict.lbl.gov/regpredict/). Genes obtained from riboswitch regulons were used as queries for BLAST searches against existing entries of ECF transporters in TCDB.

III.

Results

Sequence similarity of the S subunits of four transporter families

An initial BLAST search using the S subunit of the *Rhizobium etli* biotin transporter (BioY; TC#3.A.1.25.1) as the query sequence revealed similarities to the S domains or subunits of putative thiamin transporters (ThiW; TC#3.A.1.26.1). The IC program yielded a comparison score of 11.8 SD in a region of ~65 aas (table 2 and figure 1A), sufficient to establish homology. A similar score was obtained for the comparison between ThiW and the tryptophan transporter (YhaG; TC#2.A.88.4.1) (table 2 and figure 1B). Comparisons of a bacterial member of the phosphate/sodium symporter family (YjbB; TC#2.A.58.2.1) with ThiW and YhaG gave slightly lower but comparable scores (table 2 and figures 1C and 1D). The comparison scores for the remaining alignments showed lower degrees of similarity even after refinement, although they were still highly suggestive of homology (BioY against YhaG and YjbB; see table 2 and figures 1E and 1F). The overall relatedness was also apparent from average hydropathy plots generated for these proteins, in which peaks of hydrophobicity showed similar numbers and spatial distributions (figures 2A, 2B and 3A), with the exception that YjbB appears to contain four or five additional TMSs (figure 2C). As shown here, this is due to an intragenic duplication event that gave rise to two full repeat sequences (see the section entitled “Duplication of the permease domain within the secondary transporter, YjbB”). Our results indicate a high likelihood that all four families with diverse substrates and modes of energization belong to a single superfamily.

Internal duplications within the integral membrane subunits of the ECF sub-superfamily

As shown in our alignment results and AveHAS plot, the 5 TMSs of BioY, the S subunit of the biotin transporter, may have arisen through an intragenic duplication event since TMSs 1 and 2 are similar in sequence to TMSs 4 and 5 (figures 3A and 4A), although the alignment score (8.3 SD) is insufficient to establish homology. However, analysis of other type-I ECF transporters with 5 putative TMSs clearly established this pattern of duplication. A score of 11.0 SD for CbiM (figures 4B and 3B) was obtained when the corresponding TMSs were compared. Analyses of the two halves of the S subunits of other type-I ECF transporters revealed scores that were consistent with homology (figures 3C and 4C). Thus, the S subunits of the ECF transporter arose by intragenic duplication of a 3-TMS precursor with loss of a TMS at the C-terminus to give the prevalent 5 TMS homologues. It should be noted that many of the ECF transporters possessing 5 TMSs as predicted by AveHAS may actually contain six TMSs, with a short segment of the second TMS in the classic α-helical arrangement immediately upstream of the third TMS to form one TMS (Zhang et al., 2010; Erkens et al., 2011).

Using these same methods, intragenic duplications within the T subunits could not be detected. However, since the S and T subunits are homologous (see the section entitled “Homology between the S and T subunits of the ECF sub-superfamily”), it follows that, based on the Superfamily Principle (Doolittle, 1981; Saier, 1994), the two halves of both the S and T subunits must have arisen through intragenic duplication. Thus, both the S and T subunits had a primordial 3-TMS precursor, a

pattern that is shared with the ABC2 export transporters (Wang et al., 2009). This conclusion has been confirmed independently (Zheng et al., manuscript in preparation).

Duplication of the permease domain within the secondary transporter, YjbB

With YjbB being almost twice as large as most S homologues included in our study, we sought to explore the origin of the extra TMSs in the PNaS family. We found that an internal duplication gave rise to the two halves of YjbB; the comparison score obtained between these two halves was 14.6 SD, with putative TMSs 1 to 3 aligning with putative TMSs 5 to 7 (figures 2C and 5). Additional evidence for duplication came from comparison of the 5 and 6 TMS S subunits of ECF transporters with YjbB. Both putative permease domains of YjbB were shown to share significant sequence similarity with S subunits of ECF transporters (figures 1B, 1C and 1E). Thus, the YjbB transporter family and its homologues may function as the equivalent of the dimeric complexes of the permease domains or subunits in ECF porters. These results are in agreement with those obtained by Zheng et al. (manuscript in preparation).

Homology between the S and T subunits of the ECF sub-superfamily

Since only type-I ECF transporters, with dedicated energizing modules, have all three transporter subunits (S, T and A) encoded within single operons, they probably function together (Overbeek et al., 2005; Rodionov et al., 2009). These systems were used for sequence comparisons between the S and T subunits of ECF

homologues. The best comparison score (11.9 SD) corresponded to the alignment for the S (CbiM) and T (CbiQ) subunit homologues of a cobalt transporter (TC#3.A.1.18.1). This value is sufficient to establish homology (figures 6A and 7A). Nickel transporters (NikM and NikQ for the S and T subunits, respectively) gave lower scores (9.9 SD) but still showed significant sequence similarity (figures 6B and 7B). A comparable score (9.4 SD) was obtained for the *R. capsulatus* transporter (BioY [S] and BioN [T]; data not shown). Using the MEME suite motif analysis program (Bailey & Elkan, 1994), the presence of the EAA motif that is involved in ATPase binding to many ABC transporters was only apparent in the last halves of the T subunits, occurring just prior to the second to last TMS (see also Neubauer et al., 2009). Its presence was not detected in the S subunits, which agrees with results from the Rodionov group and others (Rodionov et al., 2009; ter Beek et al., 2011; Zhang et al., 2010). In summary, these analyses reveal a common origin for the S and T subunits.

Genomic distributions of biotin and thiamin transporter components

Genomic clustering patterns and organismal distributions of biotin and thiamin ECF transporter components were next examined in 73 fully sequenced microbial genomes. Of these genomes, ECF homologues were observed in 27 Gram-positive bacteria, 11 Gram-negative bacteria, 1 *Chlamydia* species, 1 *Treponema* species, and 12 archaea. Gene clustering, particularly when within an operon, often indicates participation within the same biological pathway or cellular structure. Genome wide analyses revealed a large degree of genomic clustering for all three components of the

ThiW homologues (42 clusters observed), with clustering of the A and T subunits being the second most common (37 clusters; see tables 3A, 3B and figure 8). In some organisms, including actinobacteria and an archaeon, rare three-domain SAA' fusion proteins were detected. Our results imply a dependency of most of the ThiW S subunits on an energizing module for transport. In contrast, many A/T clusters of ThiW homologues may serve as energizing modules for the S subunits of other ECF homologues in the same organisms (ter Beek et al., 2011). It is possible that the isolated ATPases (14 cases described for ThiW homologues) may also function in capacities unrelated to transport (Castillo & Saier, 2010).

Our analyses identified S and T clusters that are missing the ATPases in both *Lactococcus lactis* subsp. cremoris SK11 and *Haloarcula marismortui* ATCC 43049. For *L. lactis*, it is likely that the three S subunits can function as secondary transporters, share energizing modules with other ECF systems, or function via both mechanisms. For *H. marismortui*, the A subunit gene downstream of the T subunit (gi#55378961) gene is absent from NCBI annotation; however, my BLAST analysis on the intergenic region downstream of the T subunit revealed an encoded single A subunit.

For analysis of biotin transport clusters that appeared to contain S and A subunits but not a T subunit, a hypothetical 5 TMS protein (gi#11498771) was found to be encoded downstream of the S subunit (gi#11498769) and a duplicated ATPase subunit (gi#11498770) in the *Archaeoglobus fulgidus* DSM 4304 genome. BLAST results showed that it is homologous to the T subunits of ECF transporters (data not shown). In the *Methanospirillum hungatei* JF-1 genome, a 5 TMS protein-encoding

gene with remote similarity to the T subunit is encoded by a hypothetical protein (gi#88604411) based on its topology and its position downstream of an S subunit (gi#88604408) and two A subunits (gi#88604409 and gi#88604410). Similarly, a 5 TMS hypothetical protein (gi#124484963) in *Methanocorpusculum labreanum Z* was assigned the function of a T homologue based on positional analysis and BLAST results. These archaeal transport systems have been assigned TC#s 3.A.1.26.8, 3.A.1.25.2 and 3.A.1.25.3, respectively. The latter two systems appear to be biotin uptake porters based on genome context analyses.

From our initial analyses of multiple thiamin transport clusters that apparently contain S and A subunits but not T subunits, a gene (gi#84489907) encoding a potential T homologue in *Methanosphaera stadtmanae* DSM 3091 was found, which is followed by a gene (gi#84489906) encoding a duplicated ATPase and an S homologue (gi#84489905). Operons that contain the complete complement of subunit components were also observed for two other T homologues (gi#29377242 and gi#116516329). However, for gi#116516329, beside the upstream S subunit (gi#116516079), an additional S homologue (gi#116516704) was found to be encoded near the 5' end of this operon. All in all, a majority of clusters that appeared to be missing components of the transport complex turned out to encode complete systems. Genome context analyses should offer confirmation of the identities and substrate specificities of nearby subunits.

Our analyses of BioMNY transporter homologues indicated that a large proportion of S homologues function independently of the other ECF transporter components. Thirty five gene clusters were found to encode just the S homologues, in

contrast to eight that encode the entire complex, suggesting capacities for alternative modes of energization (Hebbeln et al., 2007; Rodionov et al., 2009). Further, the analyses with YjbB and YhaG homologues did not reveal obvious clustering with ATPase genes (data not shown). Our overall analyses hint at the possibility that S subunits of some ECF transporters may, depending on the cellular electrochemical gradient and local substrate concentrations, function independently of ATP-hydrolysis, while others may have evolved to function entirely by an ATP-independent mechanism. This last possibility has been confirmed (Erkens & Slotboom, 2010; Finkenwirth et al., 2010; Hebbeln et al., 2007).

Sequence divergence of the transmembrane ECF components

To provide evidence for functional divergence between S and T subunits, these homologues were compared by constructing phylogenetic trees and determining relative comparison scores. A much greater degree of sequence conservation was observed among T subunits than S subunits, with 9 out of 36 comparisons giving >11 SD for the S homologues and all 36 comparisons giving >11 SD for the T homologues (table 4). The greater sequence divergence among S subunits may correlate with the need for them to recognize a wide variety of micronutrients in the absence of periplasmic receptors, while greater conservation of the corresponding T subunits may reflect their interaction with the highly conserved A subunits.

Phylogenetics of ECF transporter components

The SuperfamilyTree (SFT) programs allowed construction of phylogenetic trees for all ECF homologues under study (Yen et al., 2009; Chen et al., 2011). Phylogenetics of type-I ECF transporters, in which the S, T and A subunits from a given transport complex could be confidently assigned, revealed that, with the exception of the closely related cobalt (CbiMNQO) and nickel (NikMNQO) transporters, there is little evidence of co-evolution, contrary to expectation for systems with dedicated components (Saier, 2003a, 2003b). This could imply either that the subunits became functionally coupled during recent evolution or that there has been substantial shuffling of constituents between these systems during their evolutionary histories (figure 9). However, branches for the biotin transporters (BioMNY), the putative queuosine precursor transporters (QrtTUVW), and the putative cobalamin precursor transporters (CbrTUV) are difficult to analyze as homologues of these permeases also contain representatives from type-II ECF transporters (Rodionov et al., 2009), and their evolutionary histories can not be conclusively established. The elimination of those transporters with just a single S subunit in the operon (presumably of type-II origin) did not improve the overall phylogenetic associations (data not shown).

Phylogenetic analyses of type-II ECF systems revealed co-evolution of the A and T subunits but not the S subunits (see the next section entitled “Phylogenetics of the two paralogous ATPase subunits in type-II ECF transporters”). Thus, type-I ECF transporters, though possessing dedicated energizing modules, must have undergone frequent gene shuffling between constituents during their evolution.

Phylogenetics of the two paralogous ATPase subunits in type-II ECF transporters

From the analyses of type-II ECF energizing modules, orthologous A subunits generally cluster more tightly together than the corresponding paralogues. As shown in the phylogenetic tree, the upstream ATPases all cluster together, separately from the downstream ATPases, showing that throughout the evolutionary histories of these systems, the order of these two genes relative to each other has not changed (figure 10A). This fact must have physiological significance, but its molecular basis remains obscure.

Closer scrutiny revealed that the upstream ATPases from *Bacillus halodurans C-125* and *Oceanobacillus iheyensis* HTE831 (Bha1 and Oih1 in figure 10A) and the upstream ATPase from *Pediococcus pentosaceus* ATCC 25745 (Ppe1 in figure 10A) are not in clusters 6 and 4, respectively, although the downstream Bha2 and Oih2 homologues and the downstream Ppe2 are in clusters 6' and 4', respectively, of the corresponding 3' ATPase cluster. This apparent discrepancy may reflect the limits of the algorithm in resolving distant evolutionary relationships. The inclusion of the T subunit homologue, MpnA, from *Mycoplasma pneumoniae* M129 in cluster 2 of figure 10B agrees with the phylogenetic clustering of its corresponding A subunits. These analyses suggest closer evolutionary associations and provide evidence that the components of the AAT module function together as parts of energizing complexes. These results indicate that ancestral ECF transporters probably arose through duplication of ATPase subunits before they diverged to energize transporters of different specificities. For those transporters that utilize heterodimeric ATPase

complexes, the two ATPase paralogues may have different binding affinities to the S or the T subunits, and this could be a general feature of all of these systems. By contrast, many BioMNY, CbiMNQO and NikMNQO operons contain only one ATPase each, possibly reflecting a subtle distinction between ECF homologues that necessitates the use of a heterodimeric ATPase as opposed to a homodimeric ATPase (Hebbeln et al., 2007; Rodionov et al., 2006, 2009).

Phylogenetics of S subunit homologues of the type-I and type-II ECF porters and of the membrane constituents of ABC2 uptake porters

The S subunit provides the minimal requirement for transport in the ECF system. To reveal phylogenetic relationships between ECF and traditional ABC2 transporters as well as between type-I and type-II ECF transporters, the SFT program was used to construct phylogenetic trees. ABC2 transporters other than ECF porters that have two integrated membrane subunits are thought to have both subunits share substrate binding and interaction with the two ATPase subunits. However, these two functions may not be shared equally by the two subunits. The *E. coli* maltose transport complex, for which high resolution X-ray structures are available, has two membrane subunits, MalF and MalG. The MalF protein appears to have a greater degree of interaction with maltose than the MalG protein, suggesting that MalF may be closer to the S subunits of the ECF transporters than MalG, while MalG may more closely resemble the T subunit of an ECF transport complex (Oldham et al., 2007). However, from structural analysis, it is not entirely clear if the degrees of substrate and ATPase interactions with MalF and MalG change during the transport cycle.

Since many TM subunits of ECF porters do not show extensive similarity to either MalF or MalG, both sequences were included for analysis. When a transport complex contains a homodimeric TM complex, the sole TM subunit was used.

As demonstrated in the superfamily tree in figure 11 (see table 6 for the query sequences used for the construction of the tree), both type-I and -II ECF transporters cluster near the root of the tree, with type-II ECF systems appearing to be more ancient (TC#2.A.88). However, there are also exceptions as TC#3.A.1.26.1 and TC#3.A.1.34.1 appear in clusters along with ABC2 porters, and TC#2.A.88.3.1, TC#2.A.88.4.1, TC#2.A.88.5.1, and TC#2.A.88.7.1 appear latter than most of the type-I ECF. With ABC2 uptake porters, their transmembrane components appear to be much more recent. Based on this distribution and the comparison scores, we concluded that the S subunit homologues of type-I/II ECF and ABC2 transporters have a common ancestry and that the S subunits of type-II ECF transporters more closely resemble the ancestral forms of the S subunits of the ECF systems than those of type-I systems.

Uptake analysis of an ECF homologue from *Mycobacterium smegmatis* str. MC2 155

When 5nM thiamine was present in the M9 minimal medium, both the complete transporter complex (SAA'T) and the S subunit alone (S) showed evidence of uptake after 5 hours as indicated by absorbance measurement at 600nm wavelength (figure 12). Under the conditions used, the uptake conferred by the ThiW homologue is more efficient than the native transporter from the *E. coli* host as is evident by the

logarithmic growth phase and the final cell density. However, in their native configuration, both SAA' and T constructs failed to show any sign of uptake.

Discovery of novel ECF transporters using riboswitch elements

The analysis of riboswitch regulons revealed two novel transporters that could be closely related to ECF transporters in three separate taxonomic groups. The first putative ECF transporter is regulated by the cobalamin riboswitch element and is composed of a core permease component (RoseRS_3102 from *Roseiflexus* sp. RS-1) that showed no sequence similarity with any existing ECF entry in TCDB. However, the other transmembrane component (RoseRS_3101) of this putative ECF transporter showed homology with the T subunit of one of the ECF entries (TC#3.A.1.26.5) with a BLAST score of e^{-7}. The TCDB BLAST result is even better for the A subunit (RoseRS_3100) with a score of e^{-59} against another ECF entry (TC#3.A.1.26.1). Based on this observation, I propose to add this transporter to the expanding list of ECF entries in TCDB.

The other putative ECF transporter found (YpdP; SA1265 from *Staphylococcus aureus* subsp. aureus N315; see also homologues from *Staphylococcus capitis* SK14, *Staphylococcus epidermidis* ATCC 12228, *Staphylococcus carnosus* subsp. carnosus TM300, *Staphylococcus haemolyticus* JCSC1435, *Staphylococcus saprophyticus* subsp. saprophyticus ATCC 15305) was under the regulation of the preQ1 riboswitch. It showed high similarity with members of ECF transporters (TC#2.A.88.8) in TCDB with a maximal score of e^{-114} for SH1474 and STACA0001_1686. SA1265 was used as a representative homologue for

Staphylococcaceae, and it showed 7 TMSs while homologues from Bacillales showed 5 (OB2209) to 6 (ABC0888) TMSs. Binary alignments showed that both OB2209 and ABC0888 lost the C-terminal TMS while OB2209 lost an additional N-terminal TMS. Cluster analysis for all YpdP homologues failed to turn up any energizing module encoded with the core permease component.

Incorporation of riboswitch discovery with comparative genome analyses

Our collection of riboswitches responds to a diverse group of cellular metabolites, and these RNA elements were grouped into six substrate-specific categories: coenzyme regulators, amino acid regulators, ribosome regulators, nucleotide derivative regulators, ion/sugar regulators, and putative riboswitches (see tables 7 to 12 for the names of specific regulators). Some of these riboswitches bind to large macromolecules including protein regulators (such as *pyrR* regulatory sequence); however, they were included in my study as these RNA sequences can form highly structured motifs that are riboswitch-like and presumably regulate their downstream genes via mechanisms similar to the more canonical riboswitches that respond to small effector molecules.

I had discovered that there is a great number of riboswitches present in firmicutes, with an average of 42.7 riboswitches per genome (with a standard deviation of 12.5 sites per genome) (table 13). However, there is also a wide prevalence of riboswitches across the bacterial domain, averaging 12.1 riboswitches per genome (with a standard deviation of 4.7 sites per genome) (table 13). Such a widespread distribution can only be explained if some of the riboswitches had arisen

in the last common ancestor of all or many bacteria. This observation is especially applicable to TPP, cobalamin, FMN, glycine, and yybP-ykoY riboswitches; TPP riboswitches are present in all the phyla analyzed while the rest of these riboswitches are missing only from two to four phyla, possibly due to elimination early in the evolution of the bacterial domain (see tables 7, 8, and 12). The metabolic pathways of these common riboswitches are all essential for cell survival and have been described in detail (Rodionov et al., 2002b/2003a; Vitreschak et al., 2002); however, the biological process that *yybP-ykoY* participates in is not yet known.

Curiously, even though firmicutes and actinobacteria had split relatively recently based on current estimate, the number of riboswitches per genome in Actinobacteria (averaging 11.8 sites per genome) is not markedly different from the other bacterial phyla analyzed (averaging 12.1 sites per genome; see figure 14 and table 13). A plausible explanation is that riboswitches had undergone rapid expansion, both in terms of types and numbers before firmicutes split from actinobacteria. The expansion in firmicutes comes mainly from riboswitches that regulate amino acid biosynthesis, with T-boxes making up the majority (43.7% of total ribswitch for firmicutes vs 19.0% for actinobacteria; see table 14 and figure 15). In contrast, the proportion of riboswitches for coenzymes appears to have contracted relative to other riboswitches (21.9% for firmicutes vs 43.5% for actinobacteria; see table 14 and figure 15). Such a difference may be explained by different metabolic requirements for the two phyla.

Emergence of rare riboswitche elements

Even though some riboswitches may have arisen before the last common ancestor of all bacteria, not all types of riboswitches are initially present. In the long history of life on Earth, there appears to be many instances where riboswitches originated independently in different phyla.

In the deeply rooted phyla of Deinococcus-Thermus, chloroflexi, and thermotogae, I did not observe the presence of certain coenzyme riboswitches. For instance, all riboswitch elements responsive for SAM and SAH coenzymes (except SAM riboswitch) are absent (see table 7). In addition, some amino acid riboswitches (glnA, Leu_leader and Thr_leader; see table 8), some ribosomal subunit riboswitches (S15, L13_leader, and L19_leader; see table 9), some nucleotide derivative riboswitches (preQ1, and preQ1-II; see table 10), some ion riboswitches (*ykoK* and Mg_sensor; see table 11), and almost all putative riboswitches (except *yybP-ykoY*; see table 12) were missing from these phyla. This suggests that these riboswitches did not originate before the last common ancestor of all bacteria. Of these riboswitches, *ylbH*, preQ1-II, SAH, *sucA*, Mg_sensor, SAM-IV, SAM-alpha, SAM-SAH, SMK box, purine, *serC*, *speF*, and *ybhL* riboswitches can only be found in one or two phyla (represented by Bacillales, Streptococcaceae, Lactobacillaceae, Mycobacteriaceae, Rhizobiales, Ralstonia, and Enterobacteriales; see tables 7, 10, 11, and 12), suggesting a much more recent evolutionary innovation.

Of the rare riboswitches, *ylbH* riboswitches are found exclusively in Bacillales and are present in all species analyzed for that taxon except *Bacillus clausii* KSM-K16 (data not shown). This riboswitch regulates the expression of the *ylbH* gene, which encodes an RNA methyltransferase for a small ribosomal subunit, and a *coaD* gene,

which encodes a phosphopantetheine adenylyltransferase gene that is involved in Coenzyme A biosynthesis.

PreQ1-II sites are found almost exclusively in Streptococcaceae and regulate the expression of a queuosine precursor transporter (*queT*) of energy-coupling factor (ECF) family (data not shown). The only streptococcal species with preQ1-II site absent in our analysis is *Streptococcus thermophilus* CNRZ1066. In Lactobacillaceae, preQ1-II sites can be found only in three (*Lactobacillus casei* ATCC 334, *Lactobacillus rhamnosus* GG, and *Pediococcus pentosaceus* ATCC 25745) out of fifteen genomes, indicating possible horizontal gene transfer (HGT) events. Interestingly, even though *queT* genes are present in *Oenococcus oeni PSU-1*, *Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293*, and *Lactobacillus salivarius subsp. salivarius UCC118*, they are not controlled by preQ1-II riboswitches and likely represent a separate HGT event.

SAH riboswitches are found only in some proteobacteria and mycobacteria and control the expression of an adenosylhomocysteinase (*ahcY*), an unknown membrane protein, a 5,10-methylenetetrahydrofolate reductase (*metF*), and a regulator of competence-specific gene (*tfoX*-like) in members of proteobacteria (data not shown). In Mycobacteria, the SAH regulon was found to consist of *metH* gene only and likely have different evolutionary history than the regulons found in proteobacteria.

sucA riboswitches are found exclusively in Ralstonia as well and participate in succinyl-CoA biosynthesis; they control the expression of components of 2-oxoglutarate dehydrogenase complex [dehydrogenase (E1 component, *sucA*),

dihydrolipoamide succinyltransferase (E2 component, *sucB*), and dihydrolipoamide dehydrogenase (E3 component, *lpdA*)].

Mg_sensors control the expression of a magnesium transporter (*mgtA*) and were found in Enterobacteriales solely, although they are present only in five out of twelve genomes analyzed and may represent an evolutionary innovation originating in the common ancestor of *Escherichia coli* str. K-12 substr. MG1655, *Citrobacter koseri* ATCC BAA-895, *Salmonella typhimurium* LT2, *Klebsiella pneumoniae* subsp. pneumoniae MGH 78578, and *Enterobacter* sp. 638. As an additional evidence for vertical transmission, *mgtA* homologues with their own Mg_sensor elements form a distinct clade near the root of the phylogenetic tree (see figure 16), lending support to the concept of independent emergence of this riboswitch in the last common ancestor of a few enterobacterial lineages.

SAM-IV riboswitches are present only in Mycobacteriaceae and regulate the expression of an o-acetylhomoserine sulfhydrylase (*metY*), a homoserine o-acetyltransferase (*metX*), and a putative methyltransferase in all the Mycobacteriaceae species analyzed except *Mycobacterium leprae* TN. In addition, a probable aminotransferase/cysteine desulfurase (gi#169627616) regulated by SAM-IV is found in *Mycobacterium abscessus* ATCC 19977; the homologue of the gene can also be found in other families in the order of Actinomycetales although whether any of them are regulated by riboswitches or participate in methionine biosynthesis has yet to be established.

SAM-alpha can be found in Rhodobacterales and Rhizobiales. Similar to SAM-IV, SAM-alpha regulates the expression of *metY* and *metX*. In addition, a

homoserine o-succinyltransferase (*metA*) gene and a methionine biosynthetic gene of unknown function (*metW*) were found to be regulated by the SAM-alpha riboswitch element, frequently in different clusters. *metX* and *metW* were typically found organized in an operon in Rhizobiales. Additional *metX-metW* operons not regulated by SAM-alpha were also found in some species of the Rhizobiales. Phylogenetic analysis of the MetX homolgoues revealed a possible deletion of the SAM-alpha regulatory elements in some lineages of Rhizobiales (AZC_0229 and Xaut_1909 in figure 17).

SerC riboswitches were found in the majority of alpha-proteobacteria analyzed. In almost all of these organisms, the regulons this riboswitch regulates were highly conserved and consisted of *serC* and *serA* genes (encoding a phosphoserine aminotransferase and a D-3-phosphoglycerate dehydrogenase, respectively) in the serine biosynthetic pathway. Their limited distribution may point to their emergence in alpha-proteobacteria.

SpeF riboswitch sequences and their regulon contents are also highly conserved. The *speF* gene regulated by this riboswitch likely participates in polyamine biosynthesis, and its orthologues are present in all the Rhizobiales species analyzed. This riboswitch element may have originated in Rhizobiales as it was found only in Rhizobiales species; however, *Bradyrhizobium japonicum* USDA 110, *Bradyrhizobium* sp. BTAi1, and *Rhizobium etli* CFN 42 did not seem to contain the regulatory element upstream of their *speF* genes. When ploted in a phylogenetic tree using the SpeF protein sequences, unregulated homologues from *Bradyrhizobium* and that from *Rhizobium* were found in two distinct clusters (blr7759, BBta_1349, and

RHE_CH03629 in figure 18). This result suggests a deletion of regulatory element in these two genera in two independent events. Like other *speF* regulons, RHE_CH03629, BBta_1349 and blr7759 contain an acetyltransferase gene (except *Brucella melitensis* 16M) following the *speF* genes. From genome context and BLAST analyses, *Bradyrhizobium* sp. BTAi1 and *Rhizobium etli* CFN 42 probably lost their riboswitch sequences recently due to homologous recombination; however, the upstream distance from the previous gene is relatively conserved (as compared to closely related species) for *Bradyrhizobium* sp. BTAi1 but has shortened considerably for *Rhizobium etli* CFN 42. *Bradyrhizobium japonicum* USDA 110 likely had lost its riboswitch sequence in the process of intrachromosomal shuffling as the neighboring genes are not conserved at all. My evidence points to recent uncoupling of *speF* riboswitch elements from the regulons they control.

ybhL riboswitches, which regulate the expression of uncharacterized *ybhL* membrane protein, were also found only in Rhizobiales although only in 8 out of 15 species analyzed. From my genome context and BLAST analyses, the riboswitch sequence in *Bartonella quintana* str. Toulouse was likely lost due to intrachromosomal shuffling as its chromosomal neighbors are not conserved. Phylogenetic analysis of YbhL protein sequences revealed high degrees of correlation between *ybhL* genes regulated with the regulatory element and those that are not regulated, with riboswitch regulated genes clustered at the bottom of the tree (figure 19). The clustering of *Bartonella quintana* str. Toulouse YbhL homologue (BQ00080) with other riboswitch regulated YbhL homolgoues supports my genome context and BLAST analyses. My

evidence points to the coevolution of *ybhL* riboswitch elements and the *ybhL* genes they control.

Evolution of functionally equivalent riboswitches

There are many homolgous genes discovered in this study to be regulated by functionally equivalent riboswitches. For instance, the *metX* gene, which regulates the conversion of homoserine to o-acetyl-homoserine, can be found with the SAM, SAM-alpha, and SAM-IV riboswitches. There is a great degree of clustering of riboswitch element with the respective *metX* homologues they regulate on a phylogenetic tree based on MetX protein sequence data (see figure 20). Similarly, *metY* homologues cluster with their respective riboswitch elements (see figure 21). However, there are two cases of Staphylococcal homologues (Sca_I_Staph and Mca_I_Staph in figure 21) clustering with the clostridial homologues. These two homologues likely were transferred to members of Staphylococcaceae as a result of horizontal gene transfer from members of Clostridiaceae.

I had also observed two major clusters for clostridial MetY homologues. The bottom cluster often have *metA* downstream of *metY* in an operon (except Ckl2_I_Clostridi in figure 21). In contrast, the top cluster usually only has *metY* in its own operon (all except Mca_I_Staph and Sca_I_Staph in figure 21). The operonic differences between these two clusters may reflect their different evolutionary histories. A Bacillale MetY homologue was also found to be clustered with one of the two major Clostridial cluster. The sole occurrence of this Bacillale homologue indicates its origin in Clostridiaceae.

In addition to the unique MetY orthologue found in each organism, three MetY paralogues (Ckl_I_Clostridi, Ckl2_I_Clostridi, and Ckl3_I_Clostridi in figure 21) can be found for *Clostridium kluyveri* DSM 555. The three MetY paralogues were highly divergent and were found in different clusters, including a paralogue (Ckl3_ I_Clostridi) not regulated by any SAM-responsive riboswitch. Ckl3_ I_Clostridi possibly functions in some as-yet undiscovered biochemical reactions not directly tied to the biosynthesis of methionine. Paralogues (Oih_I_Bacillales and Oih2_I_Bacillales in figure 21) from *Oceanobacillus iheyensis HTE831* were also highly divergent in sequence and may participate in different biochemical pathways.

Other evidence of riboswitchs coevolution with the genes they regulate can be found in *ykkC* genes and the two divergent riboswitches that regulate their expression. As with sequence divergent riboswitches that respond to SAM, there appeared to be a large degree of clustering of *ykkC-yxkD* and mini-*ykkC* riboswitches with the *ykkC* genes they regulate. However, exceptions were found in two pseudomonal homologues (Pst_mini_Pseudomona and Pae2_mini_Pseudomona in figure 22), which cluster with the solitary ralstonial (Cta_mini_Ralstonia in figure 22) and desulfovibrial (Dde_mini_Desulfo in figure 22) homologues as opposed to the main pseudomonal cluster in the right-hand side of the tree. There also appeared to be a horizontal gene transfer event from Rhizobiales that accounts for the sole appearance of a cyanobacterial *ykkC* homologue (Gvi_mini_Cyano in figure 22).

Chapter III, in full, is in press in the International Journal of Bioinformatics, 2012. Sun, Eric; Saier, Milton. The dissertation author was the primary investigator and author of this paper.

Chapter III, in part, is currently being prepared for publication. Sun, Eric; Leyn, Semen; Kazanov, Marat; Novichkov, Pavel; Saier, Milton; Rodionov, Dmitry. The dissertation author was the primary investigator and author of this material.

IV.

Discussion

I here describe my efforts to trace the evolutionary origins of what I have shown to be a recently proposed sub-superfamily of ABC transporters, the ECF transporters. These differ from typical ABC uptake porters in having no extra-cytoplasmic binding receptor and having highly divergent integral membrane subunits that are believed to perform different functions. I confirmed homology for the BioY, ThiW, and TrpP families using quantitative statistical means. Additionally, I provided evidence for homology with the PNaS symporter family. The fact that ECF transporters can function by two distinct energy coupling mechanisms (Hebbeln et al., 2007 and this dissertation) suggests that: (1) the ECF sub-superfamily might be at evolutionary crossroads between secondary and primary active transport, and (2) additional examples of transporters that are capable of utilizing alternative modes of energy-coupling may exist.

From our sequence analyses, the S subunits of the ECF sub-superfamily members appear to have duplicated from a 3-TMS peptide to give rise to 6-TMS transporters (figures 3A, 3B, and 4). To this 6 TMS unit, a TMS apparently was lost or gained in some members to give 5 or 7 TMS proteins, respectively. In some cases, the core 5 or 6 TMS S subunits may have duplicated internally to give rise to 10 or 12 TMS proteins, respectively (see Zheng et al., manuscript in preparation). Transporters of the PNaS family appear to have undergone an intragenic duplication event where a 4 TMS unit duplicated to give 8 TMS proteins (figure 5 and 3C). Even though the two integral membrane components of an ECF transporter probably function in different capacities, the S- and T-subunit alignments indicated a common origin (figures 6 and 7). When checking for internal duplications in T subunits, we observed little sequence

conservation between their two halves compared to those of the corresponding S subunits. Recognizing their common origin as shown here, however, suggests that the two halves of T-subunits have become functionally differentiated to a greater degree than the two halves of S-subunits, thereby masking their common origin. Interestingly, crystallographic data have revealed that the two halves of the S subunit may together comprise a central transport channel (Zhang et al., 2010), implying the importance of structural symmetry despite the divergence of the two transmembrane subunits in ECF systems. Despite poor sequence similarity, we did find similar numbers of transmembrane □□□□□-helices as well as similar spatial distributions in the S and T subunits. Taken together with the published size-exclusion and crystallographic data (Erkens & Slotboom, 2010; Zhang et al., 2010; Erkens et al., 2011), it seems likely that the transport mechanism of ECF transporters resembles that of traditional ABC transporters. They may have diverged in quantitative detail.

The two ECF subunits appear to have distinct functions in accordance with their extensive sequence divergence. One subunit seems to be primarily dedicated to substrate recognition while the other may function primarily to anchor the cytoplasmic ATPase to the complex and coordinate energization of the transport cycle. Such functional divergence may not be restricted to ECF transporters, as even within traditional ABC uptake transporters, there seems to be a continuum of transporters that have partial segregation of function. At one extreme, the two integral membrane components of homodimeric transporter complexes (such as the arabinose, ribose and galactose ABC transporters of *E. coli*) perform dual, shared roles of both substrate recognition and ATPase anchoring equally. On the other hand, the well-studied ABC

transporter, $MalEFGK_2$ of *E. coli*, seems to exhibit partial functional differentiation for its two integral membrane components as revealed by X-ray crystallography (Oldham et al., 2007). For $MalEFGK_2$, one subunit may primarily bind the substrate, even though both form the channel. Nevertheless, interactions of both subunits with the dimeric cytoplasmic ATPase appear to be equally critical for transport, as reflected by the existence of an ATPase-binding 'EAA' motif in both TM subunits (Oldham et al., 2007; Davidson et al., 2008). In the ECF transporters, we confirmed the existence of a modified EAA motif before the second to last TMS in the T subunit (see also Neubauer et al., 2009), but we could not find this motif in the S subunits. It is unknown whether the S subunit interacts directly with an A subunit; however, such an association is likely given our knowledge of the mechanism of transport by ABC transporters (Davidson et al., 2008). Combined with crystallographic data, it is probable that the positively charged residues in the loop near the C-terminus carry out such a function (Zhang et al., 2010). In accordance with its role as an ATPase anchor and the apparent lack of interaction with substrate, we hypothesize that a greater degree of sequence conservation could be observed among the T subunit homologues. Indeed, when we compared the sequences among all the homologues, we observed a much greater degree of sequence conservation among the T subunits than among the S subunits (table 4).

Operon structures for members of the ECF family appear quite diverse, and modes of energy coupling can often be inferred if no functional data are available. For instance, the S domain of the mycobacterial ThiW is fused to two ATP-hydrolyzing (A) domains, clearly implying ATP-dependent energization. Additionally, BioY

homologues can be encoded in operons together with ABC-type energizer-encoding genes, but there appears to be many more BioY permeases encoded in operons lacking components of ABC-type ATPases than possessing them (table 3B and figure 8). However, prokaryotic YhaG and YjbB family members were not found to be encoded in operons with ATP-hydrolyzing proteins; in the latter case, a phosphate:Na^+ symport mechanism is established for homologous eukaryotic systems (Kohler et al., 2001; Lebens et al., 2002; Murer et al., 2000; Miyamoto et al., 2007; Tenenhouse, 2005). It is reasonable to assume that some of these ECF transporters can function by either primary or secondary active transport, and possibly by both, depending on the availability of the ATP-hydrolyzing subunits, as first demonstrated for the biotin transporter of *Rhodobacter capsulatus*. (Hebbeln et al., 2007)

Meanwhile, the number of ECF members that can use either primary or secondary active transport is continuing to expand. Our preliminary functional analyses of the ThiW transporter from *Mycobacterium smegmatis* expressed in *E. coli* showed that either the S subunit alone or the complete energy-coupled transporter complex could support growth of a thiamine-auxotrophic *E. coli* strain. However, synthesis of just the T or SAA' protein did not support growth (figure 12). This result suggests that T subunit alone lacks activity, and it implied an inhibitory function for the ATPases in the absence of a complete system. A proposal for alternative modes of energy coupling has been presented for certain PTS permeases (Hvorup et al., 2003; Saier et al., 2005), and experimental data supporting such a postulate for the ArsB arsenite exporters of *E. coli* and *Staphylococcus aureus* have been presented (Bröer et al., 1993; Kuroda et al., 1997; Xu et al., 1998; see also Castillo & Saier, 2010).

Clearly, transporters capable of alternative modes of energization are more wide spread than originally thought, although the physiological relevance of these two modes of transport and their potential in transport regulation still need to be evaluated.

Like typical ABC superfamily transporters, members of the ECF sub-superfamily always require two monophyletic cytoplasmic ATPases, which are required for cooperative ATP hydrolysis. However, there appears to be divergence between the ATPase subunits or domains of the ECF porters as compared with the corresponding subunits of traditional ABC porters, which are frequently composed of homodimeric integral membrane complexes. Based on our phylogenetic analyses on components of type-II ECF energizing modules, divergence of the paralogous ATPases appears to have occurred early during evolution of this sub-superfamily before transporters of different specificities evolved, and they show high degrees of conservation across different microbial genomes (figure 10 and table 5). However, a similar degree of coevolution was not observed for components of type-I ECF transporters (figure 9). The discrepancy may reflect a consequence of different evolutionary pressures than those faced by type-II ECF transporters.

Genomic organization of the ECF transporter-encoding operons also appears to reflect some fundamental differences from ABC transporters. For example, there appear to be few instances of fusions of ECF transmembrane subunits to their cytoplasmic ATPases (S to A or T to A). Fusions of the two membrane subunits (S and T) were also rare, although fusion of the two ATPases occurred more frequently, as was first demonstrated by Rodionov et al. (2009). On the other hand, various combinations of fusions of the transmembrane subunits to the ATPase subunits and of

the transmembrane subunits to each other are common in the traditional ABC transporters, especially in exporters. The lack or scarcity of these combinations of ECF fusions may point to structural and/or functional constraints such as flexibility requirements for substrate binding or coupling of ATP hydrolysis to transport. However, it could also indicate their status as being more similar to the ancestral transporter.

Given the lack of periplasmic binding proteins and the ability of some ECF carriers to function in two different energizing capacities, we believe that the ECF transporters may indeed represent a case of transitional transporters that originally contained just the integral membrane components. This leads to the proposal that homologous traditional ABC transporters evolved via the same route. Indeed, when we constructed a superfamily tree using homologues of type-I and type-II ECF porters as well as homologues of the more conventional ABC2 uptake porters, we observed that type-II ECF porters tend to cluster near the root of the superfamily tree, again suggesting that they may more closely resemble ancestral forms of both type-I ECF and ABC2 porters (figure 11 and table 6). Members of the ECF transporter sub-superfamily probably arose from the same ancestor as ABC2 transporters, from a common prototypical 3-TMS ancestor (figures 11; see also Wang et al., 2009). The fact that the S subunits of type-II ECF porters are encoded in operons separate from those of the energizing modules indicates that these S homologues likely started out as secondary carriers and later became functionally integrated with the energizing modules, which may have originally functioned with other transporters or in some capacity other than transport. As ECF transporters evolved into the more common

ABC2 porters, they may have become functionally dependent on extracytoplasmic receptors, thus freeing the membrane subunits from the constraints of high affinity substrate binding (figure 13). The incorporation of periplasmic receptors may have imparted tighter control of the transport cycle and allowed a greater diversity of substrates to be specifically recognized. This proposal is strengthened by the observation that some ABC porters can function with any of several extracytoplasmic substrate-binding receptors (see TC#3.A.1.5.18, 3.A.1.5.25, 3.A.1.12.12 and 3.A.1.20.1)

Additional evidence for such a transition comes from an early study demonstrating that mutation of a TM subunit of a typical ABC transporter could induce periplasmic-protein-independent uptake (Treptow & Shuman, 1985). It would be interesting to compare the structures of ECF family members with the more carefully studied ABC transporters to determine if there are subtle differences in how the ATPases are arranged in the transport complex and whether such an arrangement in ECF transporters may present an impediment to the emergence of fused subunits, although their analyses are beyond the scope of this dissertation.

The evolution of ATP-dependent solute uptake occurred many times (i.e., ArsA/ArsB systems, ABC systems, cation transporting P-type ATPases [Thever & Saier, 2009; Chan et al., 2010; Xu et al., 1998] and many protein secretion systems [Saier, 2006; Saier et al., 2008]). ECF porters clearly show a common origin with ABC systems. This conclusion applies to both membrane and cytoplasmic components as shown here. Tripartite pmf-driven systems independently evolved in the TRAP transporters (Rabus et al., 1999), which share with ABC systems the

characteristic of possessing an extracytoplasmic solute-binding receptor that provides vectorial directionality (Mulligan et al., 2007, 2009). The route taken and the survival values of the specific characteristics of these systems have yet to be fully appreciated.

As many ECF transporters are regulated by riboswitches (Rodionov et al., 2009), we have conducted genomic analyses for the presence of riboswitches in fully sequence bacterial genomes to search for transporters that occur as ECF systems. Our search revealed putative ECF homologues from *Roseiflexus* sp. RS-1 and members of Staphylococcaceae and Bacillales. The putative ECF system from *Roseiflexus* sp. RS-1 appears to be composed of a full complement of ECF subunits with both the T and A subunits showing high similarity to existing ECF entries in TCDB although the putative core permease S subunit did not show recognizable similarity. We have added this additional system to TCDB. The putative ECF systems (YpdP) from members of of Staphylococcaceae and Bacillales, though lacking the energizing components of the ECF system, turned out to show similarity with existing ECF entries in TCDB. With the riboswitch identified for these putative ECF transporters, it should be possible to design experiments to test for putative substrates. We should be able to uncover additional ECF systems when more genomes are searched.

As a follow-up to the ECF investigation, I also carried out more in-depth analyses on the organismal distributions of riboswitches. Based on the distribution observed for TPP, cobalamin, FMN, glycine, and *yybP-ykoY* riboswitches, it is probable that these riboswitches originated in the last common ancestor of all bacteria. This is logical as TPP, cobalamin, and FMN are essential coenzymes for bacteria, and glycine is an essential amino acid and a basic building block for bacteria. However,

the wide distribution of *yybP-ykoY* riboswitch is unclear. The associated *ykoY* gene may code for a tellurium resistance transporter and showed high similarity to transmembrane proteins of uncharacterized function in TCDB (TC# 9.A.30.1). The co-occurrence of *ykoY* with the regulatory element is quite high with 75 homologues discovered out of 242 *yybP-ykoY* riboswitches analyzed. The next most commonly associated gene (58 homologues found) codes for a predicted membrane protein (COG2119). COG2119 showed high similarity to a TCDB entry (TC#9.B.26.1), but its function is not known. Except in Vibrionales, COG2119 did not appear to occur together with *ykoY* with appreciable frequency and may have a function complementary to that of YkoY (data not shown). The next most common genes, *atcL* (46 homologues found), code for a cation-transporting P-type ATPase homologue (TC#3.A.3.2), which likely transports calcium to the extracellular matrix to assist in biofilm formation (Sarkisova et al., 2005). A similar function is also probably carried out by another gene, COG0530 (sodium/calcium exchanger). The associated *yybP* gene supposedly encodes a putative sercreted protein, although there was no corresponding TCDB entry to allow identification of any potential transport partner, and the *yybP* gene is the only gene in its regulon. Even though the regulatory element is partially named after the *yybP* gene, the distribution of *yybP* genes appeared to be restricted to Bacillales and was not found in other taxa. In addition to putative calcium exporters, there were proteases found to be regulated by *yybP-ykoY* riboswitches, though they were found in smaller numbers than putative transporters/transmembrane proteins and were restricted to members of the Desulfovibrionales, Clostridiaceae, and Deinococcus-Thermus. In all, *yybP-ykoY*

riboswitches probably regulate genes that are associated with virulence and biofilm formation instead of an essential biochemical pathway. These results illustrate the diversity of cellular processes that riboswitches regulate.

For the organisms missing these common riboswitches, it should be possible to determine if any of these regulatory sequences are missing due to replacement by an alternative regulatory sequence. (Such a replacement scenario is likely for phyla Chlorobi and Bacteroidetes as most of their major riboswitch classes were missing from orthologous regulons [see tables 8~12, 14 and figure 15].) This can be accomplished by analyzing the upstream UTR from any of the orthologous operons. The analysis of 5' UTRs from homologous regulons should also help to refine the model used in constructing consensus RNA sequences to search for more distantly related sequences (Novichkov et al., 2010). In addition, regulon content analysis will provide complementary information to direct sequence comparisons between orthologous riboswitch elements. It may be possible to reconstruct evolutionary scenarios after more genomes are analyzed.

Although most riboswitches are fairly long and complex in their three-dimensional structures (TPP and cobalamin riboswitches for instance), it is possible that some of these regulatory sequences could have arisen spontaneously in the regulatory region of functionally relevant genes as some are fairly short (preQ1, mini-*ykkC*, L19-leader and *serC* for example), making concrete phylogenetic analyses difficult. However, the distribution of these short riboswitch sequences appeared to be more restricted in their distribution (see tables 9, 10, and 12). This would make biological sense as linkage of regulatory elements to functionally unrelated genes

should be detrimental in most cases and thererfore would not become fixed in the genome. Once a riboswitch is coupled to the regulation of functionally relevant genes, the operon structure would be expected to be preserved and transferred as a functional unit (as in the case of HGT). My observation of the co-occurrence of riboswitches with the genes they regulate held in most cases observed (even for genes controlled by riboswitches that are functionally equivalent) and strongly supports this argument (see figures 16~22).

My analyses also revealed that proportion of some riboswitch classes, noticeably those that are not responsible for coenzyme regulation, appeared to be in fluctuation when going from phylum to phylum (see table 14 and figure 15). And even within the coenzyme riboswitch class, proportions of different riboswitch orthologues tend to change (see table 7). Although the types of interaction between riboswitches and their ligands may be limited due to the absence of a protein mediator, the organization of binding domains and expression platforms is also crucial in determining the regulatory outcome upon substrate binding. This organizational flexibility may permit different regulatory mechanisms to exist in orthologous riboswitches, allowing different responses to occur in metabolically diverse organisms. A comprehensive analysis of orthologous riboswitch sequences in regard to this organizational difference is the next logical step in these continuing investigations.

V.

Figures and Tables

Table 1*:* Fully sequenced microbial genomes used for riboswitch analyses. Genomic data were obtaineded from GenBank. The results were derived from manually curated data collected on RegPredict (http://regpredict.lbl.gov/regpredict/).

Taxonomic group	Organism name
Shewanella	*Shewanella amazonensis* SB2B
	Shewanella baltica OS155
	Shewanella denitrificans OS217
	Shewanella frigidimarina NCIMB 400
	Shewanella halifaxensis HAW-EB4
	Shewanella loihica PV-4
	Shewanella oneidensis MR-1
	Shewanella pealeana ATCC 700345
	Shewanella piezotolerans WP3
	Shewanella sediminis HAW-EB3
	Shewanella sp ANA-3
	Shewanella sp MR-4
	Shewanella sp MR-7
	Shewanella sp W3-18-1
	Shewanella woodyi ATCC 51908
	Shewanella putrefaciens CN-32
Enterobacteriales	*Erwinia amylovora* ATCC 49946
	Salmonella typhimurium LT2
	Photorhabdus luminescens subsp. laumondii TTO1
	Escherichia coli str. K-12 substr. MG1655
	Edwardsiella tarda EIB202
	Erwinia carotovora subsp. atroseptica SCRI1043
	Citrobacter koseri ATCC BAA-895
	Yersinia pestis KIM
	Enterobacter sp. 638
	Serratia proteamaculans 568
	Proteus mirabilis HI4320
	Klebsiella pneumoniae subsp. pneumoniae MGH 78578

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Staphylococcaceae	*Staphylococcus aureus* subsp. aureus N315
	Staphylococcus capitis SK14
	Staphylococcus epidermidis ATCC 12228
	Staphylococcus carnosus subsp. carnosus TM300
	Staphylococcus haemolyticus JCSC1435
	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
	Macrococcus caseolyticus JCSC5402
Bacillales	*Bacillus subtilis* subsp. subtilis str. 168
	Bacillus amyloliquefaciens FZB42
	Bacillus pumilus SAFR-032
	Bacillus licheniformis DSM 13
	Anoxybacillus flavithermus WK1
	Geobacillus kaustophilus HTA426
	Bacillus cereus ATCC 14579
	Bacillus halodurans C-125
	Bacillus clausii KSM-K16
	Oceanobacillus iheyensis HTE831
	Paenibacillus sp. JDR-2
Desulfovibrionales	*Desulfovibrio vulgaris* Hildenborough
	Desulfovibrio vulgaris str. Miyazaki F
	Desulfovibrio desulfuricans G20
	Desulfovibrio desulfuricans subsp. desulfuricans str. ATCC 27774
	Desulfovibrio piger ATCC 29098
	Desulfovibrio salexigens DSM 2638
	Desulfovibrio magneticus RS-1
	Lawsonia intracellularis PHE/MN1-00
	Desulfomicrobium baculatum DSM 4028
	Desulfohalobium retbaense DSM 5692

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Thermotogales	*Thermotoga maritima* MSB8
	Thermotoga sp. RQ2
	Thermotoga neapolitana DSM 4359
	Thermotoga petrophila RKU-1
	Thermotoga naphthophila RKU-10
	Thermotoga lettingae TMO
	Thermosipho africanus TCF52B
	Thermosipho melanesiensis BI429
	Fervidobacterium nodosum Rt17-B1
	Petrotoga mobilis SJ95
	Thermotogales bacterium TBF 19.5.1
Ralstonia	*Ralstonia eutropha* H16
	Cupriavidus taiwanensis str. LMG19424
	Cupriavidus metallidurans CH34
	Ralstonia eutropha JMP134
	Ralstonia solanacearum GMI1000
	Ralstonia pickettii 12J
Cyanobacteria	*Synechococcus* sp. SYNPCC7002
	Synechocystis sp. SYNPCC 6803
	Cyanothece sp. ATCC 51142
	Cyanothece sp. SYNPCC 8801
	Cyanothece sp. SYNPCC 7425
	Microcystis aeruginosa NIES-843
	Nostoc sp. SYNPCC 7120
	Trichodesmium erythraeum IMS101
	Synechococcus elongatus SYNPCC 7942
	Prochlorococcus marinus MIT 9313
	Synechococcus sp. JA-3-3Ab
	Synechococcus sp. WH 8102
	Gloeobacter violaceus SYNPCC 7421
	Thermosynechococcus elongatus BP-1

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Bacteroidaceae	*Bacteroides vulgatus* ATCC 8482
	Bacteroides coprophilus DSM 18228
	Bacteroides dorei DSM 17855
	Bacteroides eggerthii DSM 20697
	Bacteroides fragilis NCTC 9343
	Bacteroides uniformis ATCC 8492
	Bacteroides cellulosilyticus DSM 14838
	Bacteroides plebeius DSM 17135
	Bacteroides ovatus ATCC 8483
	Bacteroides thetaiotaomicron VPI-5482
	Bacteroides stercoris ATCC 43183
Corynebacteriaceae	*Corynebacterium glutamicum* ATCC 13032
	Corynebacterium efficiens YS-314
	Corynebacterium diphtheriae NCTC 13129
	Corynebacterium aurimucosum ATCC 700975
	Corynebacterium jeikeium K411
	Corynebacterium urealyticum DSM 7109
	Corynebacterium amycolatum SK46
	Corynebacterium kroppenstedtii DSM 44385
Streptococcaceae	*Lactococcus lactis* subsp. cremoris SK11
	Lactococcus lactis subsp. lactis Il1403
	Streptococcus thermophilus CNRZ1066
	Streptococcus agalactiae 2603V/R
	Streptococcus uberis 0140J
	Streptococcus equi subsp. zooepidemicus MGCS10565
	Streptococcus dysgalactiae subsp. equisimilis GGS_124
	Streptococcus pyogenes M1 GAS
	Streptococcus gallolyticus UCN34
	Streptococcus mutans UA159
	Streptococcus suis 05ZYH33
	Streptococcus mitis B6
	Streptococcus pneumoniae TIGR4
	Streptococcus gordonii str. Challis substr. CH1
	Streptococcus sanguinis SK36

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Lactobacillaceae	*Lactobacillus sakei* subsp. sakei 23K
	Lactobacillus casei ATCC 334
	Lactobacillus rhamnosus GG
	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365
	Lactobacillus johnsonii NCC 533
	Lactobacillus helveticus DPC 4571
	Lactobacillus acidophilus NCFM
	Lactobacillus salivarius subsp. salivarius UCC118
	Pediococcus pentosaceus ATCC 25745
	Lactobacillus brevis ATCC 367
	Lactobacillus plantarum WCFS1
	Lactobacillus fermentum IFO 3956
	Lactobacillus reuteri JCM 1112
	Oenococcus oeni PSU-1
	Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293
Clostridiaceae	*Clostridium cellulolyticum* H10
	Clostridium kluyveri DSM 555
	Clostridium novyi NT
	Clostridium acetobutylicum ATCC 824
	Clostridium perfringens ATCC 13124
	Clostridium butyricum 5521
	Clostridium beijerincki NCIMB 8052
	Clostridium tetani E88
	Clostridium botulinum A str. ATCC 3502
	Clostridium bartlettii DSM 16795
	Clostridium hiranonis DSM 13275
	Clostridium difficile 630
	Clostridium sp. OhILAs
	Clostridium leptum DSM 753
	Clostridium sp. L2-50
	Clostridium sp. SS2/1
	Clostridium phytofermentans ISDg
	Clostridium nexile DSM 1787
	Clostridium scindens ATCC 35704
	Clostridium bolteae ATCC BAA-613

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Chloroflexi	*Herpetosiphon aurantiacus* ATCC 23779
	Chloroflexus aggregans DSM 9485
	Chloroflexus sp. Y-400-fl
	Roseiflexus sp. RS-1
	Roseiflexus castenholzii DSM 13941
Chlorobiales	*Chlorobaculum parvum* NCIB 8327
	Chlorobium chlorochromatii CaD3
	Chlorobium ferrooxidans DSM 13031
	Chlorobium limicola DSM 245
	Chlorobium phaeobacteroides BS1
	Chlorobium phaeobacteroides DSM 266
	Chloroherpeton thalassium ATCC 35110
	Pelodictyon luteolum DSM 273
	Pelodictyon phaeoclathratiforme BU-1
	Prosthecochloris aestuarii DSM 271
	Prosthecochloris vibrioformis DSM 265
Rhizobiales	*Xanthobacter autotrophicus* Py2
	Azorhizobium caulinodans ORS 571
	Nitrobacter winogradskyi Nb-255
	Rhodopseudomonas palustris CGA009
	Bradyrhizobium japonicum USDA 110
	Bradyrhizobium sp. BTAi1
	Rhizobium sp. NGR234
	Sinorhizobium meliloti 1021
	Rhizobium leguminosarum bv. viciae 3841
	Rhizobium etli CFN 42
	Agrobacterium tumefaciens str. C58 (Cereon)
	Mesorhizobium loti MAFF303099
	Mesorhizobium sp. BNC1
	Brucella melitensis 16M
	Bartonella quintana str. Toulouse
Deinococcus-Thermus	*Deinococcus geothermalis* DSM 11300
	Deinococcus deserti VCD115
	Deinococcus radiodurans R1
	Thermus aquaticus Y51MC23
	Thermus thermophilus HB27

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Mycobacteriaceae	*Mycobacterium abscessus* ATCC 19977
	Mycobacterium avium 104
	Mycobacterium marinum M
	Mycobacterium leprae TN
	Mycobacterium tuberculosis H37Rv
	Mycobacterium smegmatis str. MC2 155
	Mycobacterium vanbaalenii PYR-1
	Mycobacterium flavescens PYR-GCK
	Mycobacterium sp. JLS
Pasteurellales	*Haemophilus parasuis* SH0165
	Haemophilus ducreyi 35000HP
	Actinobacillus pleuropneumoniae serovar 7 str. AP76
	Haemophilus somnus 2336
	Actinobacillus succinogenes 130Z
	Mannheimia succiniciproducens MBEL55E
	Pasteurella multocida subsp. multocida str. Pm70
	Haemophilus influenzae Rd KW20
	Aggregatibacter aphrophilus NJ8700
Pseudomonadaceae	*Pseudomonas stutzeri* A1501
	Pseudomonas aeruginosa PAO1
	Pseudomonas mendocina ymp
	Pseudomonas entomophila L48
	Pseudomonas putida KT2440
	Pseudomonas syringae pv. tomato str. DC3000
	Pseudomonas fluorescens Pf-5
	Azotobacter vinelandii AvOP
Vibrionales	*Photobacterium profundum* SS9
	Vibrio angustum S14
	Vibrio salmonicida LFI1238
	Vibrio fischeri ES114
	Vibrio shilonii AK1
	Vibrio vulnificus CMCP6
	Vibrio cholerae O1 biovar el tor str. N16961
	Vibrio harveyi ATCC BAA-1116
	Vibrio parahaemolyticus RIMD 2210633
	Vibrio splendidus LGP32

Table 1: **Fully sequenced microbial genomes used for riboswitch analyses**. Continued.

Taxonomic group	Organism name
Burkholderia	*Burkholderia phymatum* STM815
	Burkholderia xenovorans LB400
	Burkholderia glumae BGR1
	Burkholderia mallei ATCC 23344
	Burkholderia pseudomallei K96243
	Burkholderia sp. 383
	Burkholderia cepacia AMMD
	Burkholderia vietnamiensis G4
Caulobacterales	*Phenylobacterium zucineum* HLK1
	Caulobacter sp. K31
	Caulobacter segnis ATCC 21756
	Caulobacter crescentus CB15
Rhodobacterales	*Rhodobacter sphaeroides* 2.4.1
	Paracoccus denitrificans PD1222
	Jannaschia sp. CCS1
	Rhodobacterales bacterium HTCC2654
	Oceanicola granulosus HTCC2516
	Loktanella vestfoldensis SKA53
	Oceanicola batsensis HTCC2597
	Roseovarius nubinhibens ISM
	Roseovarius sp. 217
	Sulfitobacter sp. EE-36
	Silicibacter TM1040
	Silicibacter pomeroyi DSS-3
	Roseobacter sp. MED193
	Hyphomonas neptunium ATCC 15444
	Oceanicaulis alexandrii HTCC2633

Table 2: Comparison scores for the transmembrane porters derived from four permease families. The scores represent the average of 6 independent GAP alignments with 500 random shuffles each.

Comparison scores (in standard deviations)			
	ThiW	YhaG	YjbB
BioY	11.8	9.5	9.4
ThiW	-	11.0	10.2
YhaG	-	-	10.3

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. The protein gi numbers were obtained from Genebank. Bold numbers indicate that the protein is composed of two fused ATPases. gi numbers listed in more than one subunit category indicate a fusion. Proteins within brackets indicate fragmented peptides that make up a functional subunit. A question mark at the end of an accession number indicates a possible homologue based on cluster analyses. Genes that purportedly function within the same complex based on cluster analysis share the same shading. Organismal names are arranged alphabetically.

Organism name	**Putative BioY complex components**			**Putative ThiW complex components**		
	S subunit	**T subunit**	**A subunit**	**S subunit**	**T subunit**	**A subunit**
Agrobacterium tumefaciens str. C58	~~159184449~~	~~15888093~~	~~159184450~~	-	-	-
Archaeoglobus fulgidus DSM 4304	~~11498769~~	~~11498771?~~	~~**11498770**~~	-	-	-
Bacillus anthracis str. Ames	30263575, 30264959	-	-	30262625, 30263272	30260332, 30262627, 30263274	30260331, 30260330, **30262626**, **30263273**
Bacillus subtilis subsp. subtilis str. 168	16078101, 16080256	-	-	255767313 (putative PBP=16078389)	255767036, 16078386	16077214, 255767035, **255767312**
Bifidobacterium longum DJO10A	189440492	-	-	189439686, 189439284	189439684, 189438918, 189439283	**189439685**, **189438918**, **189439282**
Bordetella bronchiseptica RB50	~~33602772~~	~~33602773~~	~~33602774~~	-	-	-
Brucella melitensis 16M	17986602, 17987714	-	-	17986923+17986922 (CbiM+CbiN)	17986920	17986919, 17986918
Chlamydia trachomatis 434/Bu	166154570	-	-	-	-	-
Clostridium acetobutylicum ATCC 824	15893504, 15894639, 15896009, 15896696	-	-	-	15896351	15896352, 15896353

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. Continued.

Organism name	Putative BioY complex components			Putative ThiW complex components		
	S subunit	T subunit	A subunit	S subunit	T subunit	A subunit
Clostridium perfringens ATCC 13124	-	~~110800175~~	~~110800816~~, ~~110801175~~	110800458	110800706	**110798697**
Corynebacterium diphtheriae NCTC 13129	~~38234033~~	~~38234035~~	~~38234034~~	-	38233550	**38233549**
Corynebacterium glutamicum R	~~145295863~~	~~145295865~~	~~145295864~~	145295220, 145294682, 145296815	145295222, 145294684, 145296814	**145295221**, **145294683**, **145296813**
Enterococcus faecalis V583	29377530	-	-	29376663, 29377244, 29375194, 29377251	29376661, 29374883, 29377242?, 29375192	**29376662**, 29374882, 29374881, **29377243**, **29375193**,
Escherichia coli O157:H7 str. Sakai	-	-	-	-	-	**15830042**
Haloarcula marismortui ATCC 43049	-	-	-	55378963	55378961	pseudogene downstream of 55378961?
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	-	-	-	-	-	**152969331**
Lactobacillus acidophilus NCFM	-	-	-	58337225	58336659, 58336440, 58337223	58336657, 58336658, **58336441**, **58337224**
Lactobacillus casei	191639160	-	-	191637132, 191637128, 191637253	191637130, 191637255, 191639398	**191637129**, **191637254**, 191639399, 191639400

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. Continued.

Organism name	Putative BioY complex components			Putative ThiW complex components		
	S subunit	T subunit	A subunit	S subunit	T subunit	A subunit
Lactobacillus casei ATCC 334	116495673	-	-	116493866, 116493862	116493864, 116495916	**116493863**, 116495917, 116495918
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	116513290	-	-	116514603	116514601, 116513569, 116513337	**116514602**, 116513568, 116513567, **116513338**
Lactobacillus gasseri ATCC 33323	-	-	-	116630121, 116628751	116630123, 116628992, 116628749	**116630122**, 116628991, 116628990, **116628750**
Lactobacillus sakei subsp. sakei 23K	81428438	-	-	81429447, 81428442, 81428704	81429348, 81429449	81429349, 81429350, **81429448**
Lactococcus lactis subsp. cremoris SK11	116511170, 116511172, 116512642		-	116511180	116511182, 116511106	**116511181**, 116511105, 116511104
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	116617430, 116617633, ~~116617737~~, ~~116617738~~	-	-	116618045, 116617880, 116618639, 116619047	116617360, 116617877, 116618637, 116619045	116617359, 116617358, **116617878**, **116618638**, **116619046**
Listeria monocytogenes str. 4b F2365	46906843	-	-	-	46908771, 46908806	46908772, 46908773
Methanobrevibacter smithii ATCC 35061	148642489	-	-	148642832	148642831	**148642830**

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. Continued.

Organism name	Putative BioY complex components			Putative ThiW complex components		
	S subunit	T subunit	A subunit	S subunit	T subunit	A subunit
Methanococcus vannielii SB	-	-	-	-	-	**150400111**
Methano-corpusculum labreanum Z	~~124484960~~	~~124484963?~~	~~124484961~~, ~~124484962~~	124484855	124484856, 124485087	**124484855**, **124485086**, **124485367**, **124485552**
Methanoculleus marisnigri JR1	~~126179591~~	~~126179588~~	~~126179589~~	126178546	126178548	**126178547**, **126178889**
Methanosarcina barkeri str. Fusaro	~~73668131~~	~~73668128~~	~~73668129~~, ~~73668130~~	73670290, 73670289	73669274, 73670287, 73670985	**73669275**, **73670288**, **73670984**
Methanosarcina mazei Go1	~~21227138~~	~~21227141~~	~~21227139~~, ~~21227140~~	21228488	21228097, 21228490, 21229119	**21228098**, **21228489**, **21229118**
Methanosphaera stadtmanae DSM 3091	84490142	-	-	84489905	84489907?	**84489422**, **84489906**
Methanospirillum hungatei JF-1	~~88604408~~	~~88604411?~~	~~88604409~~, ~~88604410~~	88602882	88602880	**88602881**, **88601738**
Mycobacterium leprae TN	-	-	-	15827376	15827377	**15827376**
Mycobacterium tuberculosis F11	-	-	-	148823526	148823525	**148823526**
Mycoplasma pneumoniae M129	-	-	-	-	13507934, 13508170	13507933, 13507932, 13508171, 13508172
Myxococcus xanthus DK 1622	108761790	-	-	-	-	-

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. Continued.

Organism name	Putative BioY complex components			Putative ThiW complex components		
	S subunit	T subunit	A subunit	S subunit	T subunit	A subunit
Oenococcus oeni PSU-1	116491570	-	-	116490288, 116490289, 116491471	116490688, 116490871, 116491110, 116490286, 116491468, 116491517	116490687, 116490686, 116490872, 116490873, 116491109, 116491108, **116490287**, **116491469**, **116491518**
Pasteurella multocida subsp. multocida str. Pm70	-	-	-	-	-	**15602755**
Pediococcus pentosaceus ATCC 25745	~~116493320~~, ~~116493321~~, 116493365	-	-	-	116493133	116493134, 116493135
Pyrococcus abyssi GE5	14520373	-	-	-	14520347	14520348
Rickettsia typhi str. Wilmington	51473520	-	-	-	-	-
Salmonella enterica subsp. arizonae serovar 62:z4,z23:-- str. RSK2980	-	-	-	161506375	161506374	**161504071**, 161506373, 161506372
Serratia proteamaculans 568	-	-	-	-	-	**157369538**

Table 3A: Genomic cluster analysis of BioY and ThiW transporter homologues. Continued.

Organism name	Putative BioY complex components			Putative ThiW complex components		
	S subunit	T subunit	A subunit	S subunit	T subunit	A subunit
Shigella dysenteriae Sd197	-	-	-	82778308	82778307	82778304, [82778305, 82778306], **82776034**
Staphylococcus aureus subsp. aureus NCTC 8325	88196197	-	-	88195004, 88194775, 88196638	88196132. 88194773, 88196636	88196133, 88196134, **88194774, 88196637**
Streptococcus pneumoniae D39	116515429	-	-	~~116517054~~, 116516704, 116516079, 116516407, 116515998	~~116516131~~, 116516329?,116516162, 116516756	~~**116516582**~~, **116516786**, 116517082, 116515987, **116516811**
Streptococcus pyogenes M1 GAS	15674405	-	-	15675628	15675926, 15675627	15675927, 15675928, **15675626**
Streptococcus thermophilus LMG 18311	55820751, 55821311	-	-	[55820401, 55820402]	55821974, 55820406	55821976, 55821975, 55820403, 55820404
Sulfolobus solfataricus P2	-	-	-	-	15898688, 15898823	**15897117**, **15897231**, **15898689**, **15898822**
Treponema pallidum subsp. pallidum SS14	~~189025461~~	~~189025459~~	~~189025460~~	-	-	-
Ureaplasma parvum serovar 3 str. ATCC 27815	-	-	-	-	170762077	170762231, 170762407

Table 3B: Genomic cluster analysis of BioY and ThiW transporter homologues. The numbers represent the tally of all the components in table 3A. The numbers are used to construct the Venn diagram in figure 8.

	Putative BioY complex components			**Putative ThiW complex components**		
	S subunit	**T subunit**	**A subunit**	**S subunit**	**T subunit**	**A subunit**
Total # of proteins recorded	49	10	17	56	81	124
Total # of organisms represented for each transporter	37			45		
Total # of organisms containing S subunit	36			32		
Total # of organisms containing T subunit		10			39	
Total # of organisms containing A subunit			12			44
Operon structure analysis **Legend: Number of operons found in our analysis x (subunit composition in the operon)** **S = S-subunit, T = T-subunit, A = A-subunit**						
S alone	33x(S), 2x(S+S)			5x(S)		
T alone		0			1x(T)	
A alone			0			14x(A)
S and T in various combination	1x(S+T)			1x(S+T)		
T and A in various combination	1x(T+A+A)			24x(T+A+A), 12x(T+AA), 1x(T+A)		
S and A in various combination	1x(S+AA), 2x(S+A+A)			1x(S+S+AA), 2x(S+AA)		
S, T, and A in various combination	6x(S+T+A), 2x(S+T+A+A)			34x(S+T+AA), 4x(S+S+T+AA), 4x(S+T+A+A)		

Table 4: Comparisons between transporters of different specificities in the ECF family. *Panel A* shows comparisons among substrate-recognition (S) subunits, and *panel B* shows comparisons among putative ATPase-transducer (T) subunits. Cells with darker shading represent closer sequence similarity than the ones with lighter shading. The homologues were collected from the SEED database (http://seed-viewer.theseed.org/), and the comparisons were conducted with the GAP and TreeView programs (Zhai et al., 2002). Values are expressed in standard deviations (SD). Based on the superfamily principle, the data demonstrate that all S subunits tabulated are homologous as are all T subunits.

A) S subunits

SEED	BioY	CbiM	NikM	QrtT	MtsT	YkoE	HtsT	CbrT	MtaT
BioY		8.8	9.2	14.1	9.1	7.4	10.4	11.0	8.0
CbiM			59.5	10.1	7.1	6.7	6.5	10.3	4.5
NikM				11.5	7.5	7.6	10.5	10.4	6.8
QrtT					14.6	8.4	8.2	14.1	7.0
MtsT						7.3	10.6	20.8	12.1
YkoE							12.6	8.9	4.6
HtsT								8.2	4.9
CbrT									5.3
MtaT									

B) T subunits

SEED	BioN	CbiQ	NikQ	QrtU	MtsV	YkoC	HtsU	CbrV	MtaV
BioN		27.2	34.9	26.2	43.4	30.8	21.4	16.3	35.9
CbiQ			41.3	21.4	21.9	20.8	13.9	11.6	14.0
NikQ				20.9	27.0	25.0	18.0	20.2	22.2
QrtU					42.7	29.9	17.8	24.5	42.1
MtsV						37.7	27.5	29.3	40.9
YkoC							28.7	25.6	43.0
HtsU								13.6	27.8
CbrV									33.8
MtaV									

Table 5: Collection of ECF energizing components. The gi numbers of energizing components were obtained from the SEED database (http://seed-viewer.theseed.org/). These protein sequences were used to generate the phylogenetic tree in figure 10.

Organismal name	Organismal type	T	5' ATPase	3' ATPase
Bacillus cereus ATCC 10987	Firmicutes	42779222/Bce	42779220/Bce1	42779221/Bce2
Bacillus halodurans C-125	Firmicutes	15612729/Bha	15612727/Bha1	15612728/Bha2
Bacillus subtilis subsp. subtilis str. 168	Firmicutes	221307960/Bsu	1644202/Bsu1	221307959/Bsu2
Clostridium acetobutylicum ATCC 824	Firmicutes	15896351/Cac	15896353/Cac1	15896352/Cac2
Clostridium botulinum A str. ATCC 3502	Firmicutes	148381388/Cbo	148381390/Cbo1	148381389/Cbo2
Clostridium difficile QCD-32g58	Firmicutes	145956420/Cdi	145956418/Cdi1	145956419/Cdi2
Clostridium perfringens str. 13	Firmicutes	18311354/Cpe	18311356/Cpe1	18311355/Cpe2
Clostridium tetani E88	Firmicutes	28212150/Cte	28212152/Cte1	28212151/Cte2
Clostridium thermocellum ATCC 27405	Firmicutes	125975418/Cth	125975415/Cth1	125975417/Cth2
Enterococcus faecalis V583	Firmicutes	29374883/Efa	29374881/Efa1	29374882/Efa2
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	Firmicutes	104773572/Lde	104773570/Lde1	104773571/Lde2
Lactobacillus gasseri ATCC 33323	Firmicutes	116628992/Lga	116628990/Lga1	116628991/Lga2
Lactobacillus johnsonii NCC 533	Firmicutes	42518459/Ljo	42518457/Ljo1	42518458/Ljo2
Lactococcus lactis subsp. lactis Il1403	Firmicutes	15672261/Lla	15672259/Lla1	15672260/Lla2
Listeria monocytogenes EGD-e	Firmicutes	16804637/Lmo	16804639/Lmo1	16804638/Lmo2
Lactobacillus plantarum WCFS1	Firmicutes	28377869/Lpl	28377867/Lpl1	28377868/Lpl2

Table 5: Collection of ECF energizing components. Continued.

Organismal name	Organismal type	T	5' ATPase	3' ATPase
Mesoplasma florum L1	Firmicutes	50364969/Mfl	50364967/Mfl1	50364968/Mfl2
Mycoplasma pneumoniae M129	Firmicutes	13507934/MpnA & 13508170/MpnB	13507932/MpnA1 & 13508172/MpnB1	13507933/MpnA2 & 13508171/MpnB2
Moorella thermoacetica ATCC 39073	Firmicutes	83591243/Mth	83591245/Mth1	83591244/Mth2
Oceanobacillus iheyensis HTE831	Firmicutes	23097604/Oih	23097602/Oih1	23097603/Oih2
Onion yellows phytoplasma OY-M	Firmicutes	39938666/Oni	39938665/Oni1	85057820/Oni2
Pyrococcus furiosus DSM 3638	Euryarchaeota	18976439/Pfu	18976440/Pfu1	
Pediococcus pentosaceus ATCC 25745	Firmicutes	116493133/Ppe	116493135/Ppe1	116493134/Ppe2
Rubrobacter xylanophilus DSM 9941	Actinobacteria	108804943/Rxy	108804945/Rxy1	108804944/Rxy2
Streptococcus agalactiae 2603V/R	Firmicutes	22538283/Sag	22538285/Sag1	22538284/Sag2
Staphylococcus aureus subsp. aureus Mu50	Firmicutes	15925210/Sau	15925212/Sau1	15925211/Sau2
Spiroplasma kunkelii CR2-3x	Firmicutes	34849383/Sku	56748703/Sku1	56748702/Sku2
Streptococcus mutans UA159	Firmicutes	24380476/Smu	24380478/Smu1	24380477/Smu2
Streptococcus pneumoniae TIGR4	Firmicutes	15902023/Spn	15902025/Spn1	15902024/Spn2
Streptococcus pyogenes M1 GAS	Firmicutes	15675926/Spy	15675928/Spy1	15675927/Spy2
Symbiobacterium thermophilum IAM 14863	Firmicutes	51894178/Sth	51894180/Sth1	51894179/Sth2
Thermoanaerobacter tengcongensis MB4	Firmicutes	20808628/Tte	20808630/Tte1	20808629/Tte2

Table 6: Collection of S subunit homologues from the type-I/II ECF and more traditional ABC transport systems. Under the 'TC #' column, dark and light shades indicates type-II and type-I ECF transporters, respectively, while ABC2 transporters are unshaded. Under the 'Query UniProt accession #' column, transmembrane subunits with degrees of similarity insufficient for linkage to either the MalF or the MalG subunits are indicated in grey; homodimeric transmembrane complexes are underlined. The sequences were obtained from the TCDB database and were arranged according to TC#'s in the table. The protein sequences were used to generate the superfamily tree in figure 11.

TC #	Query UniProt accession #
2.A.88.1	Q6GUB0
2.A.88.2	Q2KUS5
2.A.88.3	O32074
2.A.88.4	O07515
2.A.88.5	Q99Z31
2.A.88.6	Q38XE8
2.A.88.7	Q6YQR5
3.A.1.1	P02916
3.A.1.2	P0AGI1
3.A.1.3	P0AEQ6
3.A.1.4	P0AEX7
3.A.1.5	P08005
3.A.1.6	Q93KD5
3.A.1.7	P07654
3.A.1.8	P0AF01
3.A.1.9	P16683
3.A.1.10	P21409
3.A.1.11	P0AFK4
3.A.1.12	P14176
3.A.1.13	P06609
3.A.1.14	P06972
3.A.1.15	Q55282
3.A.1.16	P38044
3.A.1.17	Q47539
3.A.1.18	Q05594
3.A.1.19	P31549
3.A.1.20	O54371
3.A.1.22	Q9S411
3.A.1.23	Q79CJ1
3.A.1.24	P31547
3.A.1.26	A9WGB0
3.A.1.27	A4PCH7
3.A.1.28	Q8XGV9
3.A.1.29	Q74I63
3.A.1.30	O34738
3.A.1.31	Q8G6E7
3.A.1.32	Q9KXJ5
3.A.1.33	Q8R9M1
3.A.1.34	Q99ZY4

Table 7: Distribution of the coenzyme riboswitch class in major bacterial phyla. The data for individual riboswitch are normalized against the sizes of the genomes in megabases (Mb). Numbers indicate number of riboswitches per Mb. Darker shades indicate higher riboswitch density (refer to legend above column headings).

Legend:	>=0.75	0.749~ 0.50	0.499~ 0.25	0.249> 0	Units are in sites per megabase pair							
Phylum (class)	TPP	Cobalamin	SAM	FMN	MOCO_RNA_motif	SAM_alpha	SAH_riboswitch	SAM-IV	SAM-SAH	SMK box	SAM-Chlorobi	THF
Firmicutes	0.962	0.335	1.059	0.535	0.029	0	0	0	0	0.063	0	0.112
Actinobacteria	0.517	0.378	0.014	0.042	0	0	0.084	0.182	0	0	0	0
Cyanobacteria	0.234	0.318	0.033	0	0	0	0	0	0	0	0	0
Proteobacteria	0.438	0.510	0.007	0.175	0.096	0.072	0.039	0	0.024	0	0	0
(alpha)	0.413	0.722	0	0.129	0.013	0.251	0	0	0.084	0	0	0
(beta)	0.246	0.461	0	0.154	0	0	0.144	0	0	0	0	0
(gamma)	0.504	0.378	0	0.205	0.197	0	0.028	0	0	0	0	0
(delta)	0.603	0.660	0.115	0.230	0	0	0	0	0	0	0	0
Chlorobi	0.377	1.165	0.480	0.069	0.034	0	0	0	0	0	0.377	0
Bacteroidetes	0.336	1.151	0	0	0	0	0	0	0	0	0	0
Deinococcus-Thermus	0.781	0.938	0.703	0.391	0.078	0	0	0	0	0	0	0
Chloroflexi	0.503	0.467	0.539	0.288	0	0	0	0	0	0	0	0
Thermotogales	0.736	0.920	0.046	0.368	0.092	0	0	0	0	0	0	0
Phyla other than firmicutes	0.436	0.568	0.056	0.147	0.068	0.048	0.033	0.016	0.016	0	0.013	0

Table 8: Distribution of riboswitch classes responsible for amino acid metabolism in major bacterial phyla. The data for individual riboswitches are normalized against the sizes of the genomes in megabases (Mb). Darker shades indicate higher riboswitch density; refer to table 7 for format of presentation.

Phylum (class)	Glycine	Lysine	glnA	Trp_leader	His_leader	Leu_leader	Thr_leader	T-boxes
Firmicutes	0.457	0.515	0	0	0	0	0	5.195
Actinobacteria	0.294	0	0	0.028	0	0	0	0.210
Cyanobacteria	0	0	0.217	0	0	0	0	0
Proteobacteria	0.325	0.133	0	0.079	0.087	0.079	0.083	0
(alpha)	0.387	0	0	0.032	0	0	0	0
(beta)	0.431	0	0	0	0	0	0	0
(gamma)	0.268	0.283	0	0.150	0.185	0.169	0.177	0
(delta)	0.172	0	0	0	0	0	0	0
Chlorobi	0	0	0	0	0	0	0	0
Bacteroidetes	0	0	0	0	0	0	0	0
Deinococcus-Thermus	0	0	0	0	0	0	0	0.781
Chloroflexi	0.108	0	0	0	0	0	0	0.899
Thermotogales	0	0.414	0	0.138	0.046	0	0	0
Phyla other than firmicutes	0.244	0.099	0.016	0.058	0.058	0.052	0.055	0.061

Table 9: Distribution of riboswitch classes responsible for the biogenesis of ribosomal subunits in major bacterial phyla. The data for individual riboswitches are normalized against the sizes of the genomes in megabases (Mb). Darker shades indicate higher riboswitch density; refer to table 7 for format of presentation.

Phylum (class)	S15	L10_leader	L20_leader	L21_leader	L19_leader	L13_leader
Firmicutes	0.131	0.326	0.316	0.301	0.296	0.209
Actinobacteria	0.098	0.056	0	0	0	0
Cyanobacteria	0	0.184	0	0.017	0	0
Proteobacteria	0.118	0.011	0.013	0	0	0
(alpha)	0.058	0	0	0	0	0
(beta)	0	0	0	0	0	0
(gamma)	0.217	0.024	0.028	0	0	0
(delta)	0	0	0	0	0	0
Chlorobi	0	0	0	0	0	0
Bacteroidetes	0	0	0	0	0	0
Deinococcus-Thermus	0	0.313	0	0	0	0
Chloroflexi	0	0.180	0	0	0	0
Thermotogales	0	0.506	0.460	0.138	0	0
Phyla other than firmicutes	0.086	0.050	0.021	0.005	0	0

Table 10: Distribution of nucleotide derivative riboswitch classes in major bacterial phyla. The data for individual riboswitches are normalized against the sizes of the genomes in megabases (Mb). Darker shades indicate higher riboswitch density; refer to table 7 for format of presentation.

Phylum (class)	Purine	PyrR	GEMM_RNA_motif	PreQ1	preQ1-II
Firmicutes	0.612	0.933	0.199	0.301	0.083
Actinobacteria	0	0.196	0	0	0
Cyanobacteria	0	0	0.100	0	0
Proteobacteria	0.024	0	0.078	0.018	0
(alpha)	0	0	0	0.006	0
(beta)	0	0	0.010	0	0
(gamma)	0.051	0	0.142	0.035	0
(delta)	0	0	0.144	0	0
Chlorobi	0	0	0	0	0
Bacteroidetes	0	0	0	0	0
Deinococcus-Thermus	0	0.156	0.234	0	0
Chloroflexi	0	0.108	0	0	0
Thermotogales	0.092	0	0	0	0
Phyla other than firmicutes	0.018	0.023	0.062	0.012	0

Table 11: Distribution of ion/sugar riboswitches in major bacterial phyla. The data for individual riboswitches are normalized against the sizes of the genomes in megabases (Mb). Darker shades indicate higher riboswitch density; refer to table 7 for format of presentation.

Phylum (class)	Mg_sensor	ykoK	glmS
Firmicutes	0	0.156	0.175
Actinobacteria	0	0.140	0
Cyanobacteria	0	0	0
Proteobacteria	0.009	0.011	0
(alpha)	0	0	0
(beta)	0	0	0
(gamma)	0.020	0.024	0
(delta)	0	0	0
Chlorobi	0	0	0
Bacteroidetes	0	0	0
Deinococcus-Thermus	0	0	0.234
Chloroflexi	0	0	0.180
Thermotogales	0	0	0
Phyla other than firmicutes	0.006	0.019	0.010

Table 12: Distribution of functionally uncharacterized riboswitches in major bacterial phyla. The data for individual riboswitches are normalized against the sizes of the genomes in megabases (Mb). Darker shades indicate higher riboswitch density; refer to table 7 for format of presentation.

Phylum (class)	yybP - ykoY	ydaO - yuaA	mini-ykkC	ykkC - yxkD	ylbH	sucA	serC	speF	ybhL
Firmicutes	0.467	0.175	0	0.121	0.049	0	0	0	0
Actinobacteria	0.322	0.196	0	0.042	0	0	0	0	0
Cyanobacteria	0.067	0.134	0.017	0.100	0	0	0	0	0
Proteobacteria	0.199	0.002	0.100	0.041	0	0.026	0.059	0.022	0.015
(alpha)	0.052	0	0.090	0.052	0	0	0.206	0.077	0.052
(beta)	0.205	0	0.113	0.072	0	0.144	0	0	0
(gamma)	0.287	0	0.110	0.028	0	0	0	0	0
(delta)	0.201	0.029	0.029	0	0	0	0	0	0
Chlorobi	0	0	0	0	0	0	0	0	0
Bacteroidetes	0	0	0	0	0	0	0	0	0
Deinococcus-Thermus	0.391	0	0	0	0	0	0	0	0
Chloroflexi	0.180	0	0	0	0	0	0	0	0
Thermotogales	0.046	0	0	0	0	0	0	0	0
Phyla other than firmicutes	0.178	0.028	0.067	0.038	0	0.017	0.039	0.015	0.010

Table 13: Combined riboswitch distribution across fully sequenced microbial genomes. Refer to table 1 for the genomes used in this table.

Taxonomic collection	Genomes	Total sites	Total per genome	Total genome size (Mb)	Average riboswitch density per Mb
Bacteroidaceae	11	84	7.64	56.5	1.49
Bacillales	11	656	59.64	47.2	13.91
Staphylococcus	7	294	42.00	17.6	16.68
Streptococcaceae	15	397	26.47	32.0	12.41
Lactobacillaceae	15	580	38.67	32.9	17.61
Clostridiaceae	20	977	48.85	76.1	12.85
Chloroflexi	5	96	19.20	27.8	3.45
Desulfovibrionales	10	76	7.60	34.8	2.18
Cyanobacteria	14	85	6.07	59.8	1.42
Corynebacteriaceae	8	69	8.63	21.5	3.21
Mycobacteriaceae	9	131	14.56	50.0	2.62
Thermotogales	11	87	7.91	21.8	4.00
Deinococcus-Thermus	5	64	12.80	12.8	5.00
Chlorobiales	11	73	6.64	29.2	2.50
Rhizobiales	15	191	12.73	75.9	2.52
Pasteurellales	9	112	12.44	19.6	5.72
Pseudomonadaceae	8	91	11.38	46.8	1.94
Vibrionales	10	202	20.20	51.6	3.92
Ralstonia	6	66	11.00	34.1	1.93
Shewanella	16	291	18.19	79.5	3.66
Burkholderia	8	127	15.88	63.4	2.00
Caulobacterales	4	36	9.00	18.2	1.98
Rhodobacterales	15	180	12.00	61.1	2.95
Enterobacteriales	12	195	18.19	56.5	3.45
Summation for major phyla (class)					
Firmicutes	68	2904	42.71	205.77	14.11
Actinobacteria	17	200	11.76	71.52	2.80
Proteobacteria	113	1567	13.87	541.48	2.89
(alpha)	34	407	11.97	155.15	2.62
(beta)	14	193	13.79	97.52	1.98
(gamma)	55	891	16.20	253.97	3.51
(delta)	10	76	7.60	34.8	2.18
Phyla other than firmicutes	187	2256	12.064	820.84	2.7484

Table 14: Proportion of riboswitches for individual effector classes in each bacterial phylum. Key: A=coenzyme, B=amino acid, C=ribosomal subunit, D=nucleotide derivative, E=ion/sugar, F=putative riboswitch; letter followed by a "'" denotes the count for that particular class of riboswitches divided by the total riboswitch count in that phylum, expressed in percentage (for instance, A' means the percentage of coenzyme riboswitches from the total riboswitch count. The riboswitch class with the highest proportion in each phylum is darkly shaded, and the one with the next highest proportion is lightly shaded.

Phylum (class)	Total sites	A	A'	B	B'	C	C'	D	D'	E	E'	F	F'
Firmicutes	2904	637	21.94%	1269	43.70%	325	11.19%	438	15.08%	68	2.34%	167	5.75%
Actinobacteria	200	87	43.50%	38	19.00%	11	5.50%	14	7.00%	10	5.00%	40	20.00%
Cyanobacteria	85	35	41.18%	13	15.29%	12	14.12%	6	7.06%	0	0.00%	19	22.35%
Proteobacteria	1567	737	47.03%	426	27.19%	77	4.91%	65	4.15%	11	0.70%	251	16.02%
(alpha)	407	250	61.43%	65	15.97%	9	2.21%	1	0.25%	0	0.00%	82	20.15%
(beta)	193	98	50.78%	42	21.76%	0	0.00%	1	0.52%	0	0.00%	52	26.94%
(gamma)	891	333	37.37%	313	35.13%	68	7.63%	58	6.51%	11	1.23%	108	12.12%
(delta)	76	56	73.68%	6	7.89%	0	0.00%	5	6.58%	0	0.00%	9	11.84%
Chlorobi	73	73	100.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Bacteroidetes	84	84	100.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Deinococcus-Thermus	64	37	57.81%	10	15.63%	4	6.25%	5	7.81%	3	4.69%	5	7.81%
Chloroflexi	96	50	52.08%	28	29.17%	5	5.21%	3	3.13%	5	5.21%	5	5.21%
Thermotogales	87	47	54.02%	13	14.94%	24	27.59%	2	2.30%	0	0.00%	1	1.15%
Phyla other than firmicutes	2256	1150	50.98%	528	23.40%	133	5.90%	95	4.21%	29	1.29%	321	14.23%

```
ThiW     1 MKTKKLTLTAIFIAINVVLSSIIVIPLGPIKA.APMQHLINVLCAVFVGP 49
           |  : | | |:| |: |||| .  :||  |     :| |  .|  : .||
BioY     1 MTNRNLVLAALFAALMVVLSLMPPVPLPAIPVPVTLQTLGVMLAGIMLGP 50

ThiW    50 WFGLAQAFISSILRMI 65
           | | |   :  :|  |
BioY    51 WRGAAACLLYLVLAAI 66
```

ThiW (gi#73662775; putative TMSs 1 & 2) vs BioY (gi#115422164; putative TMSs 1 & 2). Comparison score = 11.8 SD; % identity = 35.4; % similarity = 47.7; # gaps = 1; # PSI-BLAST iterations = 1.

Figure 1A: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (ThiW/BioY). The GAP and IC programs were used to generate comparison scores expressed in standard deviations (SD) (see Table 2). '|' indicates an identity between residues, ':' indicates a close similarity, and '.' indicates a more distant similarity. Unless noted otherwise, all putative transmembrane segments (indicated by shading of the aligned sequences) were predicted using the HMMTOP algorithm.

```
ThiW     3 KTTLRNLILAALFAAMAVLLSGLSIPVGPTRCFPFQHAINAIAGVLLGPW 52
           :  |: ||: .|| |. |.|. :. |:       |  |.  |  :||
YhaG     2 RMNLKKLIINSLFLAVGVVLNQITPPILFGMKPDFSLAMLFII.ILLNDD 50
                    .         .         .         .         .
ThiW    53 WAGGAALTTSIIRNALGTGTLFAFP.GSIPGAL.......VVGITAKVFK 94
           :     :.| ::   |    .  || | :|  :       .| |  : ||
YhaG    51 YK..TCISTGVVAGLL.AAAVTTFPGGQLPNIIDRIVTTSLVFIALRPFK 97
                    .         .         .         .         .
ThiW    95 DK...KLYAALTEPVGTGIIGAIL..SVYILA..PSIGKEATLWLVMPAF 137
           ||   |::  :|  ||| | |.:   |  |:   |.  |   : .|:||
YhaG    98 DKINDKIHMIITTIVGTIISGSVFLGSALIIVGLPASFKALFITVVLPAT 147
                    .
ThiW   138 LLSSVPGSLLGFAL 151
           ::..: |.::  |.
YhaG   148 IINAIVGTIIFVAV 161
```

ThiW (gi#289523651; putative TMSs 1~5) vs YhaG (gi#15896851; putative TMSs 1~5). Comparison score = 11.0 SD; % identity = 29.0; % similarity = 46.2; # gaps = 8; # PSI-BLAST iterations = 1. The TMSs for the YhaG homologue were predicted with the PredictProtein algorithm.

Figure 1B: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (ThiW/YhaG). See figure 1A for format of presentation.

```
ThiW   35 MAHFINILCSVILGPWYSLLCATLIGVIRMFFMGIPPLALTGAV..FGAF 82
          | | :|:| .: |  | . |  |  |::|.|   :  .        | |
YjbB    1 MKHLLNLLAAIALLVWGTQLVRT..GILRVFGANLRQVLARSISNRFTAA 48
                .         .         .         .         .
ThiW   83 LS..GVSYRVSKGKLICAIVGEVIGTGVIGAILSYPIMTFIWGRTGLTWM 130
          ||  ||.  |  |     ||   :| |.|   |.  :|  :    | . |
YjbB   49 LSGIGVTALVQSGTATALIVSSFVGQGLIALPLALAVM..LGADIGTSLM 96
                .         .         .         .
ThiW  131 FYVPSFIMATLIGGTIAFIFLGAL.......SRTGNLAKIQRSLG 168
            | || :. |   .  ||||| .       |  | | ::   ||
YjbB   97 AVVFSFDLSWL...SPLFIFLGVVLFISRQDSNAGRLGRVLIGLG 138
```

ThiW (gi#168179942; putative TMSs 2~5) vs YjbB (gi#91789695; putative TMSs 1~3). Comparison score = 10.2 SD; % identity = 33.1; % similarity = 42.5; # gaps = 6; # PSI-BLAST iterations = 1.

Figure 1C: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (ThiW/YjbB). See figure 1A for format of presentation.

```
YhaG   64 LSIGNIQPEFVIASFCLAILLVRPNVIQGAIIGALAATLIQFNTSIPGLE 113
          |.:|.     :   |||  ..        | :|       |  | .|| |
YjbB  168 LAVGDKWIIGIFIGFCLTAVVQSSAATTGILIALAGTDQISINIAIPIL. 216
                .         .         .         .         .
YhaG  114 YACDIPAAIIMTLMLMGYLKAFPKDAARKFSVFPLISSFIATVISGMIFA 163
          : |.|   :   :  .|  |      ||| .   || . . |:|   :
YjbB  217 FGCNIGTCVTTLISSIGTSK.....KARKAAFIHLIYNVMGTIIFIPLMG 261
                .
YhaG  164 TIASFFILHSPNTV 177
          |:|   :  .|. |
YjbB  262 TLAKIVVNINPDNV 275
```

YhaG (gi#227507581; putative TMSs 2~4) vs YjbB (gi#242260858; putative TMSs 6~8). Comparison score = 10.3 SD; % identity = 25.0; % similarity = 34.3; # gaps = 2; # PSI-BLAST iterations = 2. The TMSs for the YhaG homologue were predicted with the TMHMM algorithm.

Figure 1D: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (YhaG/YjbB). See figure 1A for format of presentation.

```
BioY   22 MKIQDLTLIALMAALTCILGPMSITLPFTPVPISFTNLVIYFAVMVIGMK 71
          || .:| :..|:||:  :|   .   : |   |      ||. |  .::  |
YhaG    1 MKTKELVIMSLLAAMGAVLHTIFPPIFFGMKPDMM..LVMMFLSIILFPK 48
                   .         .         .         .         .
BioY   72 RGTISYLVYLLIGAVGLPVFSGFSGGLAKLAGPTGGYLVGFIFLALISGF 121
             :  .: |. ||:   . .|| ||  .:       .  ||||||
YhaG   49 VQHV.VVIALVTGAIS.ALTTGFPGG..QIPNMIDKPVTAFIFLALFLSC 94
                   .         .         .         .         .
BioY  122 FVEKFSGNIVMAVIGMVLGTVVTYAFGTIWLCAQMHLT....FVQGLYAG 167
             |    :|:  :   :||:|.   | |.| | : :|     .  |  |
YhaG   95 M..KIKNKVVLTAVLTAIGTIVS...GVIFLSAALLITGLPAALPALLVG 139
                   .         .         .
BioY  168 VIPYLPGDAAKIVIAIIVGSAVKKAVVKARVL 199
          |:  ||      :  : |    .  . :||.:
YhaG  140 VV..LPAAVINTIAMVFVFPIAQSILRRARMI 169
```

BioY (gi#167765719; putative TMSs 1~5) vs YhaG (gi#157691729; putative TMSs 1~5). Comparison score = 9.5 SD; % identity = 26.7; % similarity = 41.2; # gaps = 8; # PSI-BLAST iterations = 1. The TMSs for both proteins were predicted with the PredictProtein algorithm.

Figure 1E: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (BioY/YhaG). See figure 1A for format of presentation.

```
BioY   12 LISLFTALTAIMAYIVIPMPGGLPPITGQSFAVMLAGLLLGAHKGAMS.Q 60
          ||||  | | :: | |  .  |:    | ||  :| |        | |
YjbB    7 LISLAGA.TMLLLYAVRMVRTGIERSYGASFQRLLTGRQSHLQAGMMGLT 55
                   .         .         .         .         .
BioY   61 IIYVLLGMAGMPVFAGGTAGAGVLAGPTG.GFIWGFILGAFVIGKIAEMS 109
          :  ||   | . . | | | .| || |||   : |  ||. .| .|
YjbB   56 LAIVLQSSAAVALLASGFAASGYLAFPTGLAIVLGGDLGSALIIQILSFK 105
                   .         .
BioY  110 KQRSLPVLYLAAVLGGIVAVYT 131
              .|.| |||  || . | |
YjbB  106 LDWLVPML.LAA..GGYLFVKT 124
```

BioY (gi#225181990; putative TMSs 1~5) vs YjbB (gi#254466313; putative TMSs 1~3). The differences in putative TMSs represent possible misprediction.
Comparison score = 9.4 SD; % identity = 36.2; % similarity = 41.4; # gaps = 5; # PSI-BLAST iterations = 1.

Figure 1F: Binary alignments of S subunit homologues of putative members of the ECF transporter sub-superfamily (BioY/YjbB). See figure 1A for format of presentation.

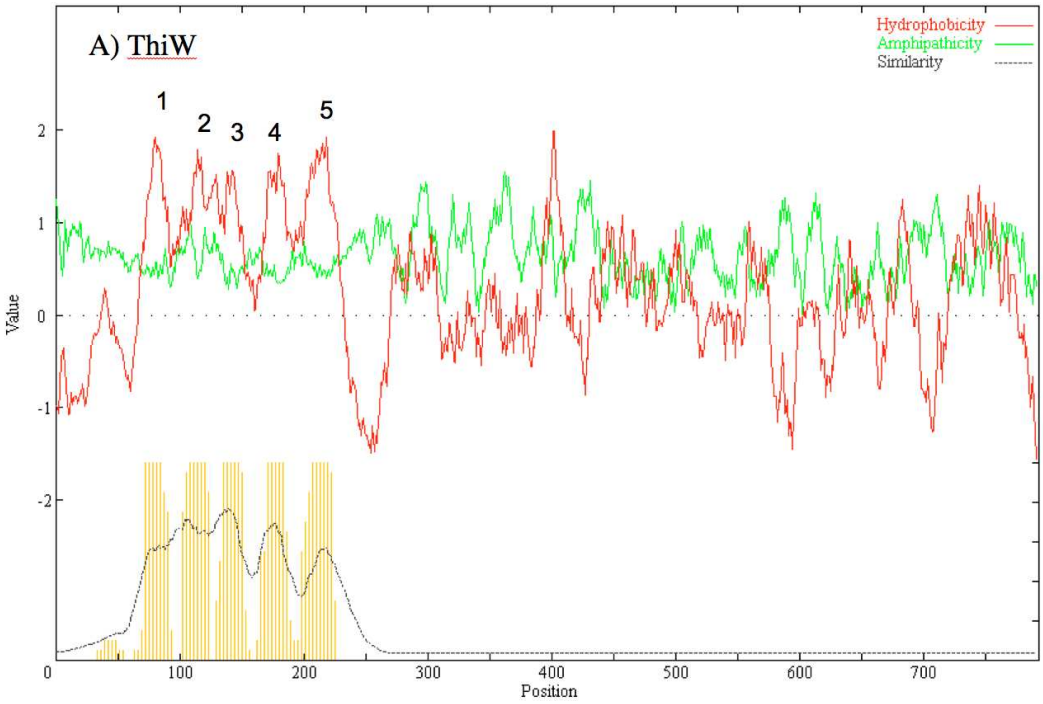


Figure 2A: Average hydropathy, amphipathicity and similarity plots (ThiW). The plots were drawn using a modified AveHAS program (Zhai & Saier, 2001; Yen et al., 2009). The vertical bars at the bottom indicate the positions of the predicted TMSs as estimated using the TMHMM program (Krogh et al., 2001). Conserved TMSs are numbered above the hydrophobicity peaks.

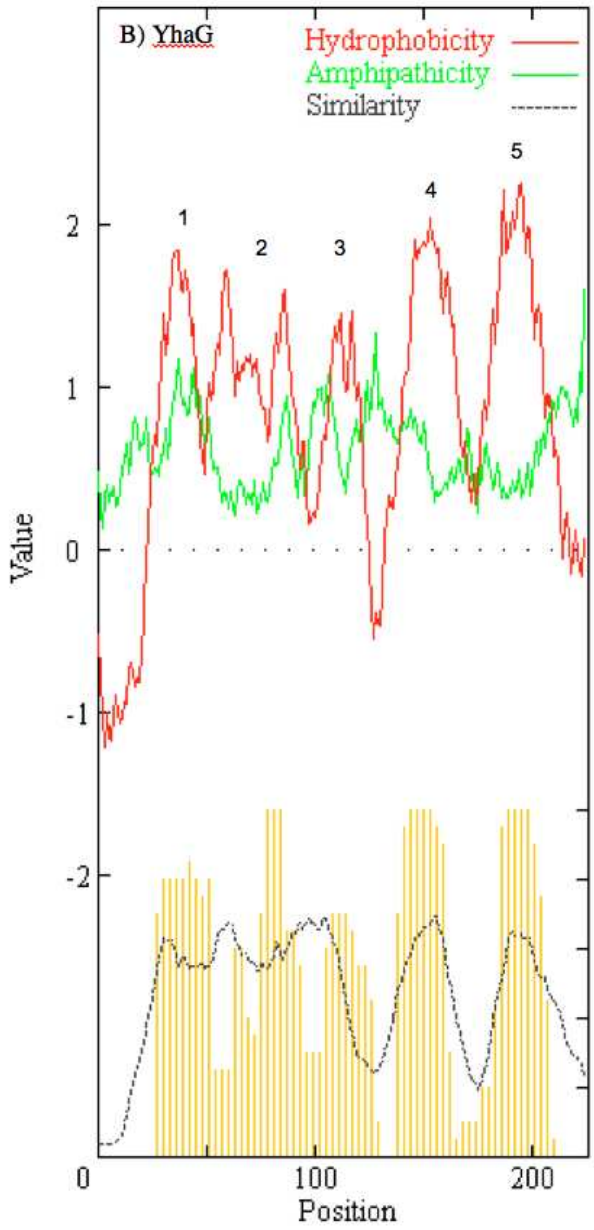


Figure 2B: Average hydropathy, amphipathicity and similarity plots (YhaG). See figure 2A for format of presentation.

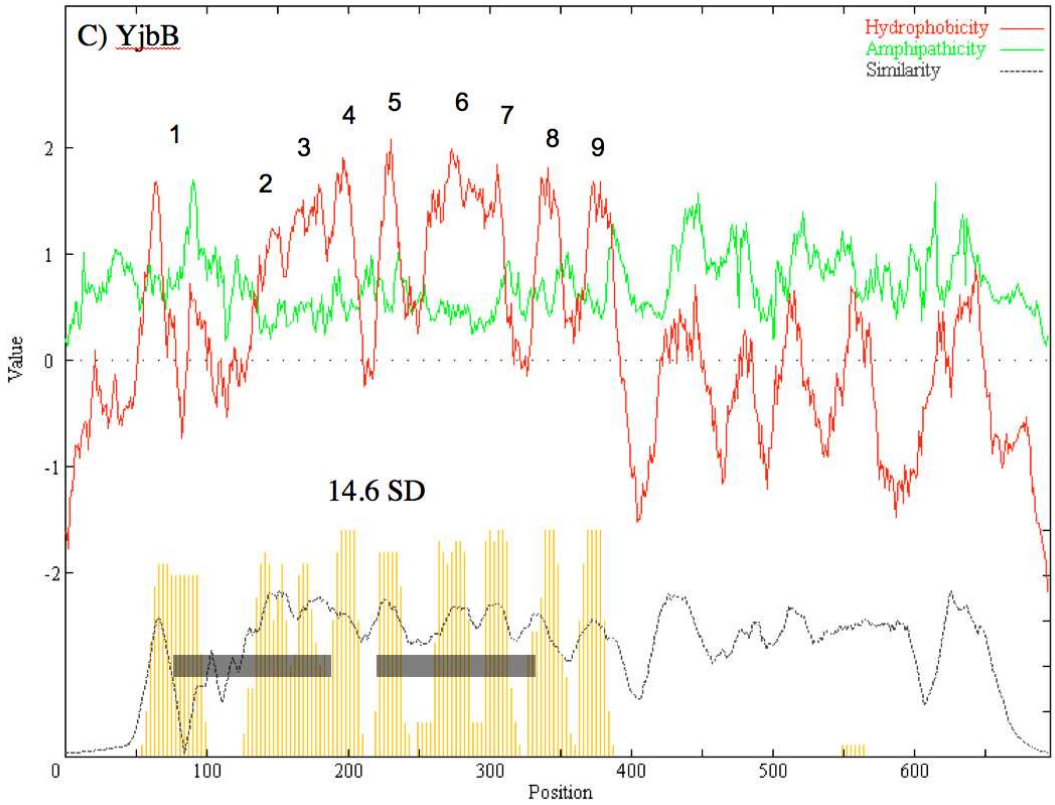


Figure 2C: Average hydropathy, amphipathicity and similarity plots (YjbB). See figure 2A for format of presentation. Horizontal bars at the bottom indicate the regions that contain the best matches using the GAP alignment program, with the number indicating the GAP score in standard deviations.

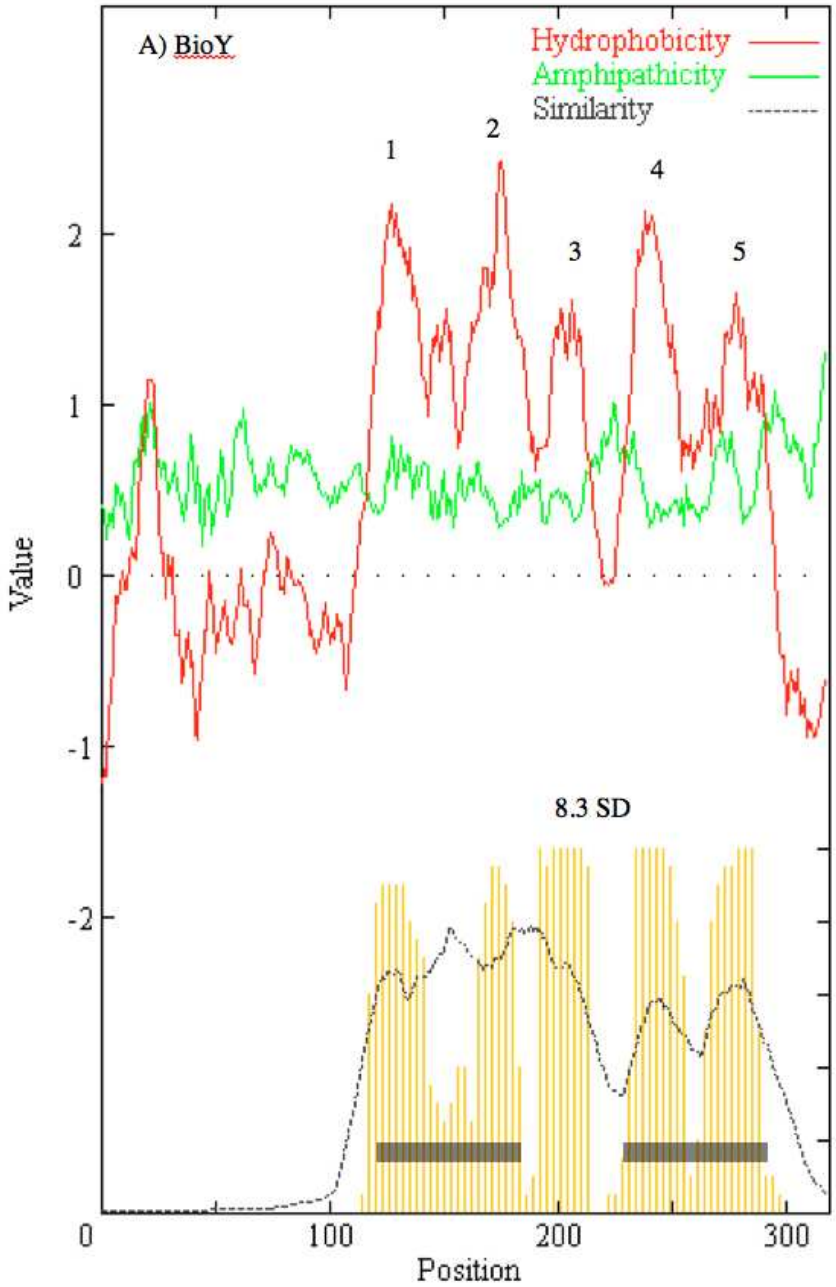


Figure 3A: The AveHAS plot for the S subunits of type-I ECF transporters (BioY). See figure 2A for format of presentation. Horizontal bars at the bottom indicate the regions that contain the best matches using the GAP alignment program, with the number indicating the GAP score in standard deviations.

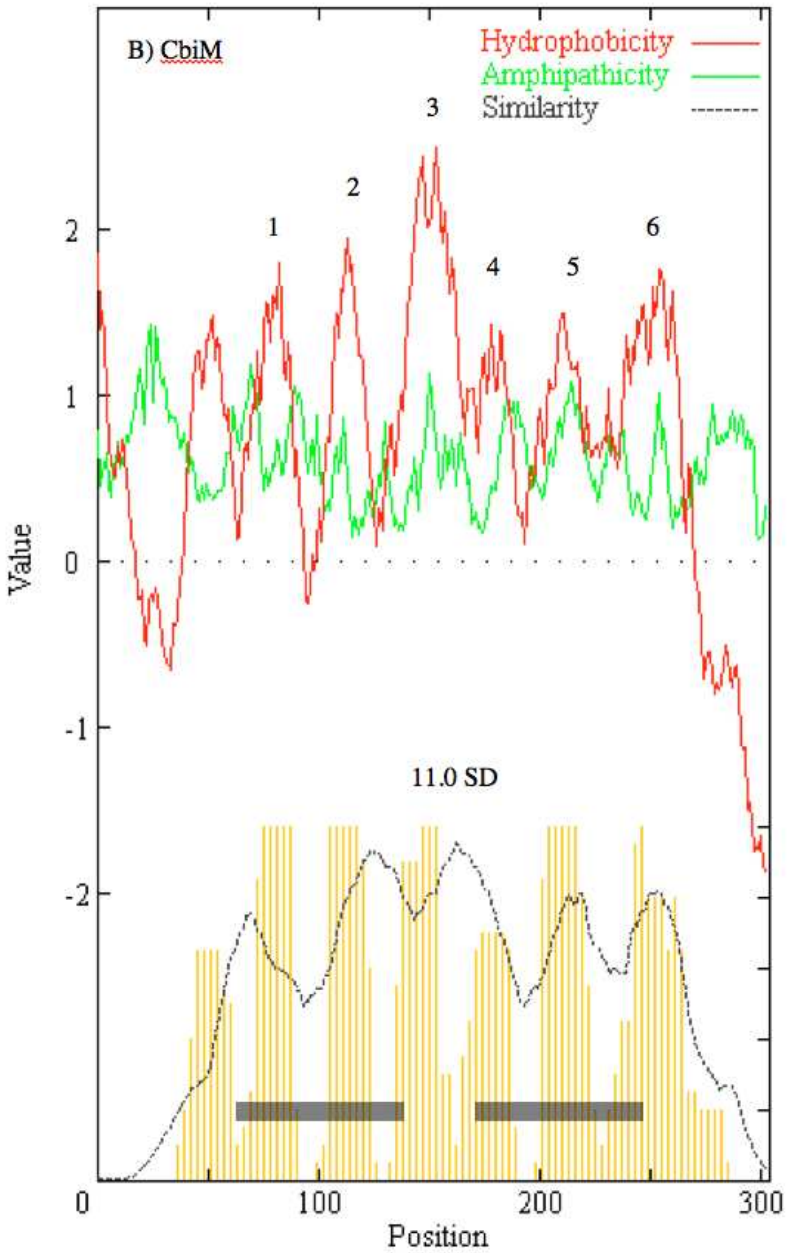


Figure 3B: The AveHAS plot for the S subunits of type-I ECF transporters (CbiM). See figure 2A for format of presentation. Horizontal bars at the bottom indicate the regions that contain the best matches using the GAP alignment program, with the number indicating the GAP score in standard deviations.

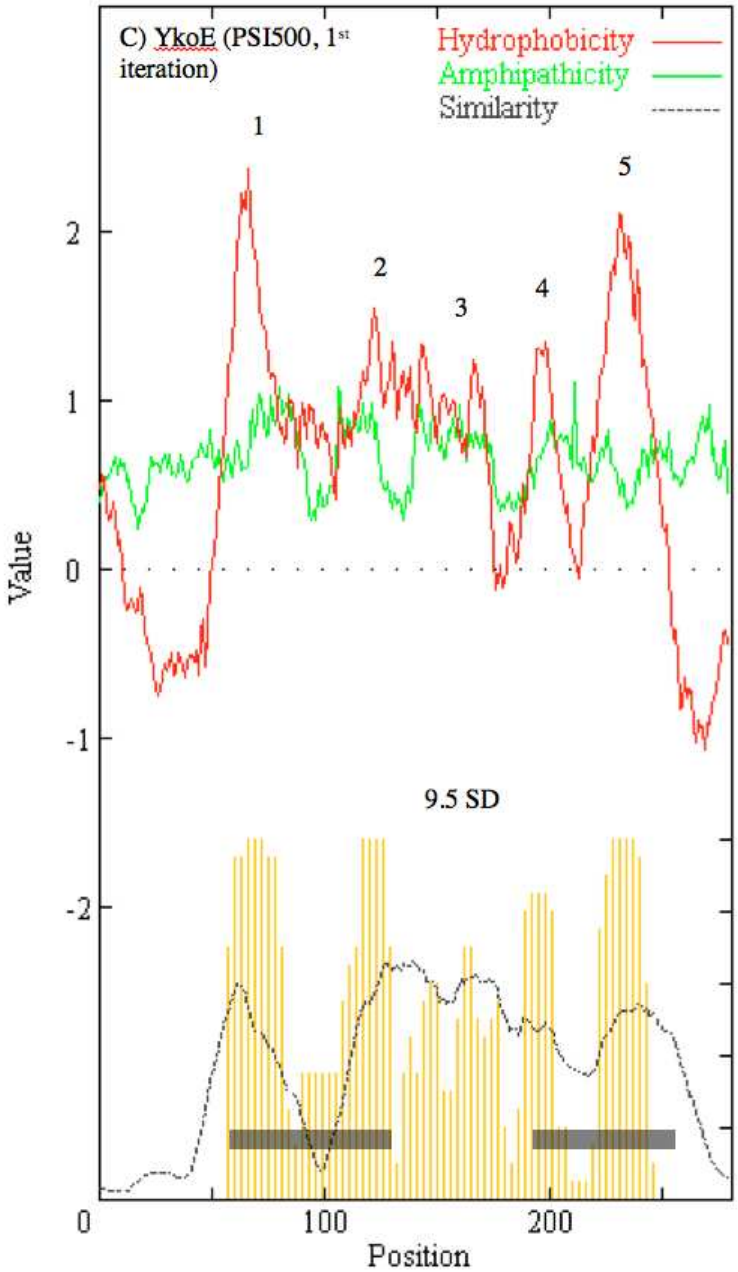


Figure 3C: The AveHAS plot for the S subunits of type-I ECF transporters (YkoE). See figure 2A for format of presentation. Horizontal bars at the bottom indicate the regions that contain the best matches using the GAP alignment program, with the number indicating the GAP score in standard deviations.

```
Cte     20 VSLFAALMAVMGLIPKIDLPFGVPITIQSLGVMLA.GVMLGPWRGMQSMV 68
           ..: | .:  .|.:  : :|: | :|  ||  .||  |   |   ....|
Rru    115 LTILACVLGGVGVLYAVGIPYWVAVTGNSLSAVLAIAVNFLPGDAIKAVV 164

Cte     69 LFLAAVAV..GLPLLS 82
              | || |  | | |
Rru    165 AGLVAVTVRRGYPALS 180
```

Alignment of a portion of the first half (putative TMSs 1 & 2) of a BioY homologue with a portion of the second half (putative TMSs 4 & 5) of another homologue.
This alignment contains the first half of the BioY protein of *Comamonas testosteroni* KF-1 (gi#221067436; Cte) and the second half of the BioY protein of *Rhodospirillum rubrum* ATCC 11170 (gi#83594715; Rru).
Comparison score = 8.3 SD; % identity = 31.7; % similarity = 42.9; # gaps = 2; # PSI-BLAST iterations = 1. The TMSs for Cte were predicted with the SOSUI program.

Figure 4A: Binary alignments of suspected intragenic duplications in S subunits (BioY homologues). The GAP and IC programs were used to generate comparison scores. '|' indicates an identity between residues, ':' indicates a close similarity, and '.' indicates a more distant similarity. Unless noted otherwise, all putative transmembrane segments (indicated by shading of the aligned sequences) were predicted using the HMMTOP algorithm. For the relative locations of TMSs, refer to the AveHAS plots in figure 3A.

```
Glo      3 IMEGFLPVKHAVAWSAASA.PFVAYGIY..SIKKRVAEHPEQRMLLGVAT 49
           :  | |     | |.| | | |||.||||   :  :  |     :   :
Cli    131 LAHGGLTTLGANAFSMAIAGPFVSYGIYRLMVMSKAPEWLAVFLAAAIGD 180
                     .         .
Glo     50 AFTFVLSALKI....PSVTGSCSHPTG 72
             |:|...|.:    |||||  .   |
Cli    181 LMTYVVTSLQLALAFPSVTGGIAASLG 207
```

Alignment of a portion of the first half (putative TMSs 2 & 3) of a CbiM homologue with a portion (putative TMSs 5 & 6) of the second half of another homologue.
This alignment contains the first half of the putative cobalamin transport protein CbiM of *Geobacter lovleyi SZ* (gi#189426704; Glo) and the second half of the putative cobalamin transport protein CbiM of Chlorobium limicola DSM 245 (gi#189346579; Cli).
Comparison score = 11.0 SD; % identity = 34.3; % similarity = 44.3; # gaps = 3; # PSI-BLAST iterations = 1.

Figure 4B: Binary alignments of suspected intragenic duplications in S subunits (CbiM homologues). See figure 4A for format of presentation. For the relative locations of TMSs, refer to the AveHAS plots in figure 3B.

```
Lde     7 EWNVKAIILMALIGIIMGVIYTYVVNPIYNSVELALNLVGLGPLAGQICA 56
          .||:   ||   :  :    | :  || :| . .   |||   :.| : |
Blo   122 KWNIGTTILSGALAGVGCWAYYWATNPGWNGLRVTWYLVG.SIISGVVIA 170
                        .
Lde    57 G..LWYMAAPLAM 67
          |  .||:.  ||.
Blo   171 GVVVWYLSRALAV 183
```

Alignment of a portion of the first half (putative TMSs 1 & 2) of a YkoE homologue with a portion (putative TMSs 4 & 5) of the second half of another homologue.
This alignment contains the first half of the hypothetical protein Ldb0513 of Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 (gi#104773636; Lde) and the second half of the hypothetical protein BL1071 of Bifidobacterium longum NCC2705 (gi#23465640; Blo).
Comparison score = 9.5 SD; % identity = 30.0; % similarity = 43.3; # gaps = 2; # PSI-BLAST iterations = 1.

Figure 4C: Binary alignments of suspected intragenic duplications in S subunits (YkoE homologues). See figure 4A for format of presentation. For the relative locations of TMSs, refer to the AveHAS plots in figure 3C.

```
Bja    12 GVALLLWGLHMVHSGIL.RAFGPDLRRLLGKALGNRVSAFAAGLGLTALL 60
          |. |:|  || .   |      |.|| |:|         | |    || .
Aca   147 GLGLMLLALHQLIEFITPYEDAPNLRLLMGAVSTQPTVAVAVAAVLTWVA 195
                      .          .          .
Bja    61 QSSTATALITSSFTAEGLVSLAAALAIMLGANVGTTL 97
           || |   |:  |  |.|.||   ||:|:.:|||.|.  |
Aca   196 HSSVAVVLLAMSLAAKGVVSPEAAMALVIGANLGSAL 233
```

Alignment of a portion of the first half of a YjbB homologue (putative TMSs 1-3) with a portion of a corresponding second half (putative TMSs 5-7).
The first half of the hypothetical protein bll4444 of Bradyrhizobium japonicum USDA 110 (gi#27379555; Bja) and the second half of the putative Na+/phosphate symporter of Azorhizobium caulinodans ORS 571 (gi#158423914; Aca).
Comparison score = 14.6 SD; % identity = 37.2, % similarity = 44.2; # gaps = 1; # PSI-BLAST iterations = 1. The TMSs for Aca were predicted with the TMHMM algorithm.

Figure 5: Binary alignment of the N-terminal and C-terminal halves of YjbB homologues. See figure 4A for format of presentation. For the relative location of TMSs, refer to the AveHAS plots in Figure 2C.

```
CbiM   46 WWVLIIPFWVIGFQKLTKITRNNPELKLLIALAGAFTFVLSA..LKIPSV 93
          || : |   | :|| |.    | |     : | |||   | |||.     .:  :
CbiQ   92 WWGMQIGHWWLGFDKVGLTTANQTLWRSLAALAATFWFVLNTPFPQLIQI 141
                   .         .         .         .         .
CbiM   94 TGSCSHPTGTGLGVVLFGPSVMTVLGSLVLLFQAVLLAHGGLTTLGANVF 143
              |   |       :.|     :  .:     : ||   |   | : .|
CbiQ  142 LKRCHAPRLLTEQILLTWRFIFILIEEAAAIHQAQSLRFGYI.SLRTGYH 190

CbiM  144 SMAIV 148
          |:|..
CbiQ  191 SLAML 195
```

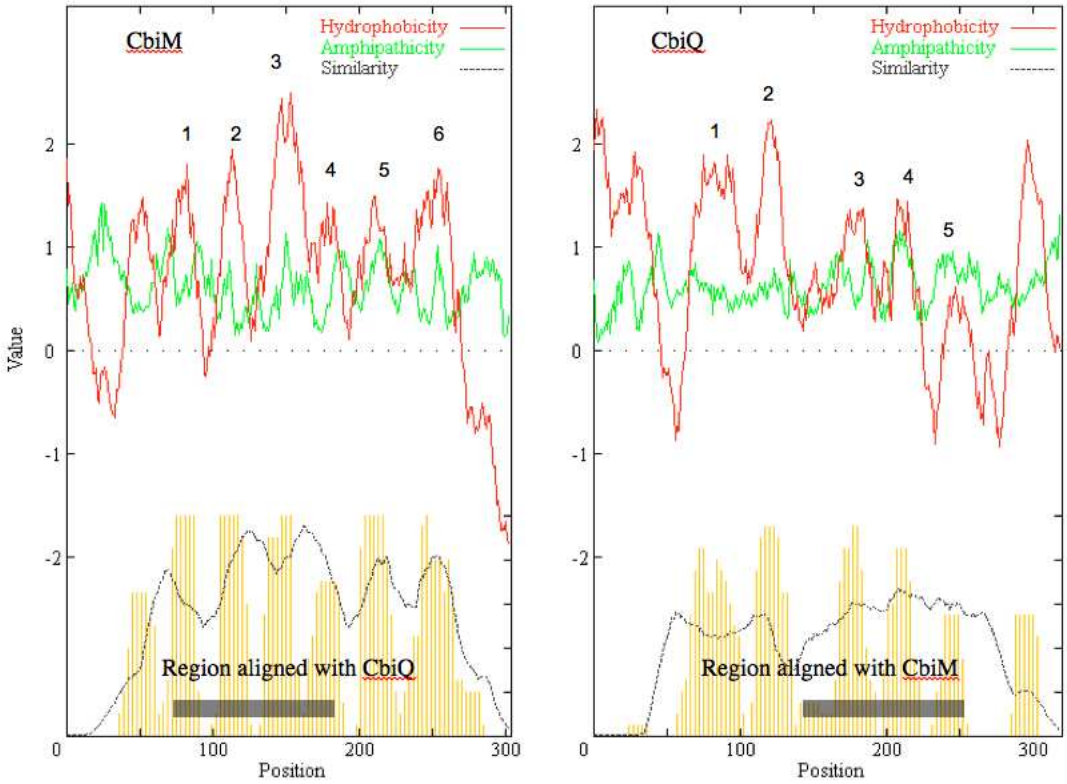
CbiM (S subunit; gi#209528164) vs CbiQ (T subunit; gi#123442928).
Comparison score = 11.9 SD; % identity = 26.5; % similarity = 37.3; # gaps = 2; # PSI-BLAST iterations = 1. The TMSs for CbiQ were predicted with the PredictProtein program.

Figure 6A: Binary alignments of S and T subunits (CbiM/CbiO). See figure 2A for format of presentation. For the relative locations of TMSs, refer to the AveHAS plots in figure 7A.

```
NikM   91 VSVVLLMQGVLFADGGLTALGVNITDMAIVTVVVSYALFRGLVKVLPRRR 140
          |... :|  ||  ||  ||  ..   : :||.| . | | | .  | |
NikQ   34 VTLLFIMSNVLLPDGAWTAFALS...LGLVTLVCTLAGF.GALFALKR.. 77
                   .         .         .         .         .
NikM  141 SSITAASFVAALISVPGAAVAFTLMYW.IGGTTDISIGKVASAMIGVHVL 189
           |   |  ||   |.|       |.  | || .  |     |.    :
NikQ   78 .SFVALPFVLVAITVLFTTPGRTVAGWQIGDSALIVTDAGATRFFSILAR 126
                   .         .         .
NikM  190 IGIGEAVITALTVGAVIAVRPDLVHGARGLEAPLKL 225
            |   .  |||  ||  : |||.|  | |  |  |
NikQ  127 SWISVQMAIALT..AVTNI.PDLLHALRHLRVPALL 159
```

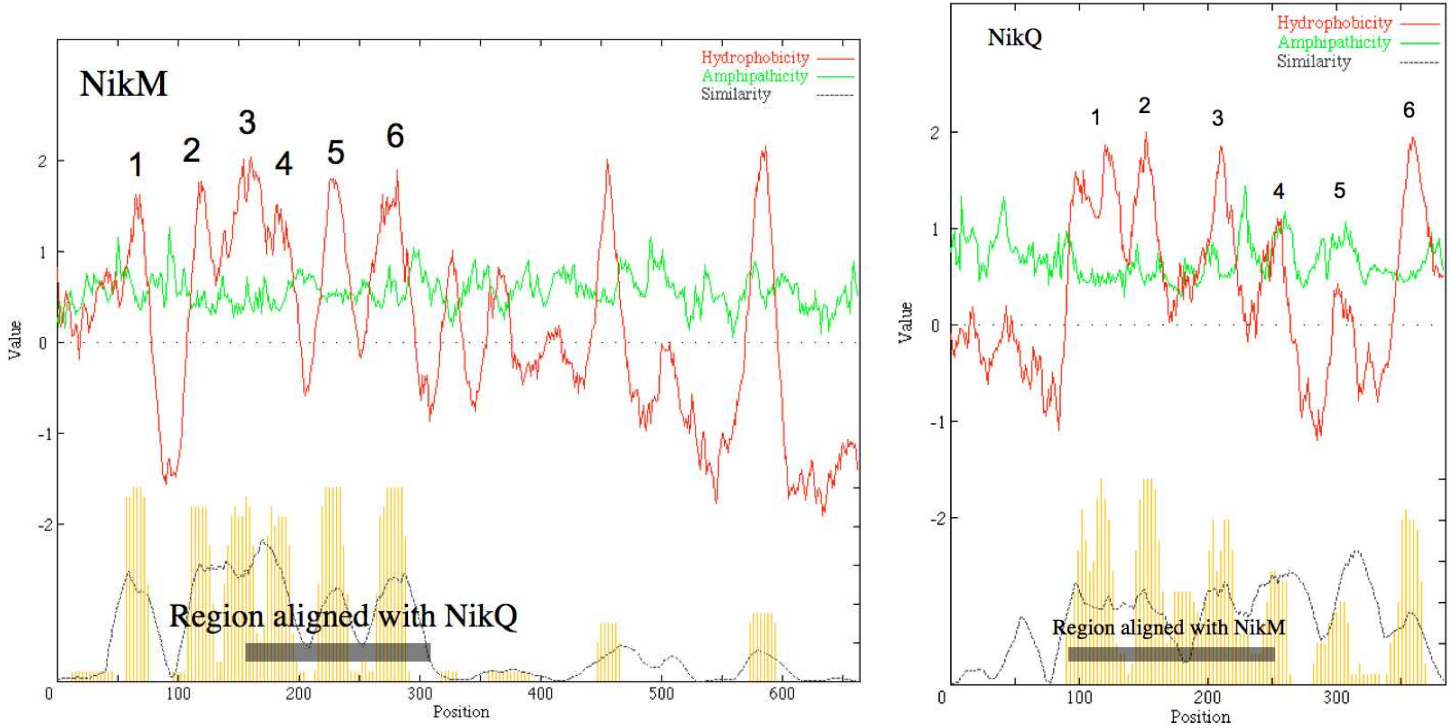
NikM (S subunit; gi#256677279) vs NikQ (T subunit; gi#156744274).
Comparison score = 9.9 SD; % identity = 33.6; % similarity = 37.6; # gaps = 6; # PSI-BLAST iterations = 2. The TMSs for both proteins were predicted with the TMHMM algorithm.

Figure 6B: Binary alignments of S and T subunits (NikM/NikO). See figure 2A for format of presentation. For the relative locations of TMSs, refer to the AveHAS plots in figure 7B.



TMS comparison of the two TM subunits of cobalt transporters. The plot on the left is CbiM (S subunit), and the one on the right is CbiQ (T subunit). The comparison score for this alignment is 11.9 SD.

Figure 7A: AveHAS plots for alignment of S (left) and T (right) subunits (CbiM/CbiQ). Comparison of the two subunits of a type-I ECF transporter (top: average hydropathy is indicated with a dark line and amphipathicity with a light line; bottom: similarity is indicated with a dashed line). The plots were generated with a modified AveHAS program (Zhai & Saier, 2001; Yen et al., 2009). The vertical bars at the bottom indicate the positions of the predicted TMSs as estimated using the TMHMM program (Krogh et al., 2001). Horizontal bars at the bottom indicate the regions that contain the best alignments as determined by the GAP alignment program. Conserved TMSs are numbered above the hydrophobicity peaks. See figure 6A for the alignment.



TMS comparison of the two TM subunits of nickel transporters. The plot on the left is NikM (S subunit), and the one on the right is NikQ (T subunit). The comparison score for this alignment is 9.9 SD. The plots were generated from 500 hits collected on the 2nd iteration of PSI BLAST.

Figure 7B: AveHAS plots for the S (left) and T (right) subunits (NikM/NikQ). See figure 7A for format of presentation. See figure 6B for the alignment.

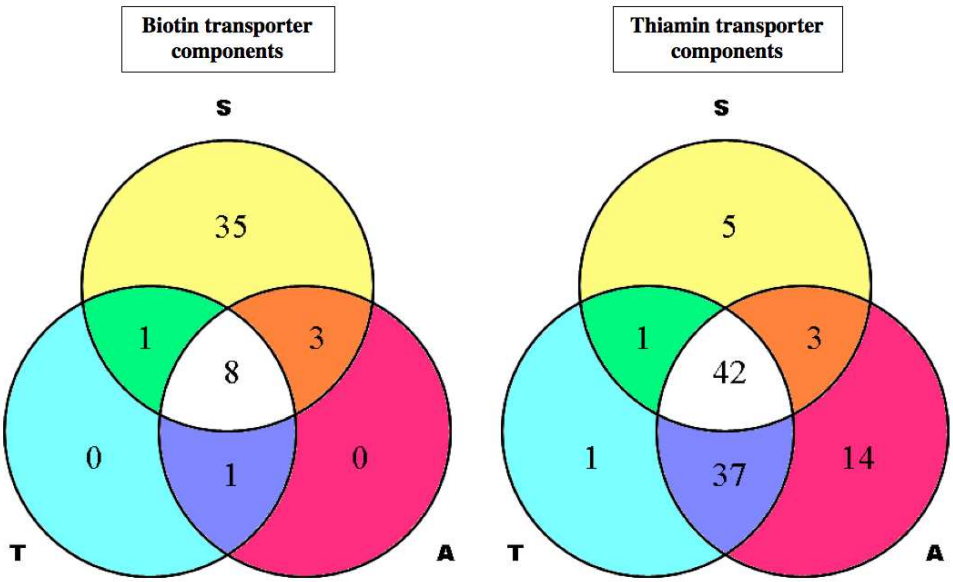


Figure 8: Venn diagram of the organismal distributions of gene clusters of the BioY and ThiW transporter components. S, T and A denote substrate recognition, ATPase-transducing, and ATPase subunits, respectively. 73 microbial genomes used for the diagram were chosen from species representative of archaea and bacteria. Refer to table 3A for the accession numbers of the proteins used and table 3B for details of cluster analysis.

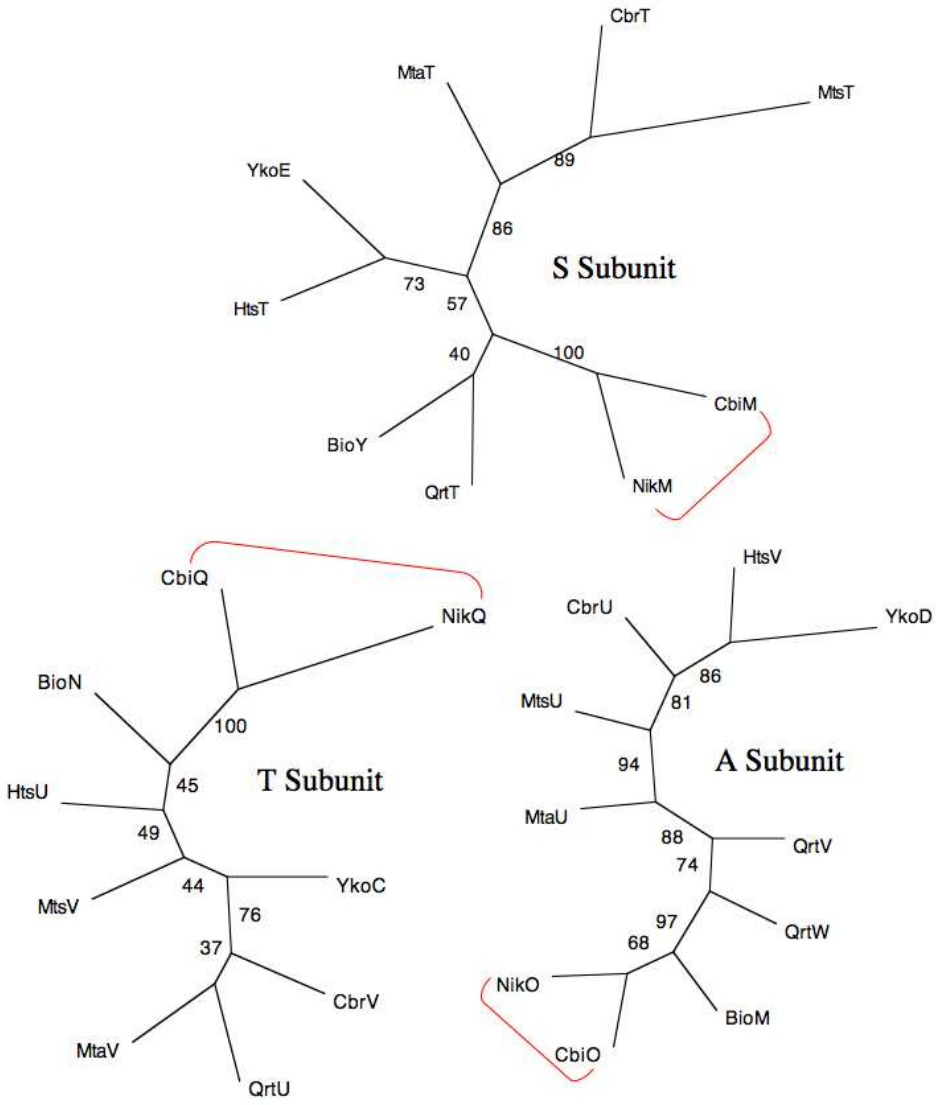


Figure 9: Phylogenetic trees for the three components of type-I ECF transporter homologues. Branch lengths estimate the average sequence divergence for all homologues of a given transporter family that have a dedicated energizing module. The number next to each node corresponds to the bootstrap value and indicates the probability of the shown branching pattern. Brackets indicate clusters that appear through all three transporter components. S indicates the substrate recognition subunit, T the ATPase anchoring subunit, and A the ATPase. Some transporters possess a fusion of the two heterologous ATPases, and for these systems both were included in the analysis. The homologues were collected from the SEED database (http://seed-viewer.theseed.org/).

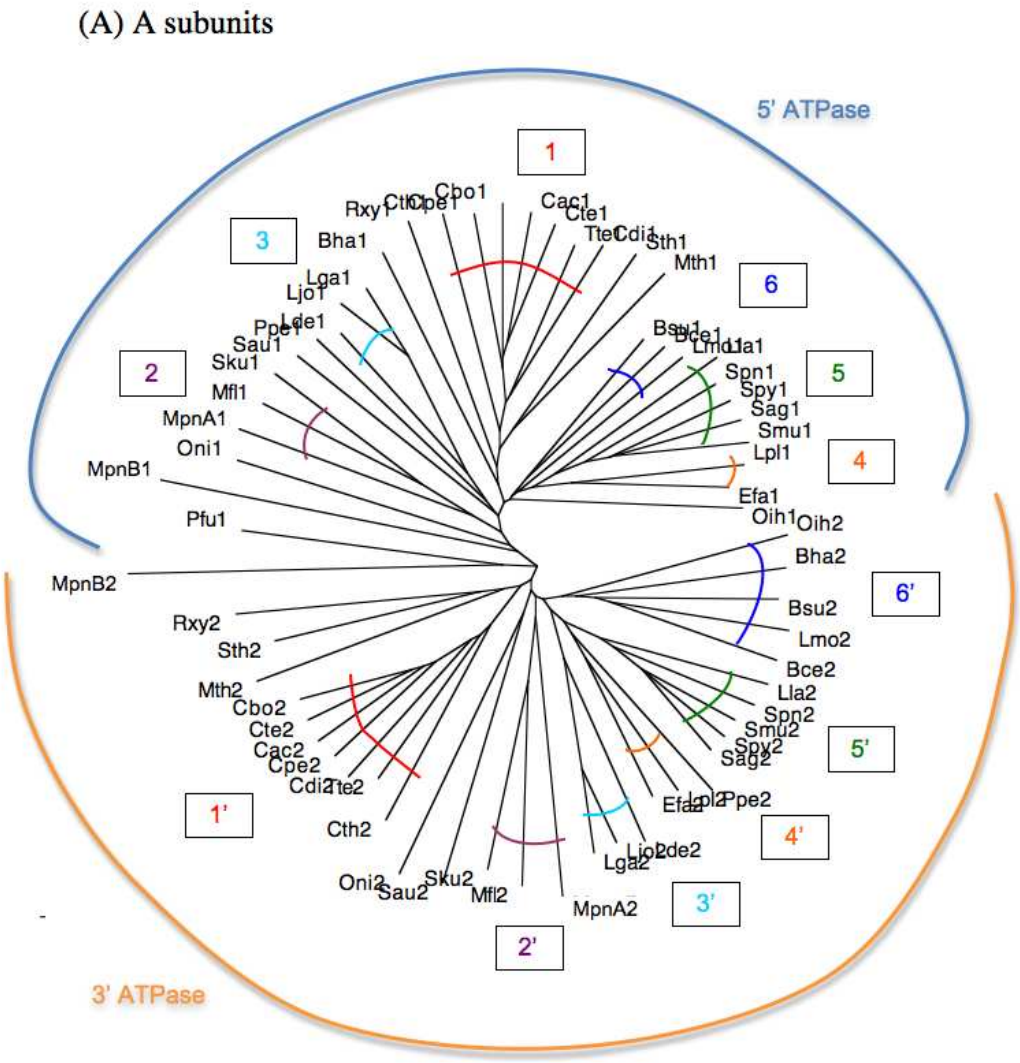


Figure 10A: Phylogenetic trees for the A subunit homolgoues of type-II ECF transporters. Each branch represents a type-II ECF energizing component from a particular organism. Inner brackets and their corresponding numbers in boxes indicate clusters with similar evolutionary relationships between the ATPase (A subunit) and ATPase-transducing (T subunit) components of the energizing modules. For the ATPase tree, two paralogues are encoded in each operon, with the upstream paralogues indicated by the number 1 and the downstream paralogues by the number 2 following the organismal abbreviation. Apostrophes indicate clusters that are associated with the downstream ATPase paralogues. The sequences of the energizing components were obtained from the SEED database (http://seed-viewer.theseed.org/). Refer to table 5 for the accession numbers of the proteins used.

(B) T subunits

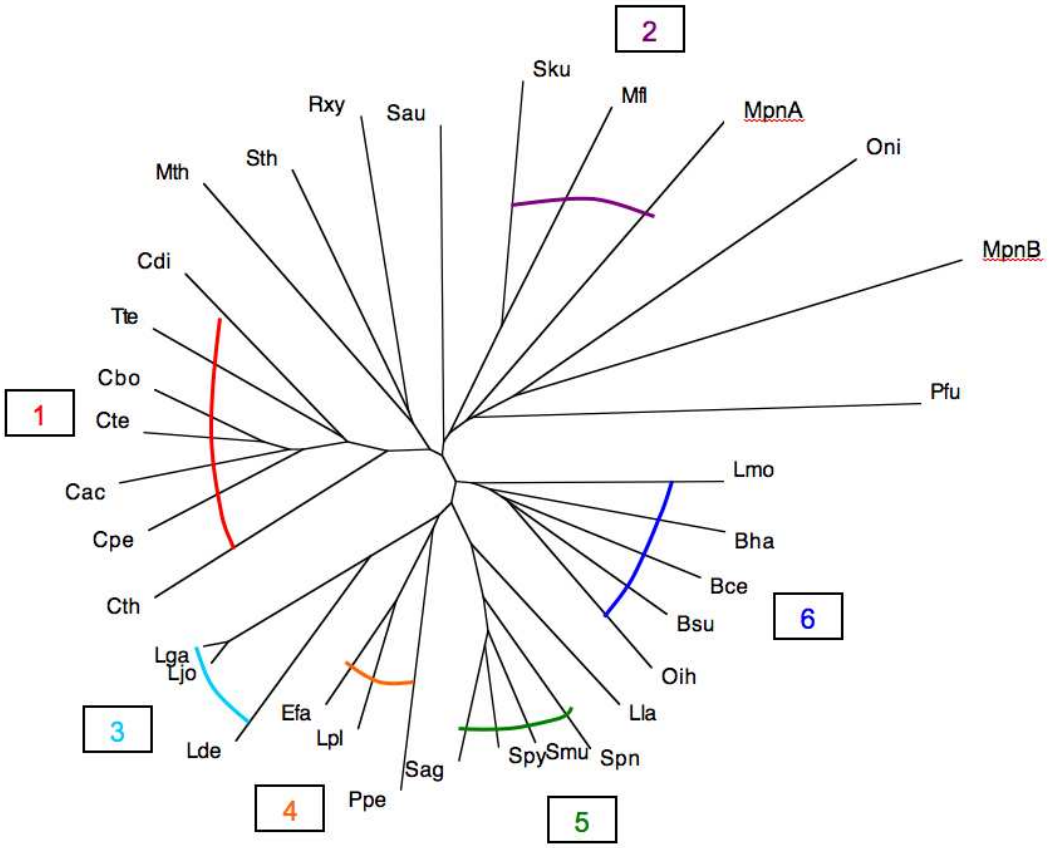


Figure 10B: Phylogenetic trees for the T subunit homolgoues of type-II ECF transporters. See figure 10A for figure description.

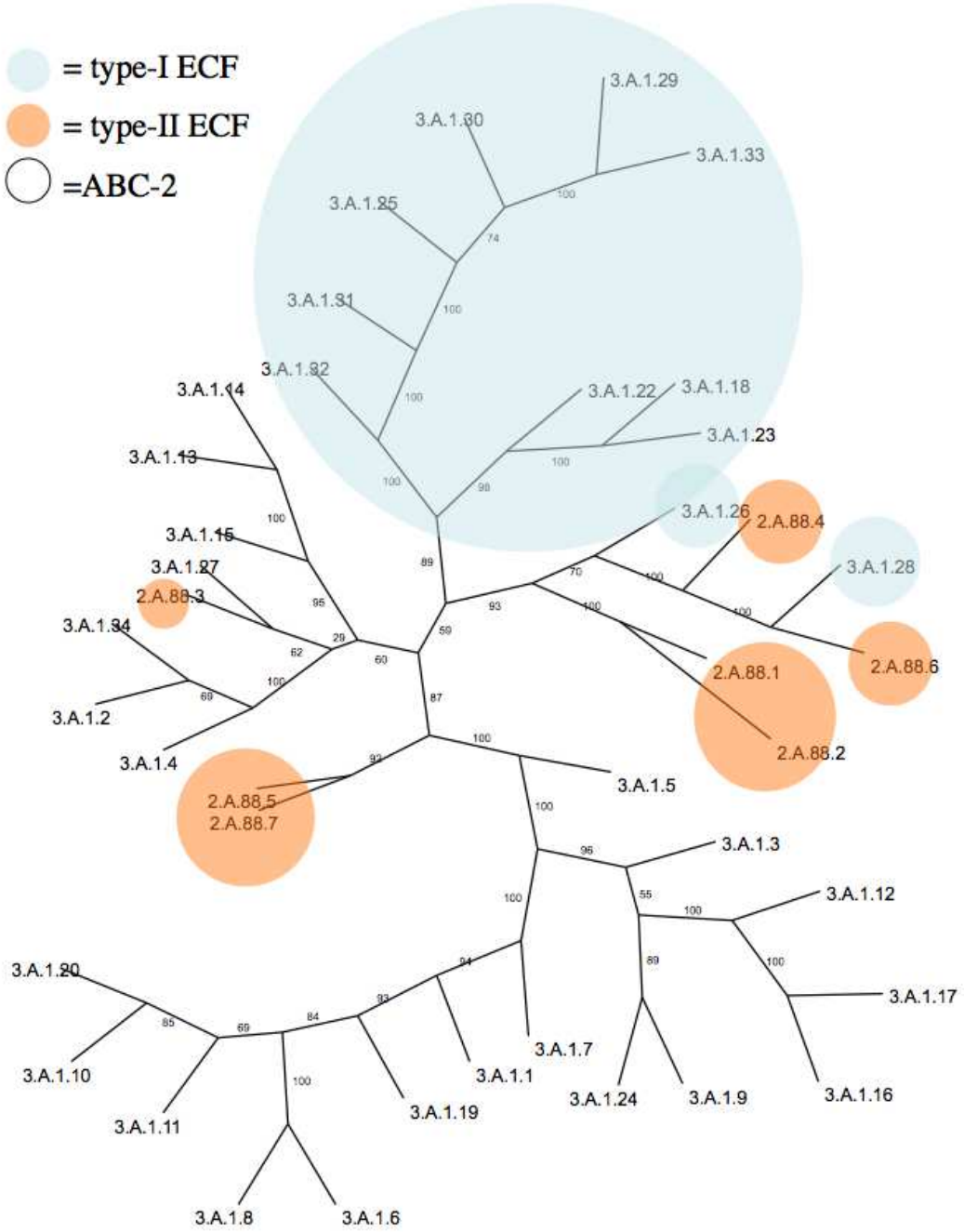


Figure 11: Superfamily trees for the S subunit homologues belonging to type-I/II ECF transporters as well as the membrane constituents of representative ABC2 members. The sequences were obtained from the TCDB database (see table 6 for the accession numbers), and their TC#'s are indicated. Refer to the text for figure description.

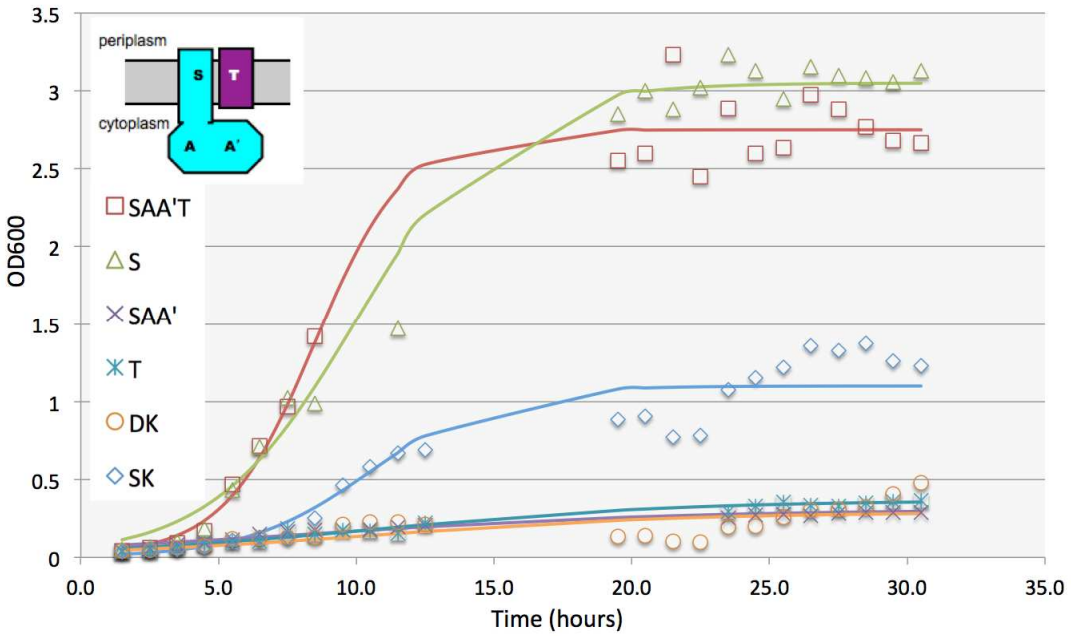


Figure 12: Growth analysis of thiamine synthesis/transport-null strains of *E. coli* expressing various transporter constituents of ThiW from *Mycobacterium smegmatis str. MC2 155*. In the sidebar, SK refers to a single knockout *E. coli* mutant unable to synthesize thiamine *de novo*, and DK indicates a double knockout mutant incapable of both thiamine synthesis and uptake. Constituents expressed include: (1) the S subunit alone (S), (2) the complete system (SAA'T), (3) the SAA' protein, and (4) the T subunit. Data are collected from two to six independent experiments, and 5 nanomolar thiamine was used in the growth media. In the inset, the native transport complex of the thiamine transporter from *Mycobacterium smegmatis str. MC2 155* is proposed.

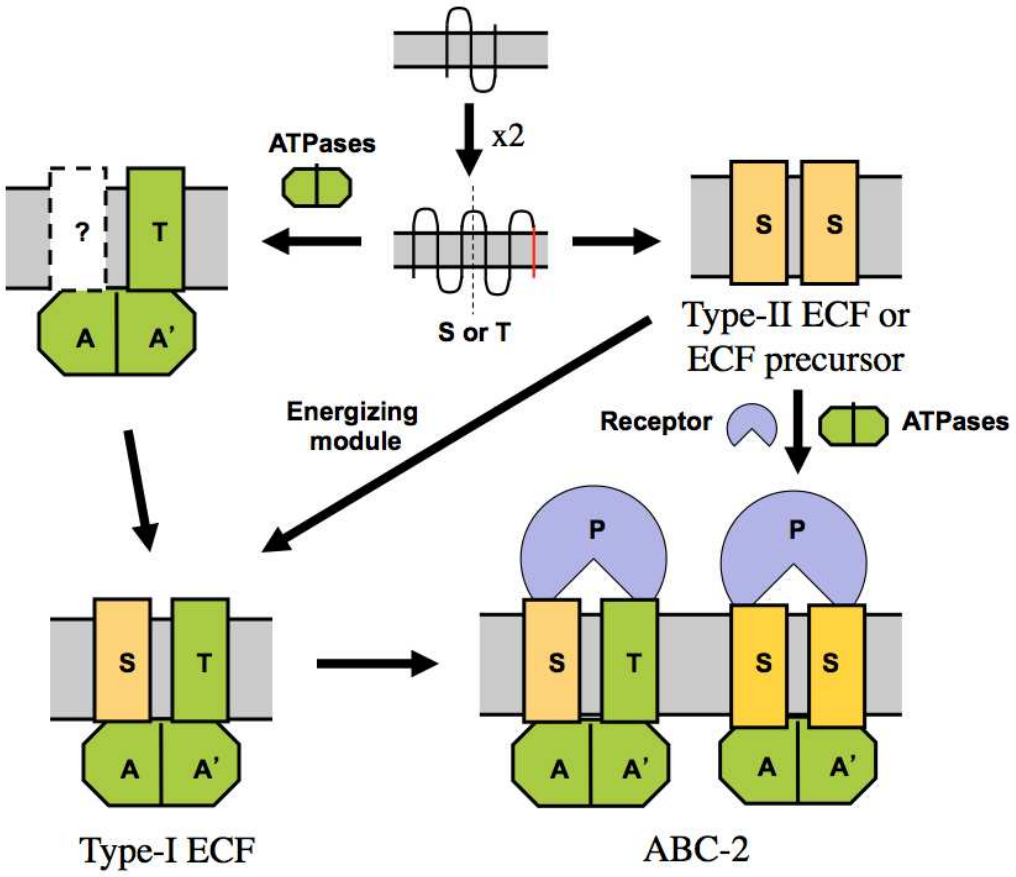
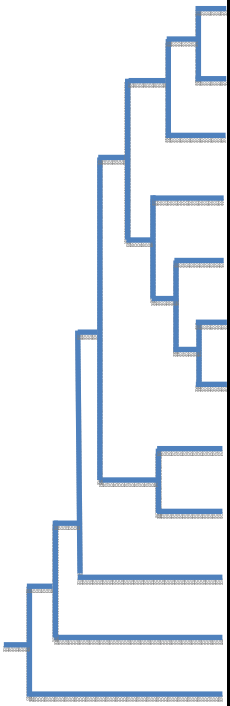


Figure 13: Proposed pathway of ECF and ABC2 evolution. Arrows indicate the proposed direction of evolution. Instead of individual subunits, transmembrane segments are shown for the progenitor of the S and T subunits of the ECF transporters before the two subunits diverged.



Phylum (Class)	Average riboswitch #/genome	Average riboswitch #/Mbps
Firmicutes	42.7	14.1
Actinobacteria	11.8	2.8
Cyanobacteria	6.1	1.4
Proteobacteria (delta)	7.6	2.2
Proteobacteria (alpha)	12.0	2.6
Proteobacteria (beta)	13.8	2.0
Proteobacteria (gamma)	16.2	3.5
Chlorobi	6.6	2.5
Bacteroidetes	7.6	1.5
Deinococcus-Thermus	12.8	5.0
Chloroflexi	19.2	3.5
Thermotogae	7.9	4.0

Figure 14: Total riboswitch distribution imposed on the phylogenetic tree of the major microbial phyla. The cells in the table are shaded to correspond to the abundance of riboswitches. The darker the shading, the more abundant the riboswitch as indicated by the boxed numbers. The phylogenetic tree is adapted from Madigan et al., 2003.

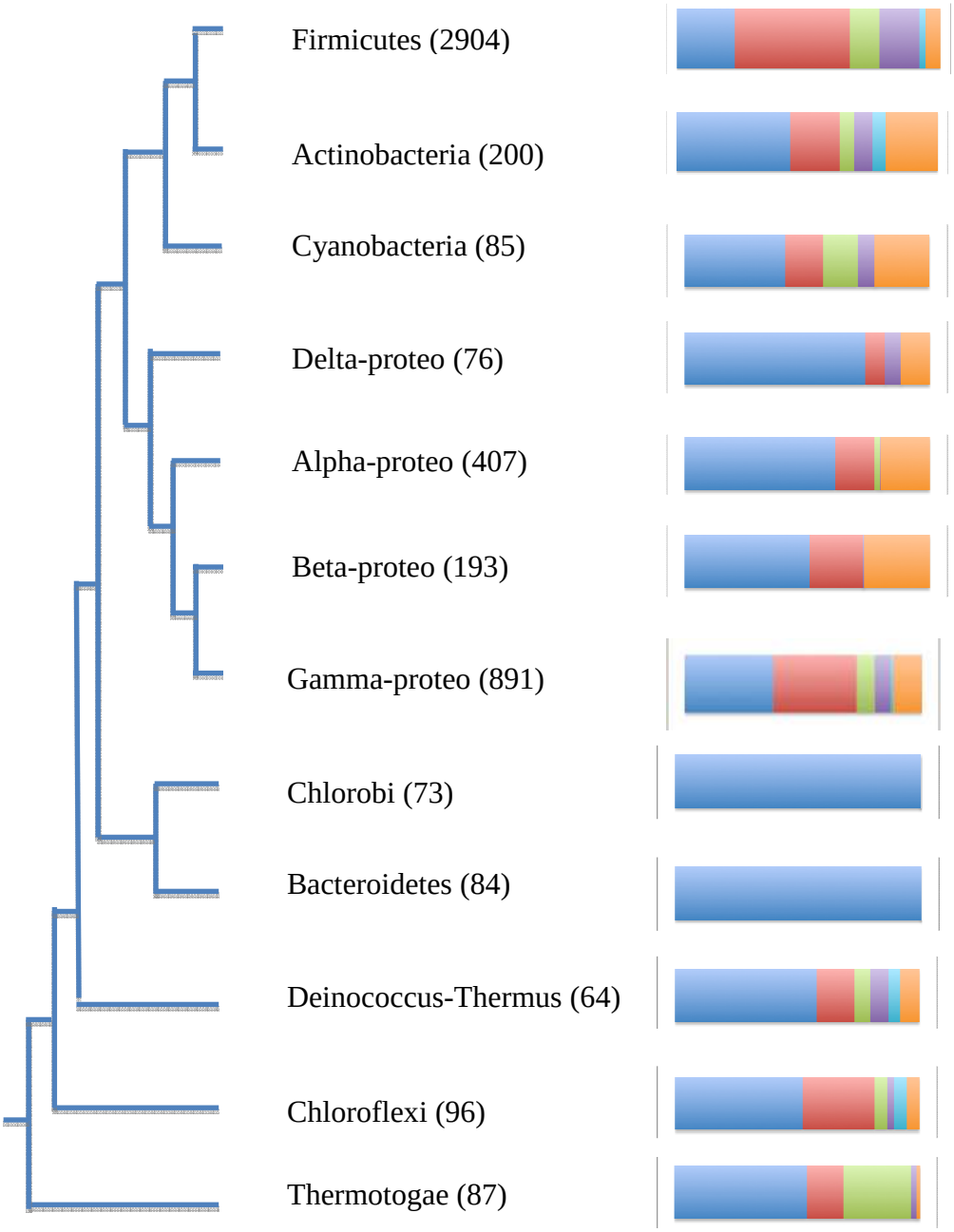


Figure 15: Proportion of riboswitches with different metabolic roles within individual phylum. Each sector in the bar graph, from left to right, represents a group of riboswitches with specificity for coenzymes, amino acids, ribosomal subunits, nucleotide derivatives, ions/sugar, and putative riboswitches, respectively. Numbers in parentheses following the phylum label indicates total number of riboswitches found for that phylum. Refer to tables 14 for the total number of sites for each phylum as well as their percentage against total sites. The phylogenetic tree is the same as from figure 14.

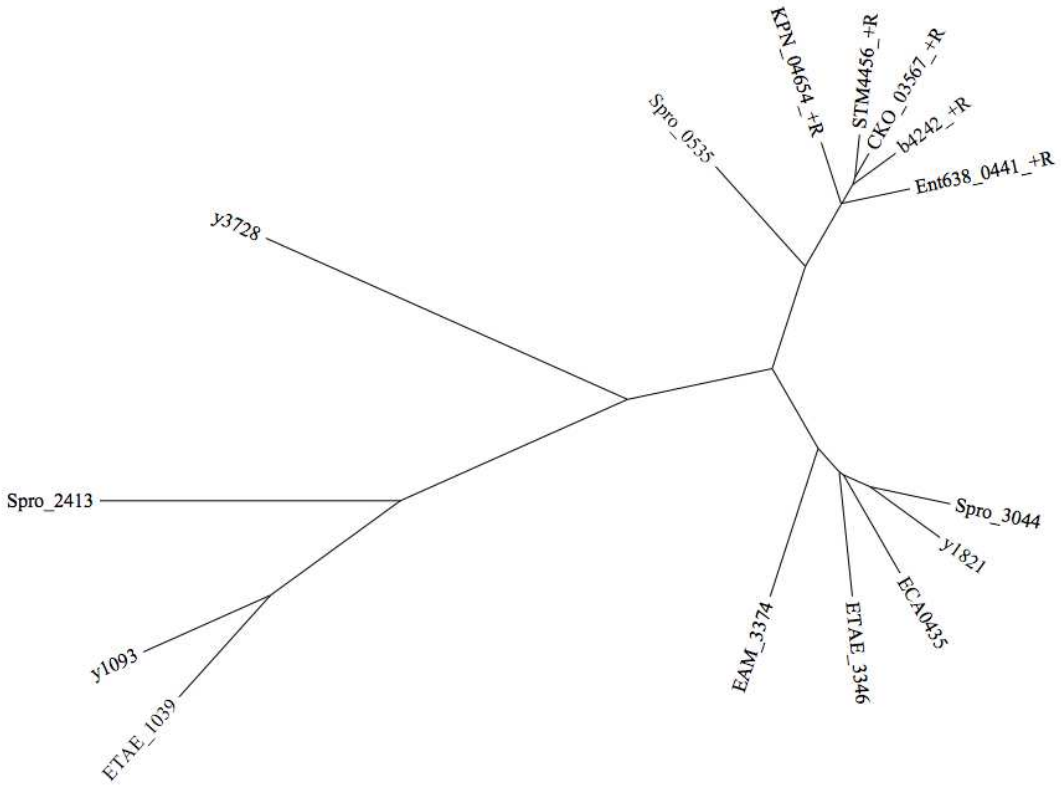


Figure 16: Phylogenetic tree of homologous MgtA proteins in Enterobacteriales. Each node represents an MgtA homologue. Homologues with a '_+R' following the locus tag are regulated by a Mg_sensor riboswitch element.

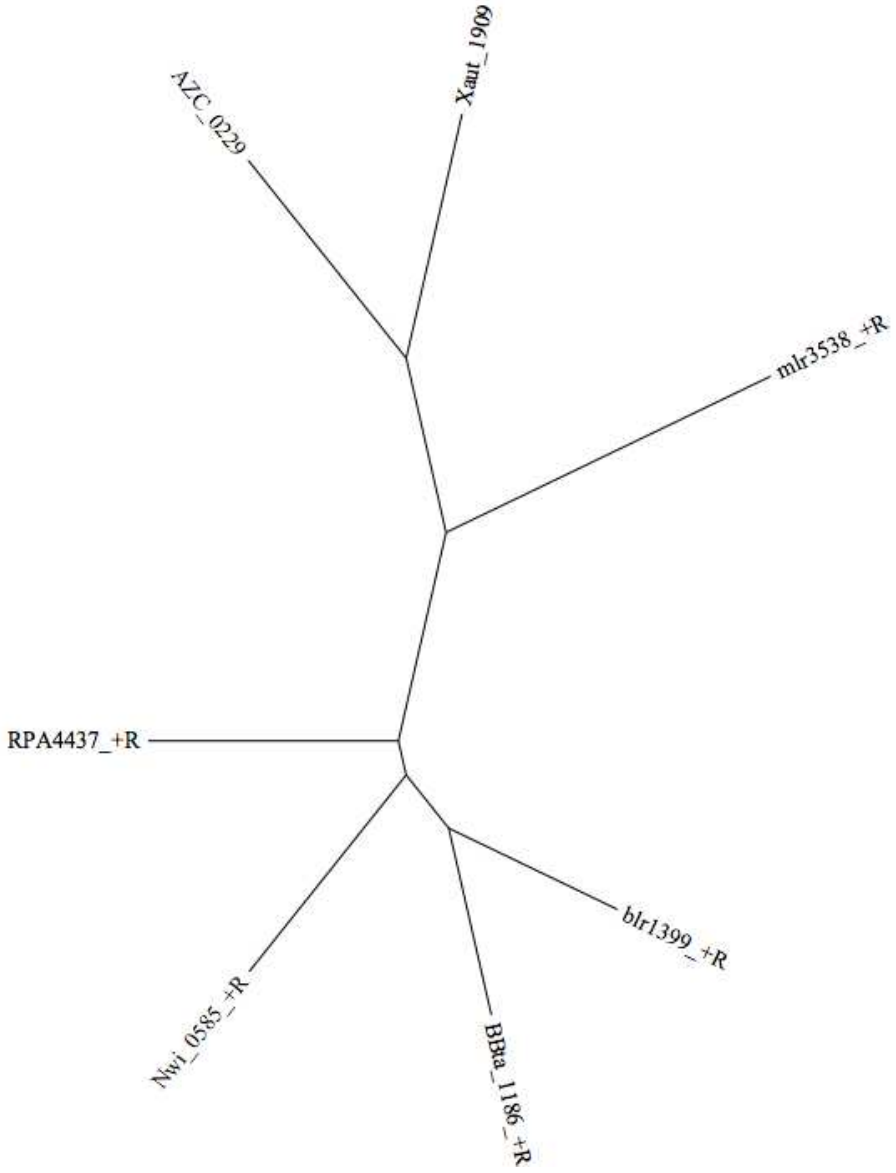


Figure 17: Phylogenetic tree of homologous MetX proteins in Rhizobiales. Each node represents a MetX homologue. Homologues with a '_+R' following the locus tag are regulated by a SAM-alpha riboswitch element.

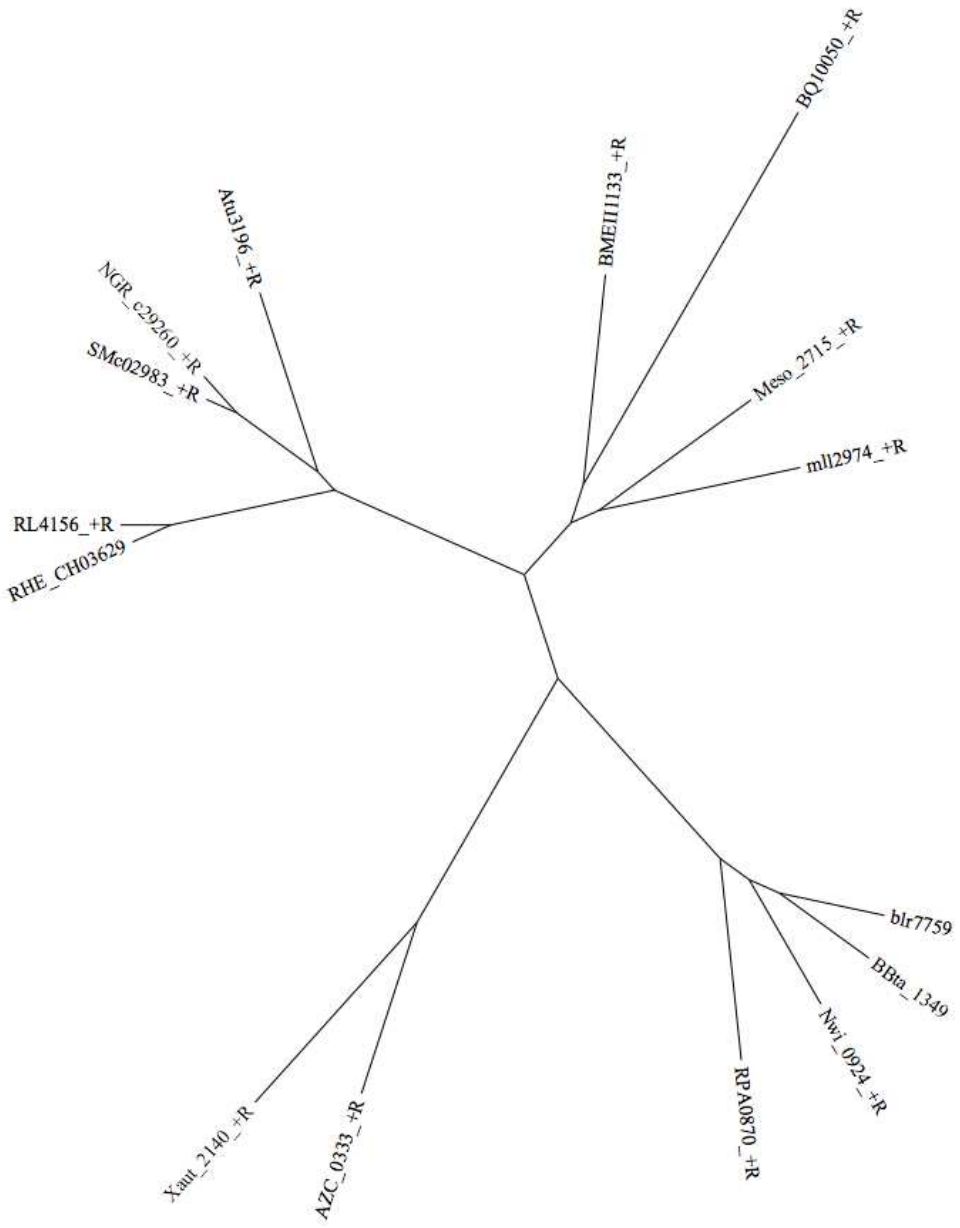


Figure 18: Phylogenetic tree of homologous SpeF proteins in Rhizobiales. Each node represents a SpeF homologue. Homologues with a '_+R' following the locus tag are regulated by a speF riboswitch element.

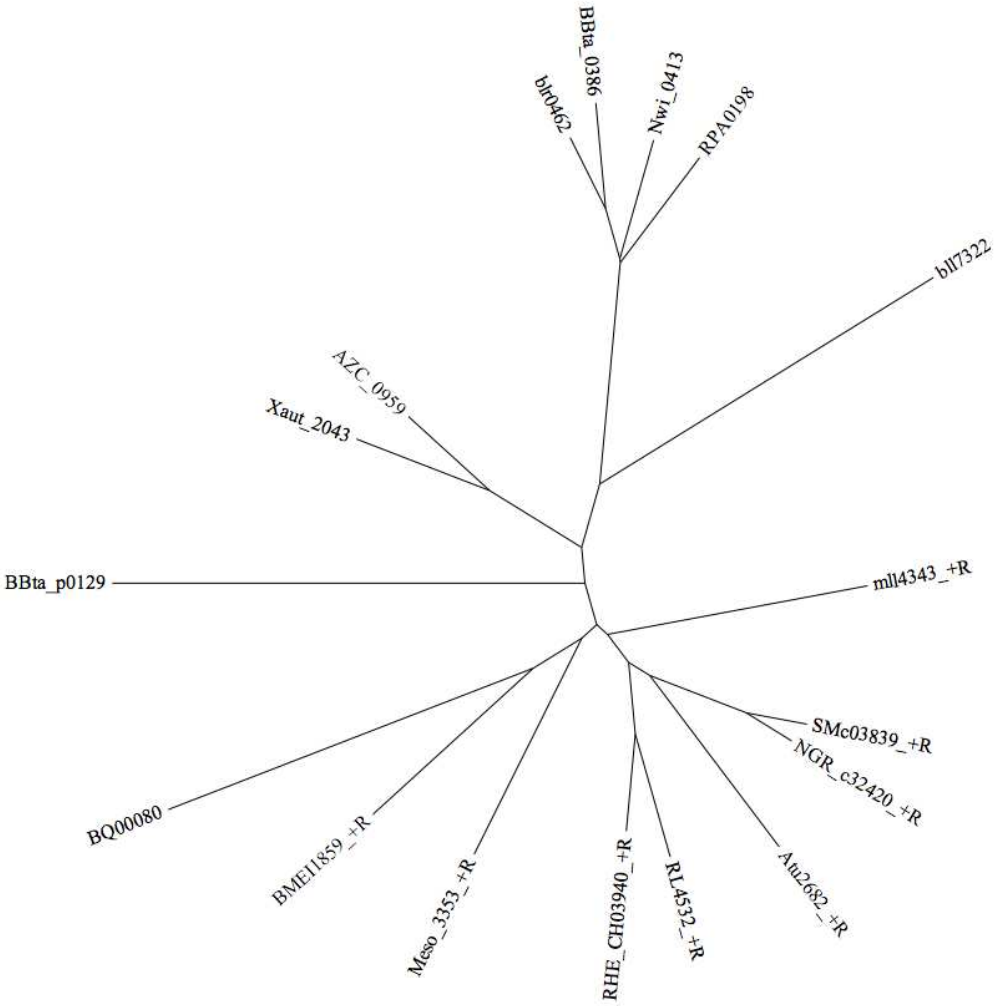


Figure 19: Phylogenetic tree of homologous YbhL proteins in Rhizobiales. Each node represents a YbhL homologue. Homologues with a '_+R' following the locus tag are regulated by a ybhL riboswitch element.

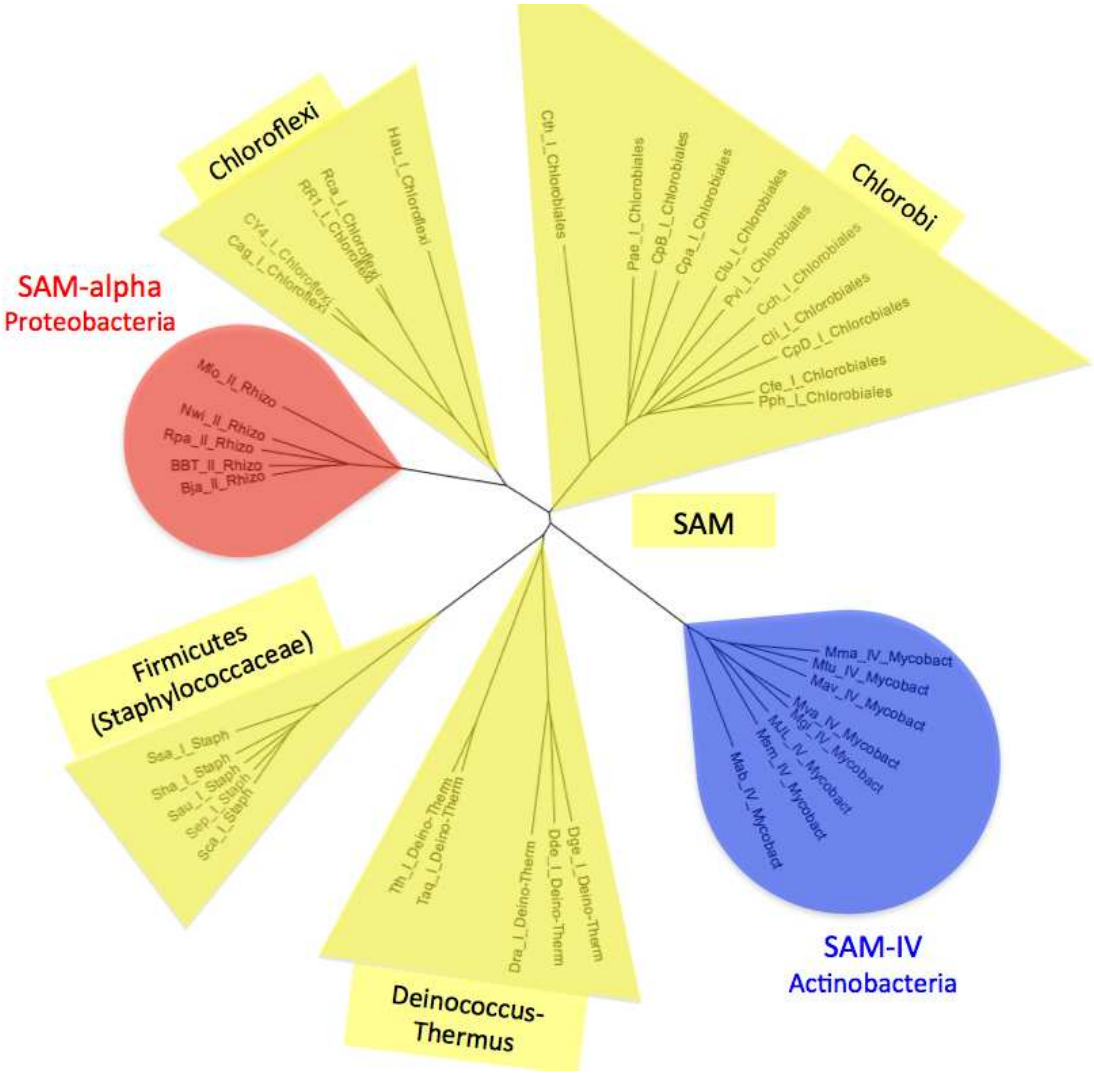


Figure 20: Phylogenetic tree of orthologous MetX proteins regulated by functionally equivalent SAM riboswitches. Each node represents a MetX homologue regulated by a single riboswitch. The Roman numeral following the first underscore indicates the identity of the riboswitch regulator (I=SAM, II=alpha, IV=IV), and the taxonomic identifier follows the second underscore. Different riboswitch classes are shaded differently. The riboswitches represented are SAM, SAM-alpha, and SAM-IV.

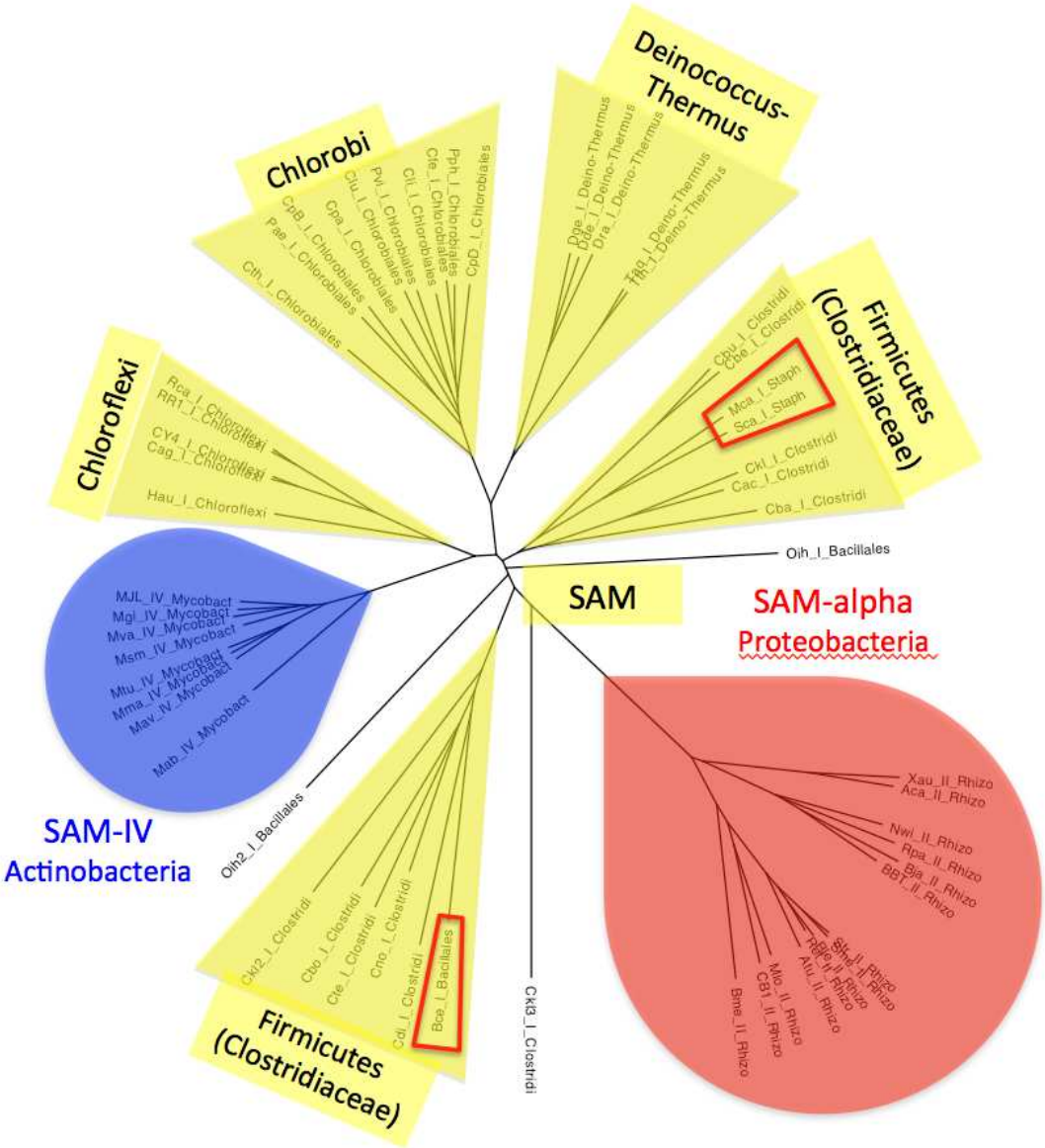


Figure 21: Phylogenetic tree of orthologous MetY proteins regulated by functionally equivalent SAM riboswitches. Each node represents a MetY orthologue regulated by a single riboswitch. Refer to figure 16 for format of presentation. Paralogues are indicated by a number immediately following the 3-letter organismal abbreviation. Homologues framed by red boxes indicate possible cases of horizontal gene transfer. Refer to text for detail.

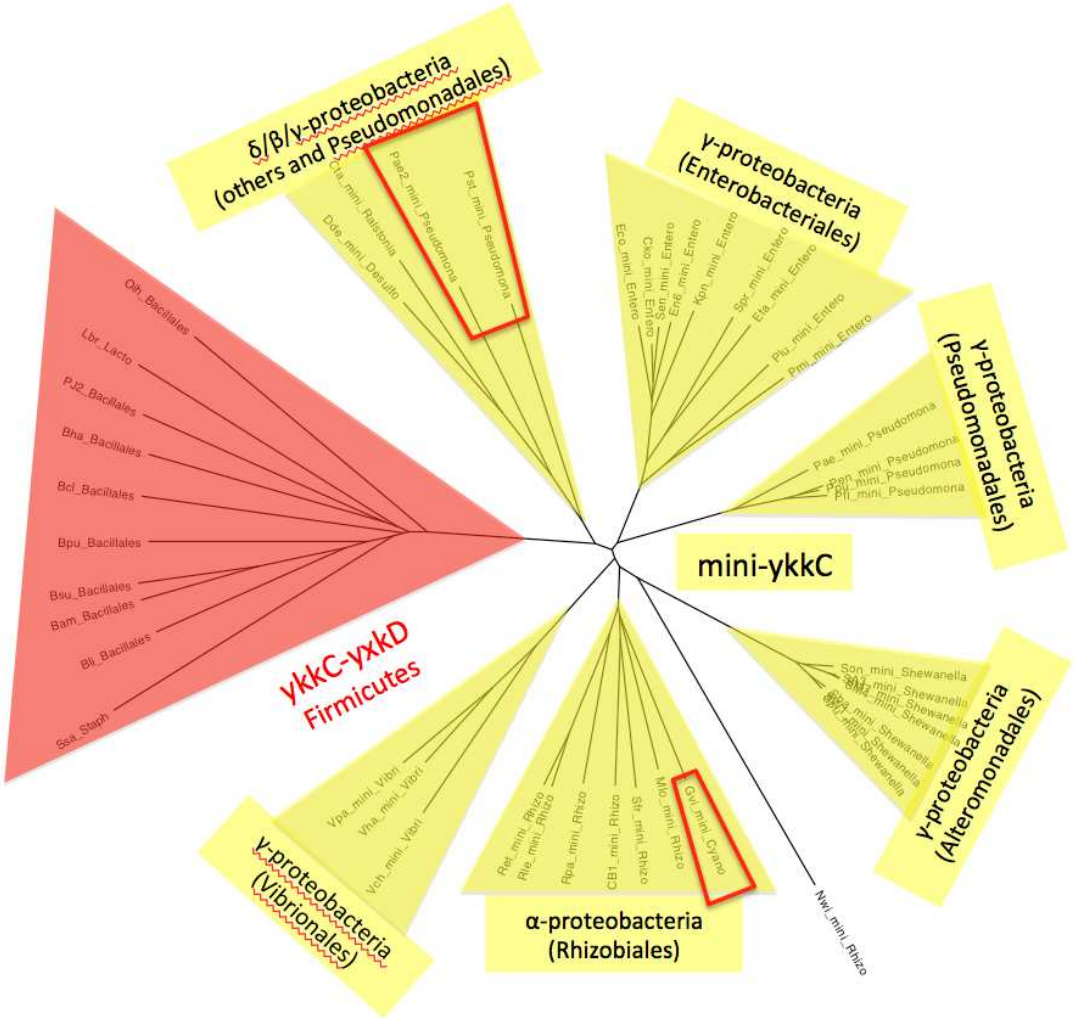


Figure 22: Phylogenetic tree of orthologous YkkC proteins regulated by functionally equivalent riboswitches. Each node represents a YkkC orthologue regulated by a single riboswitch. Refer to figure 16 for format of presentation. Homologues framed by red boxes indicate possible cases of horizontal gene transfer. The riboswitches represented are *ykkC-yxkD* and mini-*ykkC*.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389-402.

Bailey, T.L. & Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. 2:28-36.

Breaker, R.R. (2012) Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol*. 4. pii: a003566.

Bröer, S., Ji, G., Bröer, A. & Silver, S. (1993) Arsenic efflux governed by the arsenic resistance determinant of *Staphylococcus aureus* plasmid pI258. *J. Bacteriol*. 175:3480-5.

Busch, W. & Saier, M.H., Jr. (2002) The Transporter Classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol*. 37:287-337.

Castillo, R. & Saier, M.H. (2010) Functional Promiscuity of Homologues of the Bacterial ArsA ATPases. *Int. J. Microbiol*. 2010:187373.

Chan, H., Babayan, V., Blyumin, E., Gandhi, C., Hak, K., Harake, D., Kumar, K., Lee, P., Li, T.T., Liu, H.Y., Lo, T.C., Meyer, C.J., Stanford, S., Zamora, K.S. & Saier, M.H., Jr. (2010) The p-type ATPase superfamily. *J. Mol. Microbiol. Biotechnol*. 19:5-104.

Chang, A.B., Lin, R., Keith Studley, W., Tran, C.V. & Saier, M.H., Jr. (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Mol. Membr. Biol*. 21:171-81.

Chen, J.S., Reddy, V., Chen, J.H., Shlykov, M.A., Zheng, W.H., Cho, J., Yen, M.R. & Saier, M.H. Jr. (2011) Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments. *J. Mol. Microbiol. Biotechnol*. 21:83-96.

Datsenko, K.A. & Wanner, B.L. (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. USA*. 97:6640-5.

Davidson, A.L., Dassa, E., Orelle, C. & Chen, J. (2008) Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* 72:317-64.

Dayhoff, M.O., Barker, W.C. & Hunt, L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol*. 91:524-45

Debut, A.J., Dumay, Q.C., Barabote, R.D. & Saier, M.H., Jr. (2006) The iron/lead transporter superfamily of Fe/Pb2+ uptake systems. *J. Mol. Microbiol. Biotechnol*. 11:1-9.

Devereux, J., Haeberli, P. & Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res*. 12:387-95.

Doolittle, R.F. (1981). Similar amino acid sequences: chance or common ancestry? *Science*. 214:149-59.

Eddy, S.R. & Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*. 22:2079-88.

Eitinger, T., Rodionov, D.A., Grote, M. & Schneider, E. (2011) Canonical and ECF-type ATP-binding cassette importers in prokaryotes: diversity in modular organization and cellular functions. *FEMS Microbiol. Rev*. 35:3-67.

Erkens, G.B., Berntsson, R.P., Fulyani, F., Majsnerowska, M., Vujičić-Žagar, A., Ter Beek, J., Poolman, B. & Slotboom, D.J. (2011) The structural basis of modularity in ECF-type ABC transporters. *Nat Struct. Mol. Biol.* 18:755-60.

Erkens, G.B. & Slotboom, D.J. (2010) Biochemical characterization of ThiT from Lactococcus lactis: a thiamin transporter with picomolar substrate binding affinity. *Biochemistry*. 49:3203-12.

Finkenwirth, F., Neubauer, O., Gunzenhäuser, J., Schoknecht, J., Scolari, S., Stöckl, M., Korte, T., Herrmann, A. & Eitinger, T. (2010) Subunit composition of an energy-coupling-factor-type biotin transporter analysed in living bacteria. *Biochem. J.* 431:373-80.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature*. 345:60-3.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res*. 31:439-41.

Hebbeln, P., Rodionov, D.A., Alfandega, A. & Eitinger, T. (2007) Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. *Proc. Natl. Acad. Sci. USA*. 104:2909-14.

Henderson, G.B., Zevely, E.M. & Huennekens, F.M. (1979) Mechanism of folate transport in *Lactobacillus casei*: evidence for a component shared with the thiamine and biotin transport systems. *J. Bacteriol.* 137: 1308–1314.

Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*. 14:378-9.

Hofmann, K. & Stoffel, W. (1993) TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*. 374:166.

Hvorup, R., Chang, A.B. & Saier, M.H., Jr. (2003) Bioinformatic analyses of the bacterial L-ascorbate phosphotransferase system permease family. *J. Mol. Microbiol. Biotechnol*. 6:191-205.

Jack, D.L., Yang, N.M. & Saier, M.H., Jr. (2001) The drug/metabolite transporter superfamily. *Eur. J. Biochem*. 268:3620-39.

Kazanov, M.D., Vitreschak, A.G. & Gelfand, M.S. (2007) Abundance and functional diversity of riboswitches in microbial communities. *BMC Genomics*. 8:347.

Kohler, K., Forster, I.C., Lambert, G., Biber, J. & Murer, H. (2001) The functional unit of the renal type IIa Na+/Pi cotransporter is a monomer. *J. Biol. Chem*. 275:26113-26120.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*. 305:567–580.

Kuroda, M., Dey, S., Sanders, O.I. & Rosen, B.P. (1997) Alternate energy coupling of ArsB, the membrane subunit of the Ars anion-translocating ATPase. *J. Biol. Chem*. 272:326-31.

Lebens, M., Lundquist, P., Söderlund, L., Todorovic, M. & Carlin, N.I. (2002) The *nptA* gene of *Vibrio cholerae* encodes a functional sodium-dependent phosphate cotransporter homologous to the type II cotransporters of eukaryotes. *J. Bacteriol*. 184:4466-74.

Lee, J.H., Harvat, E.M., Stevens, J.M., Ferguson, S.J. & Saier, M.H., Jr. (2007) Evolutionary origins of members of a superfamily of integral membrane cytochrome c biogenesis proteins. *Biochim. Biophys. Acta*. 1768:2164-81.

Li, W. & Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22:1658-9.

Madigan, M.T., Martinko, J.M. & Parker, J. (2003) Brock Biology of Microorganisms 10th edition. Upper Saddle River, NJ: Pearson Education, Inc.

Mansour, N.M., Sawhney, M., Tamang, D.G., Vogl, C. & Saier, M.H., Jr. (2007) The bile/arsenite/riboflavin transporter (BART) superfamily. *FEBS J*. 274:612-29.

Miyamoto, K., Ito, M., Tatsumi, S., Kuwahata, M. & Segawa, H. (2007) New aspect of renal phosphate reabsorption: the type IIc sodium-dependent phosphate transporter. *Am. J. Nephrol*. 27:503-15.

Mulligan, C., Geertsma, E.R., Severi, E., Kelly, D.J., Poolman, B. & Thomas, G.H. (2009) The substrate-binding protein imposes directionality on an electrochemical sodium gradient-driven TRAP transporter. *Proc. Natl. Acad. Sci. USA*. 106:1778-83.

Mulligan, C., Kelly, D.J. & Thomas, G.H. (2007) Tripartite ATP-independent periplasmic transporters: application of a relational database for genome-wide analysis of transporter gene frequency and organization. *J. Mol. Microbiol. Biotechnol*. 12:218-26.

Murer, H., Hernando, N., Forster, I. & Biber, J. (2000) Proximal tubular phosphate reabsorption: molecular mechanisms. *Physiol. Rev*. 80:1373-409.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 25:1335-7.

Neubauer, O., Alfandega, A., Schoknecht, J., Sternberg, U., Pohlmann, A. & Eitinger, T. (2009) Two essential arginine residues in the T components of energy-coupling factor transporters. *J. Bacteriol*. 191:6482-8.

Novichkov, P.S., Rodionov, D.A., Stavrovskaya, E.D., Novichkova, E.S., Kazakov, A.E., Gelfand, M.S., Arkin, A.P., Mironov, A.A. & Dubchak, I. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res*. 38(Web Server issue):W299-307.

Nudler, E. & Mironov, A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem*. Sci. 29:11-7.

Oldham, M.L., Khare, D., Quiocho, F.A., Davidson, A.L. & Chen, J. (2007) Crystal structure of a catalytic intermediate of the maltose transporter. *Nature*. 450:515-21.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank,

E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 33:5691-702.

Prakash, S., Cooper, G., Singhi, S. & Saier, M.H., Jr. (2003) The ion transporter superfamily. *Biochim. Biophys. Acta*. 1618:79-92.

Rabus, R., Jack, D.L., Kelly, D.J. & Saier, M.H., Jr. (1999) TRAP transporters: an ancient family of extracytoplasmic solute-receptor-dependent secondary active transporters. *Microbiol.* 145:3431-45.

Rappé, M.S. & Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol*. 57:369-94.

Reddy, V.S., Shlykov, M.A., Castillo, R., Sun, E.I. & Saier, M.H. Jr. (2012) The Major Facilitator Superfamily (MFS) Revisited. *FEBS J.* 279:2022-35.

Rodionov, D.A., Hebbeln, P., Eudes, A., ter Beek, J., Rodionova, I.A., Erkens, G.B., Slotboom, D.J., Gelfand, M.S., Osterman, A.L., Hanson, A.D. & Eitinger, T. (2009) A novel class of modular transporters for vitamins in prokaryotes. *J. Bacteriol*. 191:42-51.

Rodionov, D.A., Hebbeln, P., Gelfand, M.S. & Eitinger, T. (2006) Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J. Bacteriol*. 188:317-27.

Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. (2003a) Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J. Biol Chem*. 278:41148-59.

Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. (2003b) Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res*. 31:6748-57.

Rodionov, D.A., Mironov, A.A. & Gelfand, M.S. (2002a) Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res*. 12:1507-16.

Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. & Gelfand, M.S. (2002b) Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J. Biol. Chem*. 277:48949-59.

Rost, B., Yachdav, G. & Liu, J. (2003) The PredictProtein Server. *Nucleic Acids Res.* 32:W321-W326.

Saier, M.H., Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev*. 64:354-411.

Saier, M.H., Jr. (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol. Rev. 58:71-93.

Saier, M.H., Jr. (2006) Protein secretion and membrane insertion systems in gram-negative bacteria. *J. Membr. Biol.* 214:75-90.

Saier, M.H., Jr. (2003a) Answering fundamental questions in biology with bioinformatics. *Am. Soc. Microbiol. News.* 69:175-181.

Saier, M.H., Jr. (2003b) Tracing pathways of transport protein evolution. *Mol. Microbiol.* 48:1145-56.

Saier, M.H., Jr., Hvorup, R.N. & Barabote, R.D. (2005) Evolution of the bacterial phosphotransferase system: from carriers and enzymes to group translocators. *Biochem. Soc. Trans*. 33:220-4.

Saier, M.H., Jr., Ma, C.H., Rodgers, L., Tamang, D.G. & Yen, M.R. (2008) Protein secretion and membrane insertion systems in bacteria and eukaryotic organelles. *Adv. Appl. Microbiol*. 65:141-97.

Saier, M.H., Jr., Tran, C.V. & Barabote, R.D. (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res*. 34 (Database issue):D181-6.

Saier, M.H. Jr., Yen, M.R., Noto, K., Tamang, D.G. & Elkan, C. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res*. 37:D274-8.

Sarkisova, S., Patrauchan, M.A., Berglund, D., Nivens, D.E. & Franklin, M.J. (2005) Calcium-induced virulence factors associated with the extracellular matrix of mucoid Pseudomonas aeruginosa biofilms. *J. Bacteriol*. 187:4327-37.

Sarsero, J.P., Merino, E. & Yanofsky, C. (2000) A *Bacillus subtilis* gene of previously unknown function, yhaG, is translationally regulated by tryptophan-activated

TRAP and appears to be involved in tryptophan transport. *J. Bacteriol*. 182:2329-31.

Tamang, D.G. & Saier, M.H., Jr. (2006) The cecropin superfamily of toxic peptides. *J. Mol. Microbiol. Biotechnol*. 11:94-103.

Tenenhouse, H.S. (2005) Regulation of phosphorus homeostasis by the type IIa Na/phosphate cotransporter. *Annu. Rev. Nutr*. 25:197-214.

ter Beek, J., Duurkens, R.H., Erkens, G.B. & Slotboom, D.J. (2011) Quaternary structure and functional unit of Energy Coupling Factor (ECF)-type transporters. *J. Biol. Chem*. 286:5471-5.

Thever, M.D. & Saier, M.H., Jr. (2009) Bioinformatic characterization of p-type ATPases encoded within the fully sequenced genomes of 26 eukaryotes. *J. Membr. Biol.* 229:115-30.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25:4876-82.

Treptow, N.A. & Shuman, H.A. (1985) Genetic evidence for substrate and periplasmic-binding-protein recognition by the MalF and MalG proteins, cytoplasmic membrane components of the Escherichia coli maltose transport system. *J. Bacteriol*. 163:654-60.

Tusnady, G.E. & Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*. 17:849–850.

Vitreschak, A.G., Mironov, A.A., Lyubetsky, V.A. & Gelfand, M.S. (2008) Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA*. 14:717-35.

Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*. 20:44-50.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. & Gelfand, M.S. (2002) Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res*. 30:3141-51.

Wang, B., Dukarevich, M., Sun, E.I., Yen, M.R. & Saier, M.H., Jr. (2009) Membrane Porters of ATP-Binding Cassette Transport Systems Are Polyphyletic. *J. Membr. Biol.* 231:1-10.

Xu, C., Zhou, T., Kuroda, M. & Rosen, B.P. (1998) Metalloid resistance mechanisms in prokaryotes. *J. Biochem*. 123:16-23.

Yen, M.R., Choi, J. & Saier, M.H., Jr. (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. *J. Mol. Microbiol. Biotechnol*. 17:163-76.

Yen, M.R., Chen, J.S., Marquez, J.L., Sun, E.I. & Saier, M.H. (2010) Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. *Methods Mol. Biol*. 637:47-64.

Zhai, Y. & Saier, M.H., Jr. (2002) A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. *J. Mol. Microbiol. Biotechnol*. 4:375-7.

Zhai, Y. & Saier, M.H., Jr. (2001a) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J Mol Microbiol Biotechnol*. 3:501-2.

Zhai, Y. & Saier, M.H., Jr. (2001b) A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. *J. Mol. Microbiol. Biotechnol*. 3:285-6.

Zhai, Y., Tchieu, J. & Saier, M.H., Jr. (2002) A web-based Tree View (TV) program for the visualization of phylogenetic trees. *J. Mol. Microbiol. Biotechnol*. 4:69-70.

Zhang, P., Wang, J. & Shi, Y. (2010) Structure and mechanism of the S component of a bacterial ECF transporter. *Nature*. 468:717-20.