Evolutionary Bioinformatics

ORIGINAL RESEARCH

Phylogenetic Analysis and Comparative Genomics of Purine Riboswitch Distribution in Prokaryotes

Payal Singh[1] and Supratim Sengupta[1,2]

[1]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi-110067, India.
[2]Department of Physical Sciences, Indian Institute of Science Education and Research, Kolkata, Mohanpur Campus, Mohanpur-741252, India. Corresponding author email: supratim.sen@gmail.com

Abstract: Riboswitches are regulatory RNA that control gene expression by undergoing conformational changes on ligand binding. Using phylogenetic analysis and comparative genomics we have been able to identify the class of genes/operons regulated by the purine riboswitch and obtain a high-resolution map of purine riboswitch distribution across all bacterial groups. In the process, we are able to explain the absence of purine riboswitches upstream to specific genes in certain genomes. We also identify the point of origin of various purine riboswitches and argue that not all purine riboswitches are of primordial origin, and that some purine riboswitches must have originated after the divergence of certain Firmicute orders in the course of evolution. Our study also reveals the role of horizontal transfer events in accounting for the presence of purine riboswitches in some gammaproteobacterial species. Our work provides significant insights into the origin, distribution and regulatory role of purine riboswitches in prokaryotes.

Keywords: riboswitch, evolution, phylogenetics, comparative genomics, horizontal transfer

Evolutionary Bioinformatics 2012:8 589–609

doi: 10.4137/EBO.S10048

This article is available from http://www.la-press.com.

Introduction

Riboswitches are a type of regulatory ribonucleic acid (RNA) that modulates expression of downstream genes involved in ligand biosynthesis by undergoing a conformational change on ligand binding. They are typically found in the 5' untranslated regions (UTRs) of primarily prokaryotic messenger (m)RNAs. However, some riboswitches have been found in a few eukaryotes as well.[1,2] Riboswitches are composed of two domains, an aptamer domain that contains the site for ligand binding, and an expression platform, which can switch between two conformations depending on whether the ligand is bound or unbound to the aptamer domain. They can control gene expression either by termination of transcription[3,4] or by inhibiting translation initiation[5] by sequestering the ribosomal binding site in the mRNA. Several classes[6] of riboswitches have been discovered, with the classes being distinguished by the metabolite that binds to the riboswitch. The aptamer domain is highly conserved both at the sequence and the structural level for riboswitches belonging to the same class, whereas the expression platform is highly variable even among riboswitches of the same class. In a recent work,[7] we exploited the high level of sequence conservation of the aptamer domain for a given class of riboswitches in order to develop a fast and accurate method of riboswitch classification based on profile Hidden Markov Models (pHMMs).

Riboswitches regulate various important biochemical pathways in response to the intracellular metabolic concentration, and are widespread in pathogenic bacteria, which makes them a promising drug target.[8,9] Natural as well as rationally designed structural analogs that mimic a riboswitch binding ligand have the potential to bind riboswitches and regulate gene expression. Such compounds with antimicrobial action have already been designed for TPP,[10] lysine[11] and purine[12,13] riboswitches. The possibility that riboswitches are promising drug targets is further emphasized by the finding that even natural antibiotics can act by targeting riboswitches.[14,15] Various riboswitch-based artificial regulatory systems have also been engineered to modulate ligand-dependent gene expression.[16,17] Complex riboswitches, which act according to Boolean logics and quick-responding two-way gene control systems have also been designed.[18,19] As more classes of riboswitches are discovered,[20] a detailed understanding of the riboswitch structure and function will be crucial in designing riboswitches that can precisely control the concentration of corresponding metabolites and thereby affect the functioning of an organism.

Riboswitches are also speculated[21] to be the remnants of RNA-based metabolite sensors that may have been present in an RNA world. Evidence for that hypothesis requires identifying the point of origin of riboswitches. Moreover, analyzing the distribution pattern of riboswitches across different prokaryotic genomes and the genes they regulate would shed light on the evolution of riboswitches.

The work of Rodionov et al[22,23] on the comparative genomics of thiamine biosynthesis and vitamin B_{12} metabolism has provided considerable insight into the nature of thiamin and cobalamin biosynthesis genes that are regulated by their corresponding riboswitches. More recently, Barrick and Breaker[24] carried out a study of the distribution of various classes of riboswitches across different bacterial groups. However, in order to better understand the origin and evolution of riboswitches, it is essential to first obtain a detailed picture of riboswitch distribution across all prokaryotic genomes. This has to be done not just for each riboswitch class, but also for each distinct gene (or operon) that the riboswitch of a given class regulates. By analyzing this distribution pattern and identifying the genes (or operons) regulated and their role in the metabolic pathways of ligand biosynthesis, it is possible to acquire a better understanding of the role played by riboswitches in gene regulation.

The aim of this paper was to carry out a comprehensive analysis of purine riboswitch distribution across prokaryotic genomes. In the process, we were able to identify the point of origin of the various purine riboswitches, correlate the presence of purine riboswitches with the presence and nature of genes that they regulate, as well as the metabolic pathways to which these genes belong. In some instances, we even found evidence of horizontal transfer of purine riboswitches across distant prokaryotic phyla, which nevertheless share the same environmental niche. Our work provides the first detailed analysis of the origin, evolution and comparative genomics of purine riboswitches.

Methods

Genomic sequence data retrieval and categorization

The Refseq database was downloaded from the National Center for Biotechnology Information (NCBI) FTP site. A total of 646 completely sequenced bacterial genomes were extracted from the Refseq database. Since in this study we are looking at the distribution and evolution of purine riboswitches across different taxonomic groups, the genomes were categorized into different phylums on the basis of taxonomy. Perl scripts were used to extract and categorize the genomic data.

Riboswitch Identification

A profile Hidden Markov Model (pHMM)-based method[7] of riboswitch identification was used to identify riboswitches in the bacterial genomes. A purine riboswitch-specific pHMM[7] was used to screen the bacterial genomes. A systematic analysis of the genomic context of purine riboswitches was carried out to identify the genes upstream to which riboswitches occur. The genomes with missing riboswitches were analyzed in detail to determine the precise cause of the missing riboswitches. The UTRs of the genes with missing riboswitches were also scanned using other riboswitch detection tools like Riboswitch Finder, RibEx, and Covariance Model to make sure that the riboswitches were really absent in such cases and not missed by the pHMM based detection method. This analysis did not reveal any instances of a purine riboswitch that was not detected by our pHMM method.

Phylogenetic analysis

Purine riboswitch identification in bacterial genomes pertaining to different taxonomic groups reveals that they are present predominantly in Firmicutes. With an aim to gain insight on the evolution and point of origin of different purine riboswitches, the riboswitch occurrence information was mapped onto the phylogenetic tree of the organisms belonging to the phylum Firmicutes. Out of the 646 completely sequenced genomes obtained from the Refseq database, 136 belonged to Firmicutes. For phylogenetic tree construction twenty protein families were selected. The choice of the protein sequences was made on the basis of the work of Ciccarelli and colleagues[25] that was aimed at building a highly resolved tree of life. The list of proteins is given in Supplementary file 1. The protein families selected for phylogenetic construction was the subset of the proteins used by Ciccarelli and colleagues.[25] Only those proteins that were found in all the 136 Firmicute species were used to build the sub-set. The twenty clusters of orthologous groups were selected in such a way so as to exclude any lateral transfers,[25] which is essential for obtaining a highly resolved tree.

The protein sequences corresponding to each COG were extracted from all 136 species and aligned using MUSCLE.[26] After alignment, the poorly-aligned regions with more than 20% gaps were trimmed using trimAl.[27] The aligned sequences were concatenated to produce a super-gene alignment of 4975 positions which was then used to build a phylogenetic tree using neighbor-joining (NJ) as well as maximum likelihood (ML) methods. The NJ tree was generated with Phylip version 3.69[28] using a series of sequentially executed programs. Seqboot was used to generate the data sets replicates from the alignment file. Then Protdist generated a distance matrix using the JTT model. The program Neighbor, was used to construct phylogenetic trees from each data set using the Neighbor-Joining method and the consensus tree, was built using the Consense program. The NJ tree was generated for 100 (see Figure 1) as well as 1000 bootstrap replicates (see Supplementary file 2). The two trees are consistent with one another except for some rearrangements in some of the late branches that do not affect our conclusions. This is evident from the mapping of the points of origin of the various purine riboswitches onto the NJ tree with 100 and 1000 (see Supplementary file 2) bootstrap replicates respectively.

For the ML tree, the best-fit amino acid evolution model was selected for the alignment using the ProtTest3[29] program, which uses PhyML[30] for likelihood calculations with Akaike Information Criterion (AIC). The LG[31] substitution model with invariant sites (I) and four gamma distribution (G_4) rate categories was determined to be the most appropriate for the given alignment. Maximum likelihood phylogenetic trees were generated with the substitution model, LG+I+G_4 using PhyML version 3.[30] The reliability of the trees was evaluated using 1000 bootstrap replicates. Finally the consensus tree was build using the Consense program from the Phylip package.

The ML tree groups the *Staphylococcus* genus along with the *Lactobacillales* with very low bootstrap support. This grouping is not consistent with the existing taxonomy,[32] according to which *Staphylococcus* should cluster with *Bacillales*, indicating that the NJ tree is more reliable than the ML tree in this case. The ML tree, illustrating the point of origin for different purine riboswitches, is represented in Supplementary file 3. A comparison of the NJ (Figure 1) and ML (Supplementary file 3) trees indicate that our conclusions about the origin of the most purine riboswitches (except riboswitches upstream to the COG2233 and COG1972 genes) are unaffected by the difference between the NJ and ML trees. For the ML tree, the point of origin of the purine riboswitch upstream to the COG2233 gene in the Clostridia group is located on a branch that does not include *C.novyi NT.* However, this change relative to the NJ tree does not affect our conclusions in any way as *C.novyi NT* does not possess the COG2233 gene. Also, the point of origin of the riboswitch upstream to the COG1972 gene is located in a branch that includes the *Geobacillus* species. It is worth noting that *Geobacillus* species lacks the COG1972 gene.

The Firmicute tree constructed here separates the different orders clearly with sufficiently high bootstrap values except for the class Clostridia. One reason for this may be that the members of this group are paraphyletic and do not form a phylogenetically coherent group.[33] Some of the late branches have

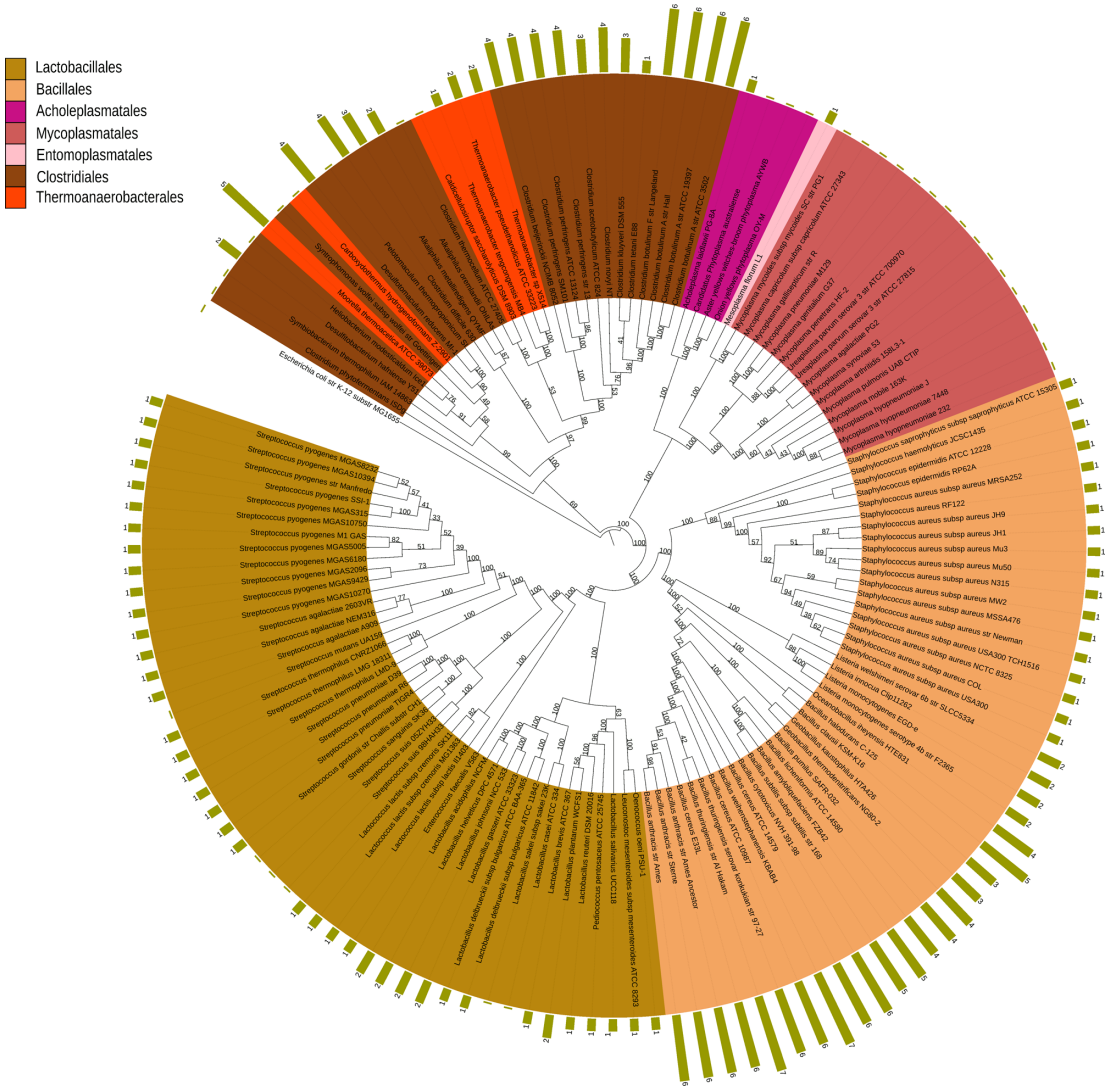

Figure 1. Phylogenetic distribution of purine riboswitches in Firmicutes.
Note: The bar length represents the number of riboswitches in each genome, which is indicated at the end of each bar.

low bootstrap support as can be seen from Figure 1. However, these pertain to evolutionary relationships within specific families, and do not affect the interpretation of our results.

To demonstrate that some of the riboswitches found in Gammaproteobacteria were horizontally transferred from the Firmicute phylum, phylogenetic analysis of the *ADA* gene was carried out. For this, the *ADA* gene was identified in Firmicutes and Gammaproteobacteria. The protein sequence for each identified gene was extracted. The sequences were then aligned using MUSCLE. The columns in the alignment with more than 20% gaps were removed using trimAl. The maximum likelihood phylogenetic tree with 1000 bootstrap replicates was constructed using the PhyML program, under the best-fit protein evolution model as selected by ProtTest ($LG+I+G_4$).

The mapping of the riboswitch distribution on the phylogenetic tree of Firmicutes was visualized using iTOL.[34]

Results

A search of the Refseq database for purine riboswitches found 263 candidates. The distribution of purine riboswitch indicates that they are widespread in Firmicutes but rare in other bacterial groups. In Firmicutes, all classes except Mollicutes use the riboswitch mode of regulation extensively. In certain genera like Bacillus and Clostridium, riboswitches occur multiple times per genome. Figure 1 gives the phylogenetic distribution of all of the purine riboswitches in Firmicutes, and Table 1 indicates the nature of the gene/operon that the purine riboswitches regulate. Analysis of the upstream genes around which purine riboswitches are found reveal that the riboswitch regulation for the purine salvage pathway genes and the transporter genes is widespread in Firmicutes. However, in a few cases, purine riboswitches have also been found to regulate transcription factors as well as the genes belonging to the de-novo synthesis of inosine monophosphate (IMP). Such instances are found primarily in the family Bacillaceae, which has the largest number of purine pathway genes under the riboswitch mode regulation. The riboswitches in these organisms are distributed upstream to a diverse set of genes, which include salvage and de-novo pathway genes as well as permeases, transcription factors and transporters. Organisms from the class Clostridia also regulate a variety of genes comprising transporters, salvage and de-novo genes for IMP synthesis via riboswitch. Riboswitches also regulate guanine monophosphate (GMP) synthesis in many organisms from the Bacillales and Clostridia groups. Streptococcaceae, Lactobacillaceae, Leuconostocaceae, *Listeriaceae* and Staphylococcaceae families possess fewer instances of purine riboswitch-regulated genes. The riboswitches are found upstream of only salvage genes and transporters in the organisms belonging to these families. The organisms belonging to the Mycoplasmatales and Acholeplasmatales orders lack the purine biosynthetic pathway, and in most cases, also lack purine-specific permeases and do not have any purine riboswitches. This can be attributed to the fact that these organisms are parasitic in nature and depend on their hosts for purine metabolism.

Outside Firmicutes, the presence of purine riboswitches was restricted to few members of the families of Thermotogaceae, Fusobacteriaceae, Shewanellaceae, Vibrionaceae and Bdellovibrionacea. The presence of purine riboswitches in a few organisms belonging to these families can be attributed to horizontal transfer of the purine riboswitch, along with the gene it regulates from the members belonging to the Firmicute group where purine riboswitches are widespread.

In the sub-sections below, we analyze the phylogenetic distribution of riboswitches found upstream to different genes involved in purine metabolic pathway and purine transportation. The riboswitches are named after the gene upstream of which they occur. If a riboswitch occurs upstream of an operon, then it is named after the first gene in the operon.

Xanthine phosphoribosyltransferase (XPT) riboswitch

The *xanthine phosphoribosyltransferase (XPT)* riboswitch is widespread in Firmicutes. Almost all the Firmicutes that carry the *XPT* gene also possess the purine riboswitch upstream to it. *XPT* is a purine salvage pathway enzyme which converts the pre-formed base xanthine, a product of nucleic acid breakdown, to *xanthosine 5′-monophosphate (XMP)*, for reutilization in RNA or DNA synthesis. In most of the cases, the *XPT* gene is organized as an operon with the *pbuX* gene, which encodes xanthine-specific permease. Figure 2 shows the phylogenetic distribution of the purine riboswitch found upstream to the *XPT* gene (or operon containing *XPT* gene) in Firmicutes.

Table 1. Gene/operon under riboswitch regulation.

Order	Gene/operon	Pathway
Lactobacillales	Xpt-pbuX	Salvage
	Transporter protein (COG2252)	Transporter
	Adenine deaminase	Salvage
	Inosine-5′-monophosphate dehydrogenase (guaB)	Salvage
	PbuX (COG2233)	Transporter
Bacillales	Xpt-pbuX	Salvage
	Guanine hypoxanthine permease (COG2252)	Transporter
	NupC family nucleoside transporter (COG1972)	Transporter
	GMP synthase (guaA)	Salvage
	Pur operon	de-novo
	GntR family transcriptional regulator	Transcription factor
	PNP family	Salvage
	Hypoxanthine efflux transporter (COG2814)	Transporter
	Xpt-permease (COG2252)	Salvage
	Amidohydrolase	Salvage
	xpt-pbuX-guaB-guaA	Salvage
Clostridiales	pbuX-xpt	Salvage
	guaB-guaA	Salvage
	Pur operon	de-novo
	PbuX (COG2233)	Transporter
	Xanthine/uracil permease family protein (COG2252)	Transporter
	Xanthine/uracil permease family protein (COG2252)-adenine Phosphoribosyltransferase	Transporter
	Xpt-bmp family membrane protein-ABC tranporter ATP binding protein-ABC transporter permease-ABC transporter permease	Salvage
	Xpt	Salvage
	Adenosine deaminase	Salvage
	FAD dependent oxidoreductase	Salvage
	Xanthine/uracil/VitC permease (COG2252)-pur operon	Transporter
	guaA	Salvage
	Basic membrane lipoprotein (COG1744)-ABC transporter-inner membrane translocator-inner membrane translocator-xpt	Transporter
	Adenine deaminase-aldehyde oxidase and xanthine dehydrogenase	Salvage
	adenine deaminase	Salvage
Thermoanaerobacterales	guaB-guaA	Salvage
	Xanthine/uracil/VitC permease (COG2252)-pur operon	Transporter
	Pur operon	de-novo
	Basic membrane lipoprotein (COG1744)	Transporter
	Xanthine/uracil/VitC permease (COG2252)	Transporter
	2Fe-2S binding protein-xanthine dehydrogenase(A)-xanthine dehydrogenase(B)-2Fe-2S binding protein-chlorohydrolase-molybdopterin dehydrogenase	Salvage
	Aldehyde oxidase and xanthine dehydrogenase-aldehyde oxidase and xanthine dehydrogenase molybdopterin binding-amidohydrolase	Salvage
Acholeplasmatales	Permease (COG2252)	Transporter
Entomoplasmatales	Permease (COG2252)	Transporter

All the *Streptococcus* species except *S.mutans UA159*, *S.sanguinis SK36*, *S.suis 98HAH33* and *S.suis 05ZYH33* possess the XPT riboswitch. *S.mutans UA159, S.suis 98HAH33* and *S.suis 05ZYH33* lack XPT gene which indicates that the corresponding riboswitch was lost along with the gene. However, *S.sanguinis SK36* carries the *XPT* gene but does not have a riboswitch upstream because its small UTR of 43 nucleotides is incapable of accommodating a purine riboswitch.

All the *Lactococcus* and *Enterococcus* species have a purine riboswitch preceding the *XPT* gene, which is the part of an operon with the nucleobase permease. However, the *Lactobacillus* species can be divided

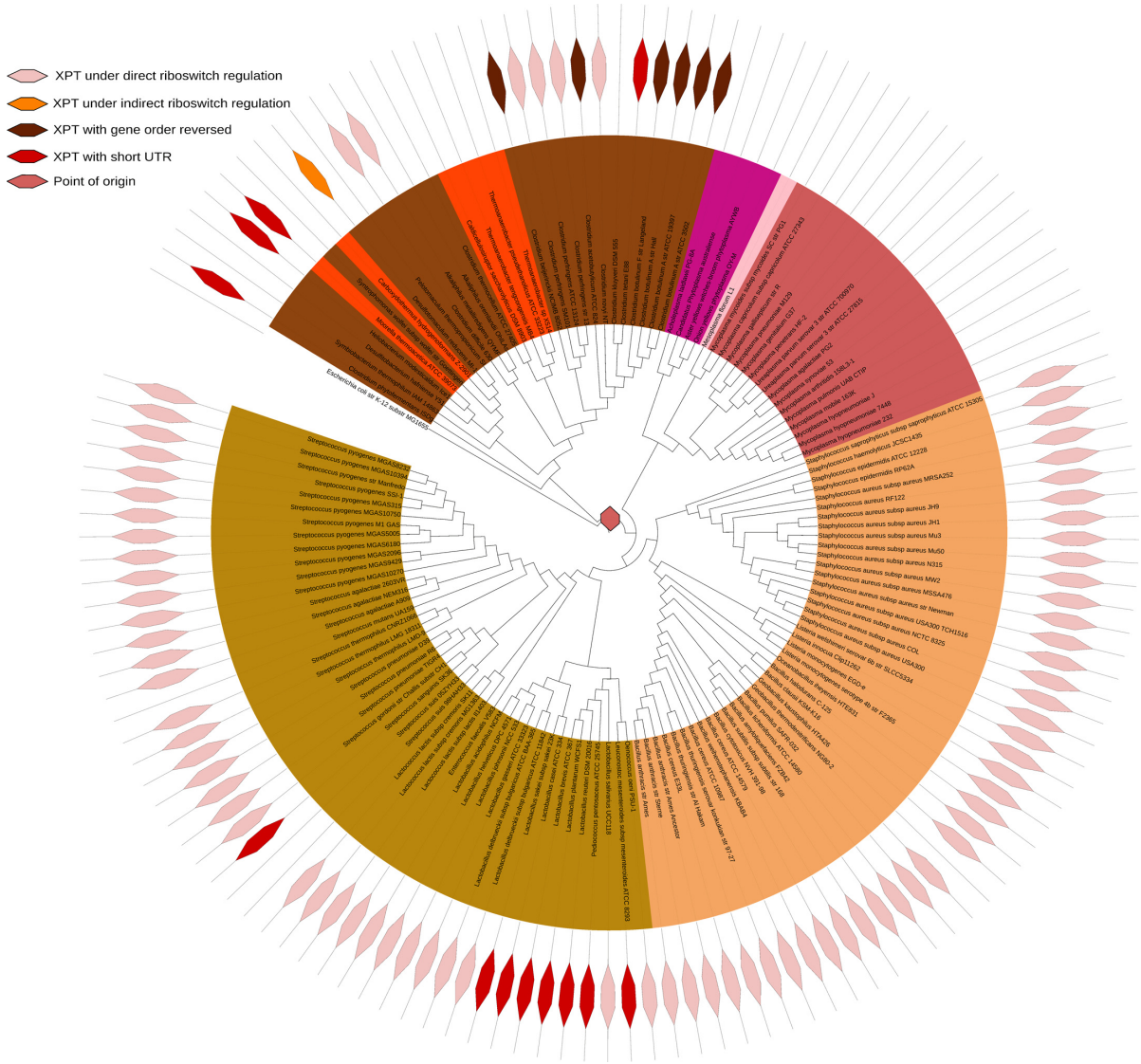

Figure 2. Phylogenetic distribution of the *XPT* riboswitch in Firmicutes.

into two groups, one with the *XPT-pbuX* operon that carries the riboswitch upstream to it and another group which possesses the *XPT* gene but not in the form of the usual *XPT-pbuX* operon and has a short UTR (<150 nucleotides) that lacks a riboswitch.

The distribution of the XPT riboswitch is more continuous in the organisms belonging to the order *Bacillales*. All the organisms of this order except *G.kaustophilus HTA426* have the *XPT* gene in an operon with a purine permease preceded by the riboswitch. (*G.kaustophilus HTA426* does not possess the *XPT* gene, either singly or as part of an XPT-pbuX operon. Moreover, the *pbuX* gene is truncated with just the C-terminal domain.) However, the *XPT* riboswitch distribution is fragmented in the order *Clostridiales*. The *XPT* gene is not organized in the form of *XPT-pbuX* operons in this order. Some of the *Clostridium* species (*C.botulinum, C.acetobutylicum* and *C.beijerinckii*) possess an operon in which the xanthine-specific permease precedes the *XPT* gene, which is the reverse of the gene order found in the *XPT-pbuX* operon of *Lactobacillales* and *Bacillales*. These *Clostridium* species have a purine riboswitch upstream to the xanthine permease, allowing regulation of the expression of the genes in the operon. Other organisms from the Clostridiales order that possess a purine riboswitch have the XPT gene occurring as a single unit, except in the case of *C.novyi NT* where *XPT* is part of an operon of five genes related to the basic membrane lipoprotein, ABC transporters and

permeases. Some *Clostridiales* species like *C.tetani E88, H.modesticaldum Ice1*, and *C.phytofermentans ISDg* have the *XPT* gene but lack the riboswitch upstream because of the small UTR. *D.reducens MI-1* possesses a riboswitch upstream to the operon containing the basic membrane lipoprotein ABC transporter and XPT. Yet other *Clostridiales* species that does not have the purine riboswitch also does not possess the *XPT* gene. Moreover, *Acholeplasmatales*, *Mycoplasmatales* and *Thermoanaerobacterales,* except *M.thermoacetica ATCC 39073* lack the *XPT* gene as well as the purine riboswitch.

Transporter riboswitch

In this section, we discuss the phylogenetic distribution and possible origin of riboswitches that regulate the transporter genes. Transporter proteins responsible for purine uptake are ubiquitous in Firmicutes. Five different transporters were found to be under riboswitch regulation in Firmicutes, namely *xanthine permease (pbuX), guanine hypoxanthine permease (pbuG), hypoxanthine efflux pump (pbuE), nucleoside transporter protein (nupC)* and, in couple of cases, *basic membrane protein*. The role of *pbuX, pbuG, pbuE* and *nupC* as purine transporters has been established in *Bacillus* subtilis.[35,36] The homologues for these genes are identified in all the Firmicutes using BLAST (sequences with bit score $\geq$ 200 are used as homologs in this study). The distribution of the transporters and the riboswitches that regulate them is shown in Figure 3. In some organisms, multiple homologs with different gene names are detected for different transporter classes; therefore, we have used the COG nomenclature instead of the gene name, to denote the distribution of transporter genes that are regulated by the purine riboswitch in Firmicutes. *pbuX, pbuG* and *pbuE* belong to COG2233, COG2252 and COG2814 respectively.

Xanthine permease (*pbuX*) from the COG2233 is widespread in Firmicutes and in most cases is regulated by a riboswitch. The sequences categorized in COG2233 belong to the category of nucleotide transport and metabolism. All *Lactobacillales* except *S.mutans UA159, S.sanguinis SK36, S.suis 05ZYH33 and S.suis 98HAH33* possess the *pbuX* gene. In *Lactobacillales,* the *pbuX* gene occurs either as an *XPT-pbuX* operon that encodes for purine salvage pathway proteins involved in the import and phosphorylation of xanthine, or as a single unit as in *L.brevis ATCC 367*, *L.plantarum WCFS1*, *L.reuteri DSM 20016* and *P.pentosaceus ATCC 25745* (exceptions are *L.casei* and *L.mesenteroides*). All *Lactobacillales* except *L.sakei 23 K*, *L.casei ATCC 334* and *L.mesenteroides ATCC 8293*, which carry the *pbuX* gene, regulate its expression either directly (by having the riboswitch upstream to the *pbuX* gene in case it occurs singly) or indirectly (by having the riboswitch upstream of the *XPT* gene, in case it occurs as an operon which also contains the *pbuX* gene) via a purine riboswitch. In *L.brevis, ATCC 367* and *P.pentosaceus ATCC 25745*, two copies of *pbuX* are found, however only one of them carries the riboswitch. The two *pbuX* genes in *P.pentosaceus ATCC 25745* are highly similar on both the protein as well as nucleotide level, suggesting that gene duplication may have resulted in the presence of two copies of the gene. Supplementary file 4 gives the sequence alignments of the *pbuX* genes in *P.pentosaceus ATCC 25745*.

In *L.casei* the UTR upstream of the operon containing the *pbuX* gene is small and thus cannot contain a riboswitch.

The *PbuX* gene is more prevalent in the order Bacillales. All the organisms of this order, except *B.clausii KSM K16* and *B.halodurans C-125*, carry the *pbuX* gene. In Bacillales, the *pbuX* gene always occurs as the part of an operon that is under riboswitch regulation (either as an *XPT-pbuX* operon or as an *XPT-pbuX-guaB-guaA* operon), except in the case of *G.kaustophilus HTA426*, in which only the C-terminal region of the *pbuX* is found. It seems plausible to surmise that the riboswitch along with the XPT and the N-terminal region of the *pbuX* was lost from this genome.

The distribution of the *pbuX* gene is more fragmented in Clostridiales. *PbuX* is subjected to a riboswitch mode of regulation in only a few organisms belonging to the order Clostridiales. *C.botulinum, C.perfringens, C.beijerinckii* and *C.acetobutylicum* possess a *pbuX* gene that is regulated by the purine riboswitch. In *C.botulinum, C.beijerinckii* and *C.acetobutylicum, pbuX* occurs as a two-gene operon along with the *XPT* gene; however the gene order is *pbuX-XPT*, which is the reverse of the gene order observed in *Lactobacillales* and *Bacillales*. *C.perfringens* possesses the riboswitch bearing *pbuX* gene as a single unit. All *C.botulinum* species except *C.botulinum F str. Langeland* carry two

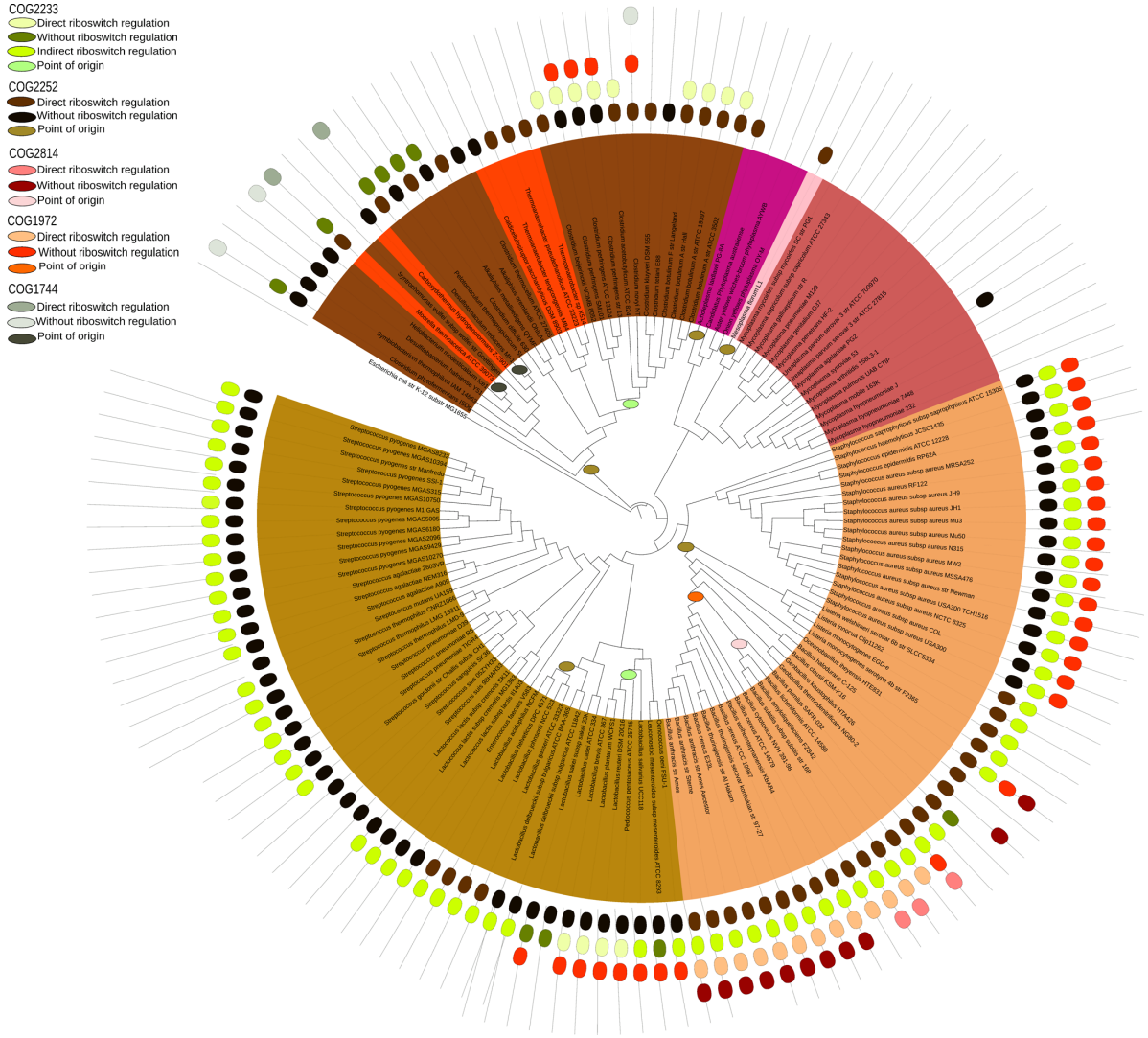

Figure 3. Phylogenetic distribution of transporter riboswitches in Firmicutes.

pbuX riboswitches, one upstream of the *pbuX-XPT* operon and another upstream of *pbuX,* which occurs singly. *C.difficile* 630*, A.metalliredigens QYMF, P.thermopropionicum SI, C.phytofermentans ISDg and D.reducens MI-1* possess the *pbuX* gene but lack the riboswitch upstream to it. Other organisms from *Clostridiales* do not have the pbuX gene altogether. Thermoanaerobacterales, Acholeplasmatales and Entomoplasmatales also lack the *pbuX* gene. In Figure 3, different color coding is used to distinguish the *pbuX* gene that is under direct riboswitch regulation from the one in which the *pbuX* gene is the second gene of a riboswitch-regulated operon.

Another class of transporter under riboswitch regulation belongs to the COG2252. The genes classified as belonging to the COG2252 are widespread in Firmicutes. COG2252 includes permeases for diverse substrates such as xanthine, uracil and vitamin C. Many members of this family are functionally uncharacterized and may transport other substrates also. All organisms of the order *Lactobacillales*, *Bacillales*, Clostridiales (except Syntrophomonas wolfei) and Thermoanaerobacterales possess the genes belonging to the COG2252. In *Lactobacillales*, the clade with the organisms *L.acidophilus NCFM*, *L.helveticus DPC* 4571*, L.johnsonii NCC* 533 and *L.gasseri ATCC* 33323 regulates one of the three COG2252 genes via the purine riboswitch. The three genes in each of the above genomes are highly similar (blast score $>$ 200 bits), suggesting that gene duplication may have been responsible for the presence of multiple copies of the gene. Supplementary file 5 gives the multiple sequence alignments of the

COG2252 gene in each of the above species. All of the organisms from *Bacillaceae* and *Listeriaceae* families employ the riboswitch mode of regulation for the transporter gene categorized as belonging to COG2252. In Bacillus subtilis, the COG2252 gene, under riboswitch regulation, has been characterized as a guanine/hypoxanthine permease. Homologues of the gene belonging to the COG2252 were found in many organisms from the orders *Lactobacillales* and *Bacillales*. However, only one of the genes homologous to COG2252 possesses the riboswitch. The COG2252 includes permeases for diverse substrates; thus it may be possible that the homologues not regulated by the purine riboswitch may have specificity for substrates other than purines. COG2252 are the most common riboswitch-regulated transporters found in Clostridiales. *C.botulinum*, *C.kluyveri, C.novyi*, *C.beijerinckii, A.oremlandii, C.difficile* and *D.reducens* possess a riboswitch upstream to the transporters from the COG2252. In a few cases like *C.botulinum, C.novyi* and *C.difficile*, more than a single gene from the COG2252 was regulated by riboswitches. Purine transport in *C.botulinum* is under tight regulation by riboswitches and four of the six riboswitches found in this organism were upstream to the transporter genes. In few organisms, riboswitch-regulated transporters occur as an operon with purine-salvage or de-novo-pathway genes, and hence also regulate their expression. In *C.botulinum*, one of the riboswitch-regulated COG2252 genes occurs as an operon with adenine phosphoribosyltransferase. *A.oremlandii OhILAs* possess a single transporter riboswitch from COG2252 that occurs as an operon with the purine de-novo pathway genes. Three transporter riboswitches were detected in *D.reducens MI-1*. Two of them were upstream to the permease from COG2252, and the third one was found upstream to the basic membrane lipoprotein which belongs to COG1744. The basic membrane lipoprotein occurs as an operon with the *ABC* transporter genes and *XPT*. One of the COG2252 transporters bearing riboswitches also occurs in an operon with the purine de-novo pathway genes. Therefore both *XPT* and the *pur* operon were under indirect riboswitch regulation in this case.

All Thermoanaerobacterales, except *C.saccharolyticus DSM* 8903 and *C.hydrogenoformans Z*-2901 possess the riboswitch bearing the COG2252 gene. This gene occurs in an operon the purine de-novo pathway genes in the case of the *Thermoanaerobacter* species. In Acholeplasmatales, Mycoplasmatales and Entomoplasmatales, only *M.florum L1, A.laidlawii PG-8 A* and *M.pulmonis UAB CTIP* possess the COG2252 gene. Both *M.florum L1* and *A.laidlawii PG-8 A* regulate this class of transporter through a riboswitch.

The organisms from the family *Bacillaceae* possess the maximum number of different transporters under riboswitch regulation. *nupC*, a nucleoside transporter protein (belonging to the COG1972) and *pbuE*, a hypoxanthine efflux pump, along with *pbuX* and *pbuG,* were found to be under the riboswitch mode of regulation. *nupC* is a Concentrative Nucleosides Transporter (CNT) family protein that mediates nucleoside uptake. The riboswitch upstream to the *nupC* gene is found specifically in the Bacillus subclade comprising of organisms *B.anthracis, B.cereus, B.thuringiensis, B.weihenstephanensis, B.cytotoxicus, B.subtilis, B.amyloliquefaciens,* and *B.licheniformis.* The exception is *B.pumilus,* where the *nupC* gene is part of a two gene operon with the pyrimidine nucleoside phosphorylase, a pyrimidine salvage gene. However, the operon does not carry a riboswitch upstream to it. This subclade has a minimum of three and a maximum of eight homologues of *nupC*, of which only one bears the riboswitch. The *Bacillaceae* organisms outside of this clade either have none, or possess only one homologue of *nupC* and no corresponding purine riboswitch.

Riboswitches upstream to transporter genes belonging to COG1744 and COG2814 are relatively rare, as is evident from Figure 3. *pbuE* is the hypoxanthine efflux transporter that belongs to the Major Facilitator Superfamily (MFS) associated with COG2814. This family includes a large and diverse group of transporters. The *pbuE* gene was found to be under riboswitch regulation in the Bacillus subclade, consisting of organisms *B.subtilis, B.amyloliquefaciens, B.pumilus*. However, *B.licheniformis* from this clade lacks the *pbuE* purine riboswitch. Amongst the *Bacillaceae* family, *B.subtilis* and *B.amyloliquefaciens* possess the maximum number of transporters subject to the riboswitch mode of regulation.

Pur operon riboswitch

The de-novo biosynthetic pathway for purine nucleotides is highly conserved and well-represented in all

the three domains of life, thereby leading to the suggestion that it was present in the last common ancestor.[37] However, the organization of the genes encoding the enzymes for the pathway and their regulation vary.[38] The pathway has been well studied in organisms like *Escherichia coli, Bacillus subtilis* and *Saccharomyces cerevisiae*. In *E.coli,* the purine de-novo genes occur as single unit or as small operons scattered across the genome.[38] In *B.subtilis,* the genes for de-novo synthesis are organized in a single operon.[39]

The genes for the de-novo pathway encode the enzymes for the inosine monophosphate (IMP) biosynthesis. IMP acts as the common intermediate in the inter-conversion between adenine and guanine.[40] This enables the cell to maintain the desired composition of the nucleotide pool.

In Bacillus, the de-novo genes for purine biosynthesis are organized as 12-member *purEKBC-SQLFMNHD* operon which encodes all enzymes required to convert phosphoribosylpyrophosphate (PRPP) to IMP. The first gene of the operon carries a purine riboswitch upstream to it. The *pur* operon in Bacillus is known to be regulated by two different mechanisms. Transcription initiation is repressed by excess adenine/adenosine. This is mediated by the purR protein repressor, which binds to the operator region of DNA and prevents transcription of the downstream genes. Alternatively, excess guanine/guanosine promotes transcription termination of the *pur* operon mediated by the purine riboswitch. A set of genes or operons regulated by *purR* repressor constitute the *purR* regulon, which consists of *XPT-pbuX, pur* operon*, purA, pbuE, guaC, pbuO*, *pbuG, glyA* and *nupG* genes.[35,36] Thus the *pur* operon and the *XPT-pbuX* operon are subjected to dual mode of regulation both by a protein repressor (*purR*) as well as by a riboswitch in *Bacillus subtilis*. *Listeria* and *Staphylococcus* also have the same organization of the de-novo genes (*purEKBCSQLFMNHD*) but lack the riboswitch upstream to it.

The distribution of the *pur* regulated riboswitches is shown in Figure 4. In *Streptococcus* the de-novo genes are organized as the *pur* cluster *(purCLFMN-vanZ-purH, purD and purEKB*), while in *Lactococcus*, the *pur* genes are separated into several unlinked clusters: *purDEK*, *purCSQL*, *purMN*, and *purH.* The purR homologue in *Streptococcus* seems to act as a repressor of the transcription of all the genes in the *pur* cluster.[41] However, the *purR* homologue in *L.lactis* seems to act as an activator of *purC* and *purD* expression.[42] In *Lactobacillus, Enterococcus and Leuconostoc* all the de-novo genes are present but organized in a single operon. Exceptions are seen in the case of *L.johnsonii, L.gasseri* and *L.brevis,* where the de-novo pathway genes for purine biosynthesis are missing. However, none of the organisms from the order *Lactobacillales* possess a purine riboswitch upstream to the de-novo genes.

In *Clostridia,* the genes for de-novo synthesis of purine are organized as a single operon. However, the order of genes in the operon can vary across organisms belonging to this class. (For instance, in *C.perfringens, C.beijerinckii* and *D.hafniense* the first gene of the operon is *PurL* instead of *PurE*.) As is evident from Figure 4, only a few organisms belonging to this class carry a purine riboswitch upstream to the *pur* operon.

Thermoanaerobacter species, D.reducens and *A.oremlandii* have purine de-novo pathway genes arranged in an operon with the first gene being a permease that carries riboswitch upstream to it. Hence, the purine riboswitch in these organisms is characterized as a riboswitch upstream to a permease (which is marked as a *pur* operon under indirect riboswitch regulation in Fig. 4), instead of a riboswitch upstream to a de-novo pathway gene, even though the latter are also regulated by the riboswitch.

guaA and *guaB* riboswitches

The de-novo purine biosynthesis reactions convert PRPP to IMP, which is the first purine nucleotide and acts as a common purine precursor. Inosine monophosphate can either synthesize *adenosine monophosphate (AMP)* or *guanosine monophosphate (GMP)* via separate, two-step pathways. Inosine monophosphate is converted to guanosine monophosphate by the oxidation of IMP to *xanthine monophosphate (XMP)*, which is catalyzed by *IMP dehydrogenase (guaB)*, followed by the amination of *XMP* to *GMP* catalyzed by *GMP synthase* (*guaA*). *IMP* dehydrogenase and *GMP* synthase are the two key enzymes in the purine salvage pathway.[43] These genes are present in all Firmicutes except in parasites like mycoplasma and phytoplasma, where all the purine biosynthesis genes are missing. The two enzymes synthesizing *GMP* from *IMP* are organized differently in different groups across Firmicutes. All the *Lactobacillales* except *L.delbrueckii*

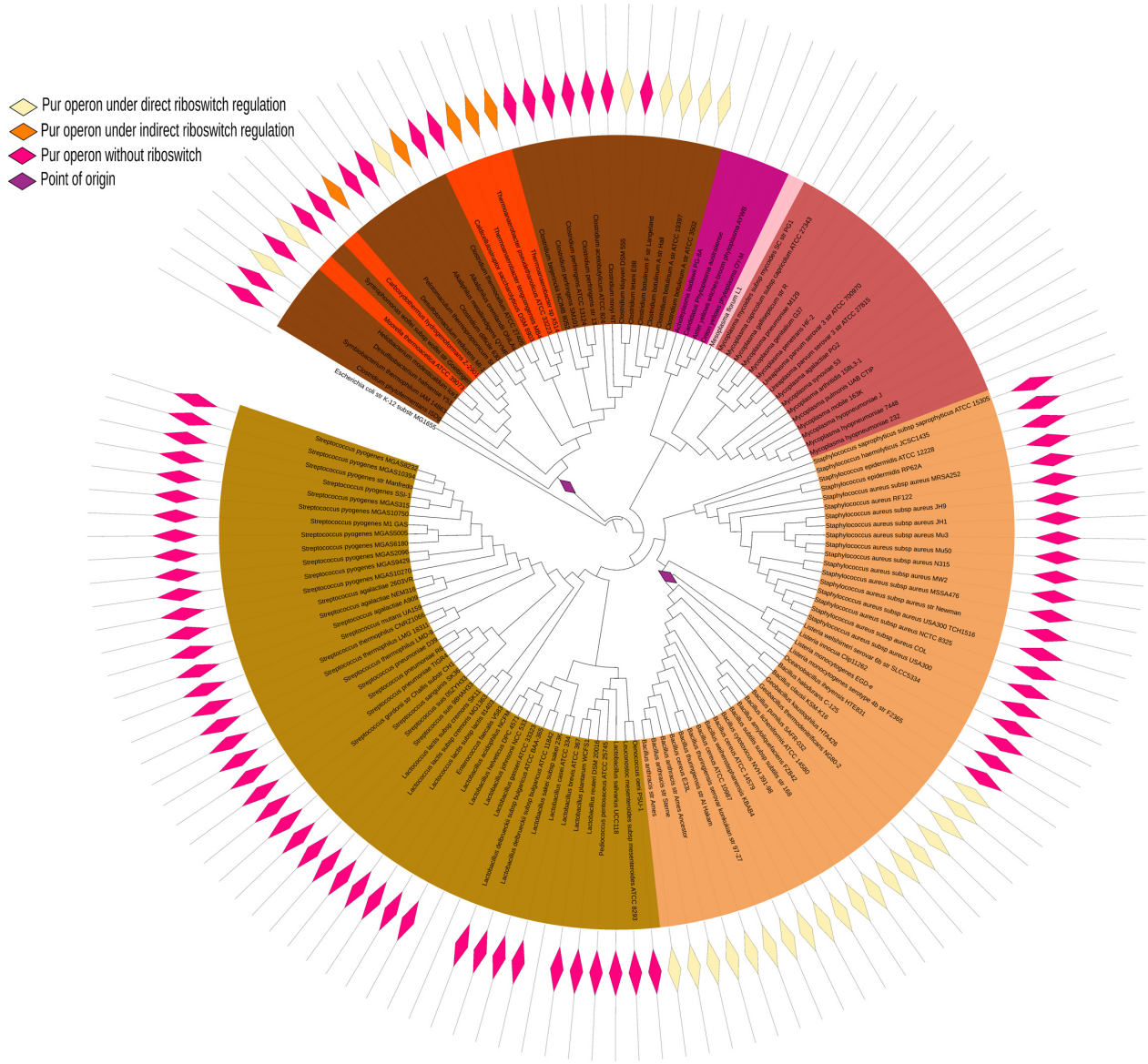

Figure 4. Phylogenetic distribution of the *pur* operon riboswitch in Firmicutes.

have these genes as single units. In *L.delbrueckii*, the IMP dehydrogenase and GMP synthase are organized in an operon. This is indicated in Figure 5 by the placement of the blue rectangles closer to the purple rhombus. None of the organisms belonging to the order *Lactobacillales* are under riboswitch regulation except *L.mesenteroides*, which has riboswitch upstream to the IMP dehydrogenase gene. The *guaA* gene is missing in this organism. All the organisms belonging to the order Bacillales except the *Staphylococcus* species have these genes as single unit. In *Staphylococcus,* these genes occur as the part of the four-gene operon *XPT-pbuX-guaB-guaA* which is regulated by the purine riboswitch. This is indicated in Figure 5 by the placement of the cyan rectangle closer to the orange pentagon in all *Staphylococcus* species.

Clostridiales possess *guaA* and *guaB* either as a two-gene operon (*guaB-guaA*) or as single units. The operon in *Clostridiales* is always regulated by the riboswitch and this is indicated in Figure 5 by the placement of the cyan rectangle (denoting indirect regulation) closer to pink pentagon. In *Clostridiales*, only *C.difficile* possesses a *guaA* riboswitch, which is not the part of an operon.

In Thermoanaerobacterales, only the *Thermoanaerobacter* species possess the operon which is organized as *guaB-guaA*. In other species the *guaA* and *guaB* occur as discrete units. Only two

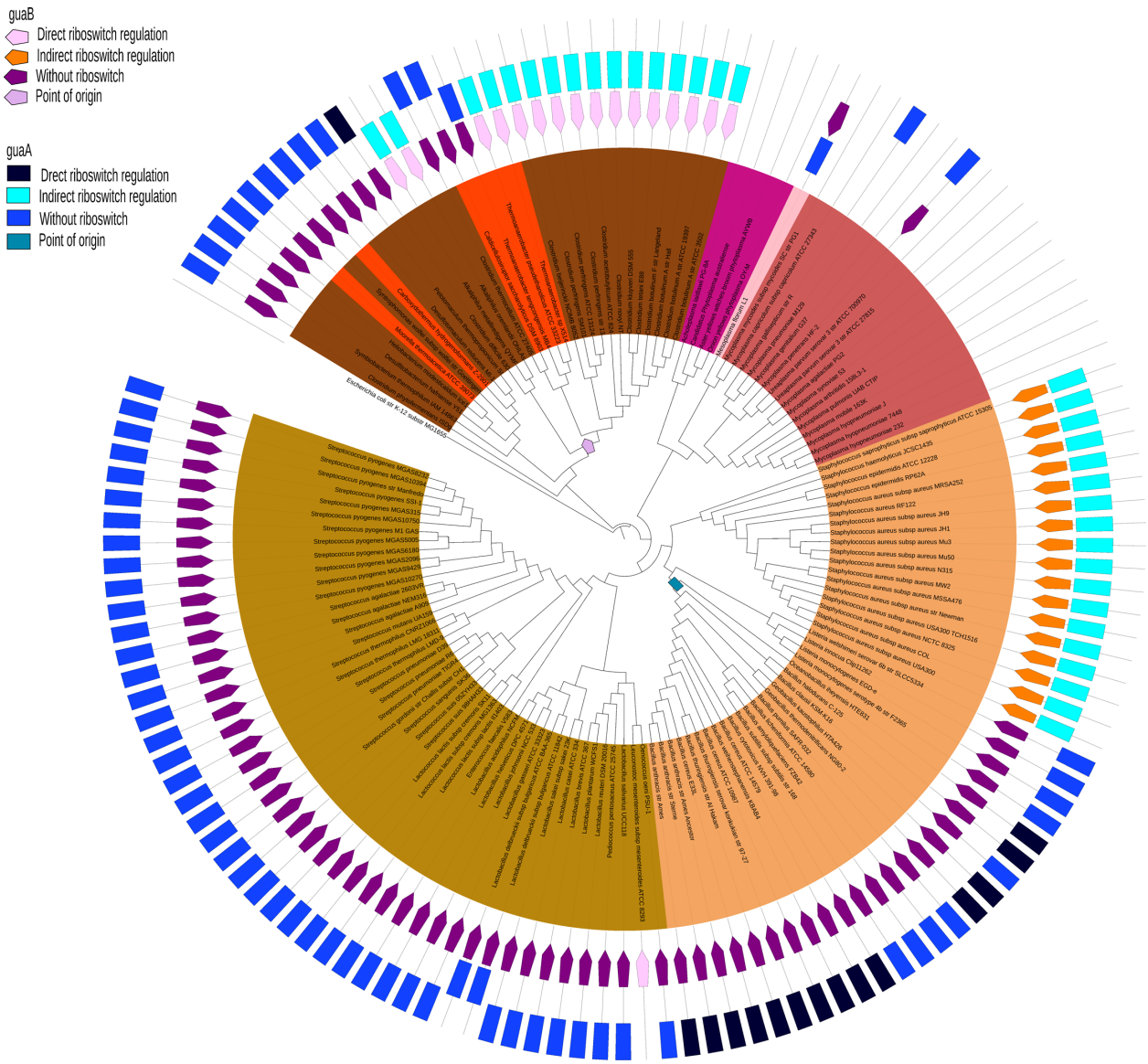

Figure 5. Phylogenetic distribution of *guaA* and *guaB* riboswitches in Firmicutes.

Thermoanaerobacter species, *Thermoanaerobacter sp X514* and *Thermoanaerobacter pseudethanolicus,* regulate the operon via the purine riboswitch.

GMP synthase gene is subject to the riboswitch mode of regulation in all the organisms belonging to the *Bacillaceae* family except *B.clausii, B.pumilus, B.licheniformis, B.amyloliquefaciens* and *B.subtilis.* These organisms possess short UTRs (having less than 200 nucleotides), which may account for the absence of the riboswitch.

GntR riboswitch

GntR is a family of bacterial transcription regulators. The transcription factors of this family have an N-terminal DNA-binding domain, a C-terminal effector-binding domain, and/or an oligomerization (E-b/O) domain. The DNA-binding domain is well conserved. However, the effector-binding domain is variable and heterogeneous, on the basis of which GntR regulators are classified into different subfamilies.[44] The distribution of riboswitches that regulate GntR is shown in Figure 6. In Bacillales, a specific subclade comprising of *B.anthracis, B.cereus, B.thuringiensis, B.weihenstephanensis* and *B.cytotoxicus* possess a riboswitch regulating the GntR transcription factors. These riboswitches appear to have originated at the root of this clade and regulate the expression of the GntR regulators.

Other riboswitches

There were some riboswitches that were not widespread but appeared in a few organisms across the Firmicutes, as shown in Figure 6. All of these riboswitches regulate the genes categorized as the salvage-pathway genes.

Adenine deaminase (ADE) catalyzes the deamination reaction of adenine to hypoxanthine, which is important for adenine utilization as purine and nitrogen source. This reaction is also essential for the conversion of the adenine compounds to GMP.[45] This gene is found primarily in *Lactobacillus*, *Bacillus* and *Clostridia;* however the riboswitch is found only in few organisms. *L.plantarum, D.reducens* and *D.hafniense* possess a riboswitch upstream to the *ADE* gene.

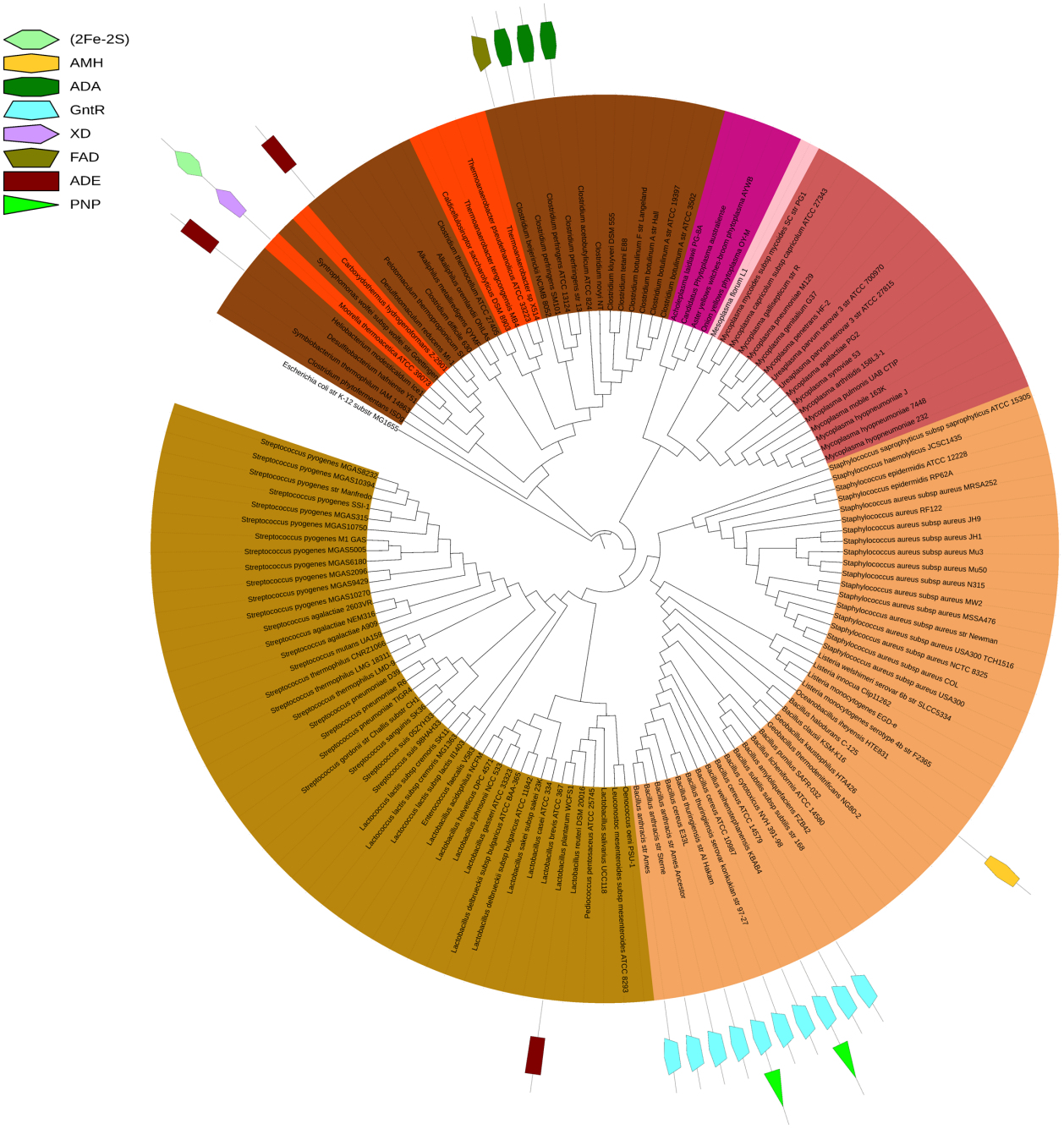

Figure 6. Phylogenetic distribution of rare riboswitches in Firmicutes.
Abbreviations: 2Fe-2S, 2Fe-2S binding protein; AMH, Amidohydrolase family protein; ADA, Adenosine deaminase; GntR, GntR transcription regulator; XD, Xanthine dehydrogenase; FAD, Flavin Adenine Dinucleotide dependent protein; ADE, Adenine deaminase; PNP, purine Nucleoside Phosphorylase.

Purine *Nucleoside Phosphorylase (PNP)* is a key enzyme in the purine salvage pathway which cleaves a nucleoside by phosphorylating the ribose to produce a nucleobase and ribose-1-phosphate. In *B.thuringiensis str Al Hakam* and *B.weihenstephanensis*, PNP was under the riboswitch mode of regulation.

Riboswitches were also found upstream to amidohydrolase (AMH) family protein in *B.halodurans*, a flavin adenine dinucleotide (FAD)-dependent oxidoreductase in *C.beijerinckii* and a xanthine dehydrogenase (XD) and 2Fe-2S binding protein in *Moorella.* The rarity of their occurrence in Firmicutes suggests that these riboswitches must have originated in specific lineages or clades.

The riboswitch upstream to the *adenosine deaminase (ADA)* gene (a purine salvage pathway enzyme) is found in a few organisms belonging to the *Clostridiales* order of Firmicutes. However, it also makes an appearance in some organisms belonging to the Shewanellaceae and Vibrionaceae family of Gammaproteobacteria. Figure 7 provides evidence of horizontal gene transfer of the *ADA* gene, since the phylogenetic tree constructed from the *ADA* gene shows that Shewanella and Vibrio clusters with organisms from

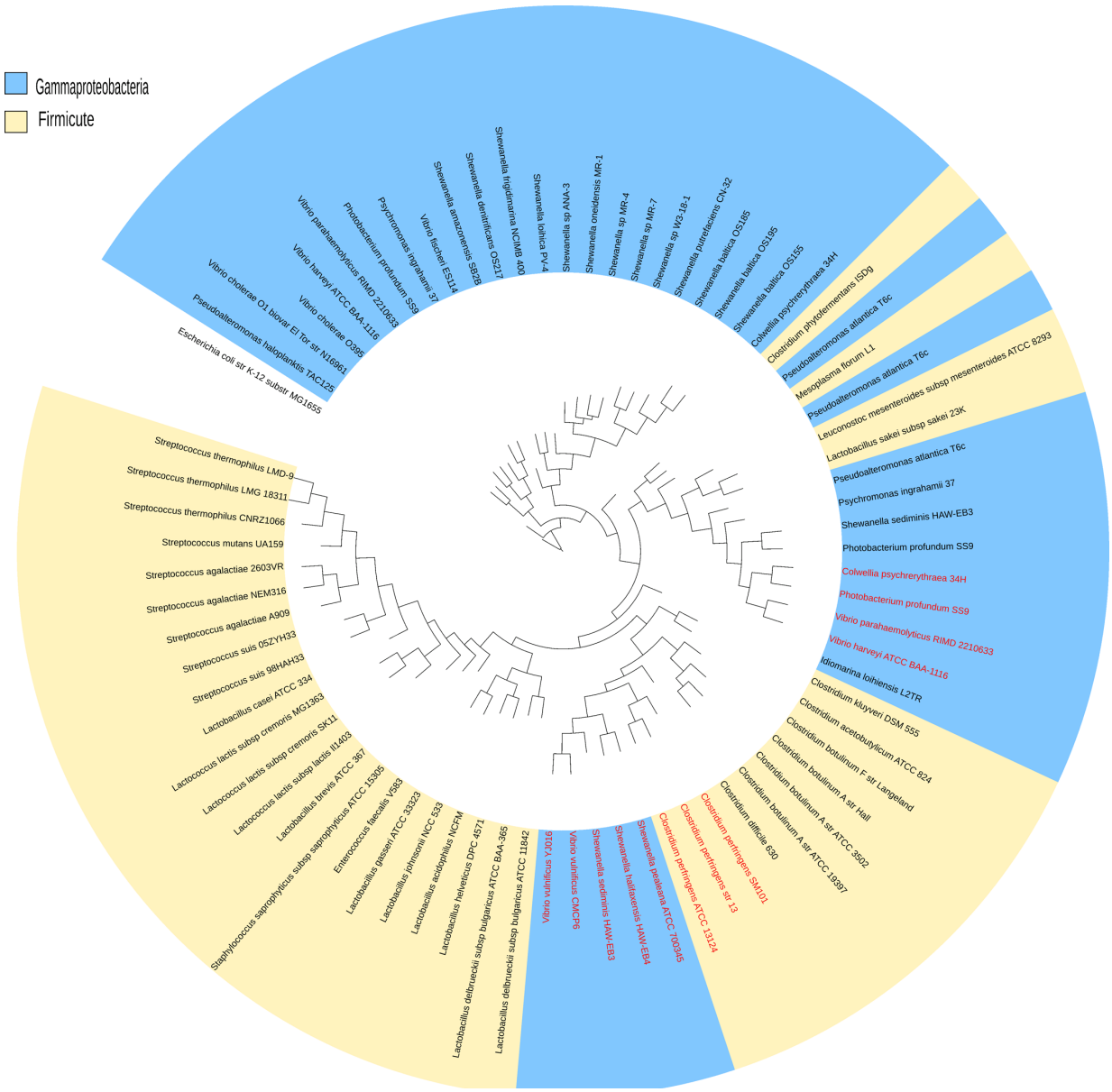

Figure 7. Phylogenetic tree of the adenosine deaminase gene found in Firmicutes and Gammaproteobacteria.
Note: The genomes labeled in red possess the riboswitch upstream of the *ADA* gene.

the Clostridiales order. Figure 7 shows two such clusters within the Clostridiales order. In one such cluster, all the purine-riboswitch-carrying members of the Shewanellaceae and Vibrionaceae families share the same clade with *C. perfringens*. This result strongly suggests that the ADA riboswitch found in these organisms belonging to the Shewanellaceae and Vibrionaceae family must have been horizontally transferred along with the gene it regulates. In the other cluster, only some of the members of the Vibrionaceae family carry a purine riboswitch upstream to the *ADA* gene. While it is quite plausible that the purine riboswitch may have been horizontally transferred from *C. perfringens* to these organisms, the evidence for HRT in this case is not conclusive, since this cluster is in a different clade. We therefore cannot rule out the possibility that they may have evolved independently in these organisms.

Riboswitch distribution outside Firmicutes

The purine riboswitches outside Firmicutes are scarce and restricted to a few species in the *Deltaproteobacteria*, *Gammaproteobacteria*, *Fusobacterium* and *Thermotogaceae* families. In *Deltaproteobacteria,* riboswitches are present only in *Bdellovibrio bacteriovorus.* Two instances of riboswitches are found in this genome, one upstream of the secreted nuclease and another upstream of a hypothetical protein. Conserved domain searches for these genes reveal that the secreted nuclease belongs to the endonuclease I superfamily, and a hypothetical gene shows similarity to *adenylsuccinate lyase*. The Endonuclease I superfamily is thought to normally generate double strand breaks in DNA, except in the presence of high salt concentrations, and RNA, when it generates single strand breaks in DNA. Its biological role is unknown. *Adenylsuccinate lyase* is a known purine de-novo synthesis gene. The hypothetical gene showing similarity to *adenylsuccinate lyase* is truncated and does not show similarity to *adenylsuccinate lyase* from other organisms. However another gene annotated as *adenylsuccinate lyase* in *Bdellovibrio bacteriovorus* is highly conserved. The genes carrying riboswitches in *Bdellovibrio bacteriovorus* are not conserved across different genomes.

The purine riboswitch is found upstream to the *pur* operon in *Fusobacterium nucleatum.* The first gene of the *pur* operon in *F.nucleatum* is the *purL* gene. In few *Clostridiales* species like *C.perfringens, C.beijerinckii* and *D.hafniense,* the first gene of the *pur* operon is also *purL*. Similarity searches indicated that the *purL* gene of *F.nucleatum* shares a strong similarity with the *purL* gene from *D.hafniense* (score = 1280 bits, e-value = 0 and coverage = 99%). The *Fusobacterium* branches out at the base of the lineage leading to Firmicutes and are known to have undergone massive gene transfer events to or from Firmicutes (particularly *Clostridiales* and *Streptococci*).[46] There is also evidence of a horizontal transfer of RFN-regulated genes from *D.hafniense* to *F.nucleatum.*[47] This leads us to speculate that the presence of the *pur* riboswitch in *F. nucleatum* may be another case of this horizontal gene transfer, where the *purL* gene along with the riboswitch upstream may have been transferred from *D.hafniense* to *F.nucleatum.* However this hypothesis needs to be further investigated once the number of fully sequenced genomes of this class increases.

Thermotoga lettingae and Petrotoga mobilis from Thermotogaceae family possess a riboswitch upstream of the *purE* gene, which is usually the first gene of the *pur* operon. In *Thermotoga lettingae,* the purine de-novo genes occur as a ten-gene operon, with the riboswitch upstream of the first gene in the operon, the *purE* gene. However, *Petrotoga mobilis* possesses purine de-novo genes as clusters. A six-gene cluster, the first gene of which is *pure,* is regulated by the riboswitch.

A few organisms belonging to the Shewanellaceae and Vibrionaceae family of Gammaproteobacteria also carry the purine riboswitch. *S.pealeana* and *S.halifaxensis* possess a riboswitch upstream to the permease gene. *S.pealeana, S.halifaxensis,* and *S.sediminis* from Shewanellaceae family, as well as *V.vulnificus*, *V.harveyi*, *V.parahaemolyticus* and *P.profundum* from the Vibrionaceae family also carry a riboswitch upstream of the *ADA* gene. Few of these instances are likely to be a result of horizontal riboswitch transfer along with the gene, as discussed earlier (Fig. 7).

Discussion

It is tempting to attempt to infer the evolutionary origin of the various purine riboswitches, given our knowledge of their detailed phylogenetic distribution. It has been argued[21,48] that riboswitches are remnants

of primordial regulatory machinery that may have been operational in an RNA world. Therefore, it seems interesting to ascertain whether the origin of the riboswitches can be traced back to the root of the tree of life.

In Firmicutes, the *XPT* gene occurs either as a single unit or in an operon with two or more genes. However, the absence of the corresponding purine riboswitch can always be correlated with either the absence of the *XPT* gene (or operon) or an UTR that is too short to accommodate a riboswitch. It then seems reasonable to conclude that the purine riboswitch was lost (sometimes along with the *XPT* gene) in these species. Moreover, the distribution (Fig. 2) also indicates that it is more parsimonious to infer a single origin for the purine riboswitch found upstream to the *XPT* gene. That point of origin of the riboswitch can be placed at the base of the Firmicutes phylum (Fig. 8) owing to its widespread presence (except in the parasitic groups). Since the Firmicute phylum is considered to be the earliest branching prokaryotic group,[25] it seems reasonable to conclude that the

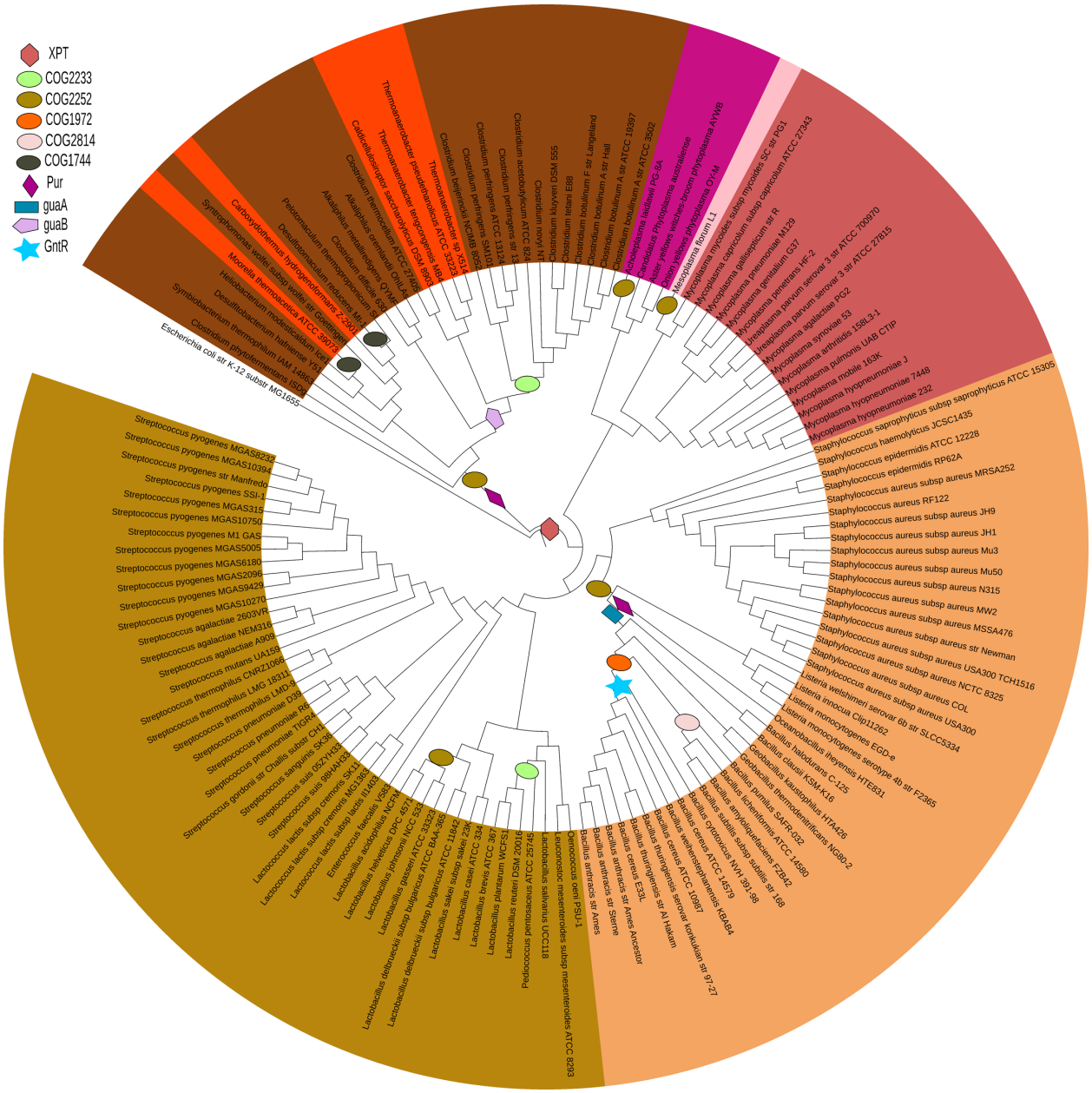

Figure 8. Possible point of origin of the various purine riboswitches.
Note: The possible points of origin for the different purine riboswitches are indicated by different shapes and colors.

purine riboswitch found upstream to the *XPT* gene has a primordial origin.

The distribution of riboswitches upstream to transporter genes belonging to COG1972, COG2814 and COG1744 indicates they are relatively rare and are likely to have originated in specific clades of the Bacillus genus or a few organisms belonging to Clostridiales and Thermoanaerobacteriales (see Fig. 8) quite late in the evolution of Firmicutes. Riboswitches upstream to the transporter genes belonging to COG2233 and COG2252 are more widespread. Even so, Figure 8 shows the multiple independent origins of the riboswitch upstream to some COG2252 genes, two of which can be placed at the root of the *Bacillaceae-Listeriacea* families and Clostridiales-Thermoanaerobacteriales orders.

The purine de-novo genes are found in almost all the organisms belonging to the phylum Firmicutes (except Mesoplasma, Phytoplasma and Mycoplasma). However, only a small fraction of them carry a purine riboswitch upstream to them. It is difficult to infer the origin of the *pur* riboswitch unambiguously. The fragmented nature of the distribution of the *pur* riboswitch, as evident from Figure 4, might suggest that they may have at least two independent origins associated with the appearance of different orders of Firmicutes during the course of evolution of Firmicutes. They may have appeared once prior to the divergence of Clostidiales from Thermoanaerobacteriales. Another independent point of origin of the riboswitch can be placed at the root of the *Bacillaceae* family (Fig. 8).

For the purpose of determining the point of origin of the *guaA* riboswitch, there is a need to distinguish between the purine riboswitch that is found upstream of a standalone *guaA* gene and that which is upstream to an operon containing the *guaA* gene. The distribution of the purine riboswitch upstream to the standalone GMP Synthase (*guaA*) gene (see Fig. 5) leads us to place the likely origin of that riboswitch at the root of the *Bacillaceae* family. The presence of a purine riboswitch upstream to a standalone *guaA* gene (that is not part of a *guaB-guaA* operon) in *C.difficile* is difficult to explain. The fact that the standalone *guaB* gene does not possess a riboswitch suggests the possibility that a reversal of the gene order followed by splitting of the *guaB-guaA* operon may have been responsible.

The distribution of the riboswitch upstream to the *guaB-guaA* operon in Clostridiales and some *Thermoanaerobacter* species (see Fig. 5) suggests that the riboswitch could either have originated prior to the divergence of the Clostridiales-Thermoanaerobacteriales order and been subsequently lost in some organisms belonging to this order or it may have originated much later in a specific clade of Clostridiales and Thermoanaerobacteriales as indicated in Figure 8.

The presence of the riboswitch upstream to the *GntR* transcription factors is an interesting case. Usually riboswitches regulate the expression of the genes/operons involved in the metabolism and uptake of a particular metabolite. However, the existence of purine riboswitches regulating the *GntR* transcription regulators expands our knowledge on the types of genes that can be regulated by riboswitches. Thus, riboswitches can also regulate the expression of the metabolic genes indirectly by controlling the expression of the transcription factors, as in the case of *GntR*. The riboswitch upstream to the transcription factor gene *GntR* is restricted to a set of organisms belonging to the Bacillus genus, even though the gene is more widely distributed across organisms belonging to the order Bacillales. It therefore seems reasonable to conclude that this riboswitch originated after the divergence of the Bacillales order from other orders of Firmicutes. The likely point of origin of the purine riboswitch upstream to the *GntR* gene is indicated in Figure 8.

Horizontal transfer of riboswitches along with the regulated genes has been well documented.[47–50] In our analysis, we found examples of horizontal riboswitch transfer (HRT) from Firmicutes to other phyla. It appears that the *pur* riboswitch in *Fusobacterium nucleatum* is one such example of HRT from *D.hafniense*. Another case is that of the *ADA* riboswitch in a few organisms belonging to the *Shewanellaceae* and *Vibrionaceae* families, which appears to have been transferred from *C.perfringens*. A number of examples of horizontal riboswitch transfer from Bacillus/Clostridia to other groups for riboswitch classes like RFN and TPP have been previously reported.[47,50] Thus, it seems that the horizontal transfer of the riboswitches between different organisms is quite prevalent and can be helpful in the dispersion of riboswitches across diverse prokaryotic phyla.

Some of the purine salvage pathway genes, de-novo pathway genes and transporter genes belonging to COG2233 and COG2252 are ubiquitous in all orders of Firmicutes (with the exception of parasitic orders like Mycoplasmatales). Hence, while discussing the evolutionary origin of the corresponding riboswitches, it is difficult to rule out the possibility that these riboswitches originated (along with the genes or operons they regulate) at the root of the Firmicute phylum (or perhaps even earlier) but were eventually lost, sometimes along with the gene (or operon), in some groups of organisms during the subsequent evolution of Firmicutes.

Another question that is raised by our analysis deals with the extent to which riboswitches are essential for the functioning of the organism. If a purine riboswitch was lost in several lineages or groups without having an adverse effect on the viability of the organisms, then it is likely that those organisms possessed alternative means of regulating purine metabolism, and did not need to rely exclusively on riboswitches.

Conclusions

Our work on the detailed distribution of purine riboswitches reveals that it regulates a wide variety of genes, ranging from purine biosynthesis genes to transporters and transcription factors. Hence, the evolutionary origin of the purine riboswitches has to be considered in the context of the many different types of genes that are regulated by these riboswitches. Our analysis suggests that the origin of the purine riboswitch upstream to the *XPT* gene can be traced back to the root of the Firmicute phylum. However the distribution of purine riboswitches upstream to transcription factor genes and some transporter genes suggests a more recent origin for those riboswitches. We also discovered cases where the *pur* operon (as well as the *xpt-pbuX* operon in some species of Bacillus) is subject to regulation both by a riboswitch as well as by the *purR* protein repressor. It would be interesting to explore the relative roles of these two distinct regulatory mechanisms. Our analysis also reveals how horizontal riboswitch transfer, along with the gene regulated by it, may have played a role in the presence of the purine riboswitch upstream to the *ADA* gene in some organisms belonging to the Shewanellaceae and Vibrionaceae families of Gammaproteobacteria. Our work provides the first comprehensive study of purine riboswitch distribution in prokaryotes.

Acknowledgements

We thank Sudha Bhattacharya and L. Aravind for valuable discussions.

Funding

The work was funded in part by a grant given to SS by the Department of Biotechnology (DBT), Government of India.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Author Contributions

PS contributed to the design of the study, wrote the programs, analyzed the data and wrote the manuscript. SS contributed to the design of the study, analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Sudarsan N, Barrick JE, Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*. 2009;9:644–7.
2. Cheah MT, Wachter A, Sudarsan N, Breaker RR. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*. 2007;447:497–500.
3. Miranda-Rios J, Navarro M, Soberon M. A conserved RNA structure (thi box) is involved in the regulation of thiamin biosynthetic gene expression in bacteria. *Proc Natl Acad Sci*. 2001;98:9736–41.

4. Mirinov AS, Gusarov I, Rafikov R, et al. Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. *Cell*. 2002;111: 747–56.
5. Fuchs RT, Grundy FJ, Henkin TM. S-adenosylmethionine directly inhibits binding of 30S ribosomal subunits to the SMK box translational riboswitch RNA. *Proc Natl Acad Sci U S A*. 2007;104:4876–80.
6. Wachter A. Riboswitch-mediated control of gene expression in eukaryotes. *RNA Biol*. 2010;7:67–76.
7. Singh P, Bandyopadhyay P, Bhattacharya S, Krishnamachari A, Sengupta S. Riboswitch detection using profile hidden Markov models. *BMC Bioinformatics*. 2009;10:325–38.
8. Deigan KE, FerrÉ-D'Amare AR. Riboswitches: discovery of drugs that target bacterial gene-regulatory RNAs. *Acc Chem Res*. 2011;12:1329–38.
9. Blount KF, Breaker RR. Riboswitches as antibacterial drug targets. *Nat Biotechnol*. 2006;24:1558–64.
10. Sudarsan N, Cohen-Chalamish S, Nakamura S, Emilsson GM, Breaker RR. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem Biol*. 2005;12:1325–35.
11. Blount KF, Wang JX, Lim J, Sudarsan N, Breaker RR. Antibacterial lysine analogs that target lysine riboswitches. *Nat Chem Biol*. 2007;3:44–9.
12. Kim JN, Blount KF, Puskarz I, Lim J, Link KH, Breaker RR. Design and antimicrobial action of purine analogues that bind Guanine riboswitches. *ACS Chem Bio*. 2009;4:915–27.
13. Mulhbacher J, Brouillette E, Allard M, Fortier LC, Malouin F, Lafontaine DA. Novel riboswitch ligand analogs as selective inhibitors of gunanine-related metabolic pathways. *Plos Pathog*. 2010;6:e1000865.
14. Mansjö M, Johansson J. The riboflavin analog roseoflavin targets an FMN-riboswitch and blocks Listeria monocytogenes growth, but also stimulates virulence gene-expression and infection. *RNA Biol*. 2011;4:674–80.
15. Lee ER, Blount KF, Breaker RR. Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulate gene expression. *RNA Biol*. 2009;6:187–94.
16. Wieland M, Benz A, Klauser B, Hartig JS. Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angew Chem Int Ed Engl*. 2009;48:2715–8.
17. Dixon N, Duncan JN, Geerlings T, et al. Reengineering orthogonally selective riboswitches. *PNAS*. 2010;107:2830–5.
18. Sharma V, Nomura Y, Yokobayashi Y. Engineering complex riboswitch regulation by dual genetic selection. *J Am Chem Soc*. 2008;130:16310–5.
19. Jin Y, Huang JD. Engineering a portable riboswitch-LacP hybrid device for two-way gene regulation. *Nucleic Acids Res*. 2011;39:e131.
20. Gardner PP, Daub J, Tate J, et al. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*. 2011;39:D141–5.
21. Breaker RR. Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol*. 2012;4:a003566.
22. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of thiamin biosynthesis in prokaryotes. New genes and regulatory mechanisms. *J Biol Chem*. 2002;277:48949–59.
23. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem*. 2003;278:41148–59.
24. Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol*. 2007;8:R239.
25. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311: 1283–7.
26. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
27. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analysis. *Bioinformatics*. 2009;25:1972–3.
28. Felsenstein J. PHYLIP: Phylogeny Inference Package. *Cladistics*. 1989;5: 164–6.
29. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*. 2011;27:1164–5.
30. Guindon S, Gascuel O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003;52: 696–704.
31. Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol*. 2008;25:1307–20.
32. Garrity GM, Bell JA, Lilburn TG. *Taxonomic Outline of the Procaryotes*. *Bergey's Manual of Systematic Bacteriology,* 2*nd Edition, Release* 5.0. 2004; New York: Springer-Verlag.
33. Gupta RS, Gao B. Phylogenomic analysis of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I). *Int J Syst Evol Microbiol*. 2009;59:285–94.
34. Letunic I, Bork P. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 2007;23: 127–8.
35. Saxild HH, Brunstedt K, Nielsen KI, Jarmer H, Nygaard P. Definition of the Bacillus subtilis PurR operator using genetic and bioinformatic tools and expansion of the PurR regulon with glyA, guaC, pbuG, xpt-pbuX, yqhZ-folD, and pbuO. *J Bacteriol*. 2001;183:6175–83.
36. Johansen LE, Nygaard P, Lassen C, Agerso Y, Saxild HH. Definition of a second Bacillus subtilis pur regulon comprising the pur and xpt-pbuX operons plus pbuG, nupG (yxjA), and pbuE (ydhL). *J Bacteriol*. 2003;185: 5200–9.
37. Becerra A, Lazcano A. The role of gene duplication in the evolution of purine nucleotide salvage pathways. *Orig Life Evol Biosph*. 1998;28:539–53.
38. He B, Shiau A, Choi KY, Zalkin H, Smith JM. Genes of the Escherichia coli pur regulon are negatively controlled by a repressor-operator interaction. *J Bacteriol*. 1990;172:4555–62.
39. Zalkin H, Ebbole DJ. Organization and regulation of genes encoding biosynthetic enzymes in Bacillus subtilis. *Journal Biol Chem*. 1998;263: 1595–8.
40. Demain AL, Shigeura HT. Dependence of diaminopurine utilization on the mutational site of purine auxotrophy in Bacillus subtilis. *J Bacteriol*. 1968;95:565–71.
41. Ng WL, Kazmierczak KM, Robertson GT, Gilmour R, Winkler ME. Transcriptional regulation and signature patterns revealed by microarray analyses of Streptococcus pneumoniae R6 challenged with sublethal concentrations of translation inhibitors. *J Bacteriol*. 2003;185:359–70.
42. Kilstrup M, Martinussen J. A transcriptional activator, homologous to the Bacillus subtilis PurR repressor, is required for expression of purine biosynthetic genes in Lactococcus lactis. *J Bacteriol*. 1998;180: 3907–16.
43. Jewett MW, Lawrence KA, Bestor A, Byram R, Gherardini F, Rosa PA. guaA and GuaB are essential for Borrelia burgdorferi survival in the tick-mouse infection cycle. *J Bacteriol*. 2009;191:6231–41.
44. Rigali S, Derouaux A, Giannotta F, Dusart J. Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR and YtrA subfamilies. *J Biol Chem*. 2002;277:12507–15.
45. Nygaard P, Duckert P, Saxild HH. Role of adenine deaminase in purine salvage and nitrogen metabolism and characterization of the ade gene in Bacillus subtilis. *J of Bacteriol*. 1996;178:846–53.
46. Mira A, Pushker R, Legault BA, Moreira D, Rodriguez-Valera F. Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol Biol*. 2004;4:50–67.
47. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res*. 2002;30:3141–51.
48. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*. 2004;20:44–50.
49. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Regulation of vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA*. 2003;9:1084–97.
50. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of thiamin biosynthesis in prokaryotes. *J Biol Chem*. 2002;227: 48949–59.

Supplementary Data

Supplementary file 1. Genes used for the construction of the phylogenetic tree of Firmicutes.

Supplementary file 2. Phylogenetic tree drawn using the NJ method with 1000 bootstrap replicates showing the points of origin of the various purine riboswitches in different shapes and colors.

Supplementary file 3. Phylogenetic tree drawn using the ML method with 1000 bootstrap replicates showing the point of origin of the various purine riboswitches in different shapes and colors.

Supplementary file 4. Sequence alignment of *pbuX* genes in *Pediococcus pentosaceus.*

gi|116491818:1818151-1819476 is the identifier indicating the gene coordinates of the pbuX gene that possesses the riboswitch. gi|116491818:c1427741-1426431 is the identifier indicating the gene coordinates of the *pbuX* gene without riboswitch.

Supplementary file 5. Sequence alignment of COG2252 genes.

(a) Sequence alignment of COG2252 genes in *L.acidophilus*. gi|159162017:1972139-1972925 is the identifier indicating the gene coordinates of the COG2252 gene that possesses the riboswitch. gi|159162017:1963679-1964999 and gi|159162017:1967562-1968829 are the identifiers indicating the gene coordinates of the COG2252 genes that are without riboswitches. (b) Sequence alignment of COG2252 genes in *L.gasseri*. gi|116628683:1860545-1861861 is the identifier indicating the gene coordinates of the COG2252 gene that possesses the riboswitch. gi|116628683:1858608-1859975 and gi|116628683:1855564-1856874 are the identifiers indicating the gene coordinates of the COG2252 genes that do not possess riboswitches. (c) Sequence alignment of COG2252 genes in *L.helveticus*. gi|161506634:2064345-2065649 is the identifier indicating the gene coordinates of the COG2252 gene that possesses the riboswitch. gi|161506634:2057069-2058379 and gi|161506634:2052112-2053419 are the identifiers indicating the gene coordinates of the COG2252 genes that do not possess riboswitches. (d) Sequence alignment of COG2252 genes in *L.johnsonii.* gi|42518084:1949566-1950876 is the identifier indicating the gene coordinates of the COG2252 gene that possesses the riboswitch. gi| 42518084:1947903-1949213 and gi|42518084:1945469-1946779 are the identifiers indicating the gene coordinates of the COG2252 genes that do not possess riboswitches.