2007

Origin of the Bacterial *SET* Domain Genes: Vertical or Horizontal?

Raul Alvarez-Venegas
Centro de Investigación y de Estudios Avanzados—Unidad Irapuato

Monther Sadder
University of Jordan

Alexander Tikhonov
Invitrogen Corporation

Zoya Avramova
University of Nebraska-Lincoln, zavramova2@unl.edu

Origin of the Bacterial *SET* Domain Genes: Vertical or Horizontal?

Raul Alvarez-Venegas,† Monther Sadder,*‡ Alexander Tikhonov,*§ and Zoya Avramova**

*School of Biological Sciences, University of Nebraska–Lincoln; †Department of Genetic Engineering, Centro de Investigación y de Estudios Avanzados—Unidad Irapuato, Guanajuato, CP 36821, México; ‡Faculty of Agriculture, University of Jordan, Amman, Jordan; and §Protein Array Center, Invitrogen Corporation, Branford, Connecticut

The presence of Supressor of variegation-Enhanser of zeste-Trithorax (SET) domain genes in bacteria is a current paradigm for lateral genetic exchange between eukaryotes and prokaryotes. Because a major function of SET domain proteins is the chemical modification of chromatin and bacteria do not have chromatin, there is no apparent functional requirement for the existence of bacterial *SET* domain genes. Consequently, their finding in only a small fraction of pathogenic and symbiotic bacteria was taken as evidence that bacteria have obtained the *SET* domain genes from their hosts. Furthermore, it was proposed that the products of the genes would, most likely, be involved in bacteria–host interactions. The broadened scope of sequenced bacterial genomes to include also free-living and environmental species provided a larger sample to analyze the bacterial *SET* domain genes. By phylogenetic analysis, examination of individual chromosomal regions for signs of insertion, and evaluating the chromosomal versus *SET* domain genes' GC contents, we provide evidence that *SET* domain genes have existed in the bacterial domain of life independently of eukaryotes. The bacterial genes have undergone an evolution of their own unconnected to the evolution of the eukaryotic *SET* domain genes. Initial finding of *SET* domain genes in predominantly pathogenic and symbiotic bacteria resulted, most probably, from a biased sample. However, a lateral transfer of *SET* domain genes may have occurred between some bacteria and a family of Archaea. A model for the evolution and distribution of *SET* domain genes in bacteria is proposed.

Introduction

In eukaryotes, chromatin structure provides an additional level of gene regulation by modulating genes' accessibility to the transcriptional machinery. Factors that alter chromatin structure are defined as epigenetic. They provide "memory" of transcriptional states that are faithfully reproduced after each round of cell division and throughout the development of an organism. Two antagonistically acting groups of genes, the Polycomb repressors and the Trithorax activators, are responsible for maintaining the activity of homeotic genes in higher eukaryotes. Ever since the discovery of a highly conserved (~130 amino acids) peptide Supressor of variegation-Enhanser of zeste-Trithorax (SET) in proteins belonging to both the repressor Su(var)3-9, E(z) and the activator (Trithorax and Ash1) groups, it has been expected that the SET domain peptide plays some important role. However, this role has remained a mystery until the recent discovery that SET domain peptides can methylate lysines at specific locations on the histone tails (Rea et al. 2000). Modified amino acids on the histone tails provide tags that are "read" by other complexes creating or destroying affinities for chromatin regulators. The combinatorial nature of these modifications commands transitions between active and inactive states, extending the informational potential of the genetic code (Felsenfeld and Groudine 2003). These amino-terminal modifications constitute a "histone code" and a molecular basis of epigenetic regulation (Jenuwein and Allis 2001).

Because SET domain–containing proteins may modify chromatin structure, not surprisingly, *SET* domain–encoding genes have been found in all eukaryotic genomes sequenced so far, from the unicellular primitive eukaryotes to the multicellular animals and plants. Absence of *SET* domain genes from a large number of bacterial genomes has provided support for an assumption that they have appeared with the occurrence of the eukaryotes (Stephens et al. 1998; Alvarez-Venegas and Avramova 2002; Aravind and Iyer 2003). Analyzing the genome of the obligate intracellular pathogen *Chlamydia trachomatis*, Stephens et al. (1998) identified a *SET* domain gene and suggested that it has originated via horizontal gene transfer (HGT) from a eukaryotic host. As additional bacterial genomes were sequenced, *SET* domain genes were identified in more pathogenic and symbiotic bacterial species. A logical assumption was made that the presence of *SET* domain genes in bacterial species is a consequence of their contacts with eukaryotic cells and that bacteria have acquired the gene through HGT (Alvarez-Venegas and Avramova 2002; Aravind and Iyer 2003). Several facts supported this assumption: First, bacteria lack chromatin structure (no need for epigenetic regulation); second, among more than 390 sequenced bacterial genomes (at the time of this study), only 83 carry *SET* domain sequences, and the majority of these bacteria are pathogens or symbionts; third, closely related species differ in whether they carry a *SET* domain–encoding gene; fourth, with the exception of 3 *Methanosarcina* species, sequenced Archaebacteria lack *SET*-related genes.

Transfer of eukaryotic genes is considered common in parasitic and symbiotic bacteria. For instance, the intimate association between *Chlamidiaceae* and host cells might favor horizontal gene flow (Koonin et al. 2001). However, analysis of multiple *Chlamidiae* genomes has indicated little genomic exchange with other genera (Read et al. 2000; Brinkman et al. 2002). The idea of a widely spread HGT, especially between species from the different domains of life, has become a hotly debated issue. Opinions range from HGT being overwhelming and rampant, obscuring possible phylogenetic relationship between the species, to being overemphasized, limited, and insufficient to "unroot" the tree of life (reviewed in Glandsdorff 2000; Ochman et al. 2000; Woese 2000; Brown 2003).

Here, we revisited the paradigm for a gene transfer across the eukaryotic and bacterial domains of life and analyzed the phylogeny of the *SET* domain–encoding genes

Key words: horizontal gene transfer, SET domain genes, chromatin proteins in bacteria, bacterial domain of life, lateral gene transfer.

E-mail: zavramova2@unl.edu.

Mol. Biol. Evol. 24(2):482–497. 2007
doi:10.1093/molbev/msl184
Advance Access publication December 5, 2006

found in bacteria. The broadened scope of sequenced bacterial genomes, to include also free-living and environmental species, provided an impetus to reexamine the distribution of *SET* domain sequences in a larger sample of bacterial genomes. Our goals were to determine, first, whether the presence of a *SET* domain in closely related species would be connected to their lifestyles, free versus parasitic; second, whether the earlier conclusion that *SET* domain genes have been horizontally acquired by pathogenic and symbiotic bacteria were biased by the available sample of sequenced genomes selectively representing pathogenic and agronomicaly significant species; and third, whether phylogenetic relationships between the bacterial *SET* domain genes could suggest occurrence and evolution unrelated to the eukaryotic *SET* domain genes.

Our analyses indicate that *SET* domain genes have existed in the bacterial domain of life and that their initial finding in pathogens and symbionts resulted from a biased sample. Importantly, bacterial *SET* domain genes have undergone an evolution of their own, unconnected to the evolution of the eukaryotic *SET* domain genes. Absence of SET domain sequences in the majority of currently available bacterial genomes, apparently, reflects gene loss. Phylogenetic and chromosome analyses of the *SET* domain gene–containing regions of *Chlorobium*, *Bacillus*, and *Methanosarcinal* genomes, however, suggest a possible HGT between some bacteria and Archaea.

Materials and Methods

Phylogenetic Analyses of Bacterial SET Domain–Containing Proteins

From approximately 400 completely and partially sequenced bacterial genomes in the National Center for Biotechnology Information (NCBI) database, we have retrieved 83 bacterial species encoding putative SET domain proteins (table 1). Excluding identical sequences from very closely related genomes, we analyzed 45 SET domain proteins found in 39 bacterial species; duplicate genes representing paralogous functions within the same species' genomes are included. Eukaryotic entries were selected to represent a broad cross section of proteins found in unicellular, filamentous, and multicellular organisms. Members from different SET domain families, as recognized in plant and animal systems (Baumbusch et al. 2001; Alvarez-Venegas and Avramova 2002; Marmostein 2003), are included. SET domain peptides have a tripartite structure of conserved N- and C-boxes separated by an inserted middle module of variable length and composition (Aravind and Iyer 2003; Marmostein 2003; see also fig. 3). To achieve optimal alignment, eukaryotic proteins with insertions comparable to the lengths of insertions in the bacterial proteins (20–30 amino acids) were selected. A full list of aligned sequences is summarized in the supplementary figure SF1, Supplementary Material online. Database searches were performed with Blast and PSI-Blast programs on the NCBI nonredundant database. SET domain–containing proteins were collected by TBlastX and PSI-Blast searches (*E* value 0.001). Pairwise alignments were compiled using the ClustalW program (Chenna et al. 2003). Phylogenetic and bootstrap analyses using the Protpars method from the PHYLIP package and the Seqboot program (500 pseudoreplicates) were employed. Unrooted majority rule consensus trees were built with the CONSENSE and plotted using the TREEVIEW programs (Page 1996). The fitch function was used to make minimum evolution trees using Phylip software (Felsenstein 1989). Minimal evolution (ME) trees with 10 global rearrangements and 10 random jumbles are shown in supplementary figures SF2 and SF3, Supplementary Material online.

Phylogenetic analyses for supportive bacterial 23S ribosomal RNAs and bacterial 50S ribosomal protein L3 were performed as specified in the legends of the supplementary figures SF4 and SF5, Supplementary Material online.

Genome analyses and localization of bacterial *SET* domain genes were carried individually for each of the bacterial genomes carrying *SET* domain genes. To outline syntenic regions flanking the SET gene, bacterial genomes were compared by pairwise alignment. Comparisons were carried out at both DNA sequence levels and at amino acid level using published genome data.

Results and Discussion

Distribution of *SET* Domain–Encoding Genes among Bacteria

Analysis of overall distribution of *SET* domain–encoding genes among sequenced bacterial genomes (NCBI) revealed the following facts (summarized in table 1): First, the retrieved bacterial *SET* domain genes are present in most of the known bacterial domains (Cyanobacteria, photosynthetic green sulfur, Flexobacter–Bacteroides, Spirochaetaes, Chlamidiae, Planctomycetaes, and low G + C Gram-positive and α, β, γ, and δ Gram-negative bacteria). Clearly, *SET* domain genes have existed before the separation of these branches. Second, although most species represent obligatory pathogens and symbionts, SET domain genes are found also in environmental species (opportunistic pathogens or symbionts) and in free-living organisms for which no symbiotic relationships have been found. Examples of the first group are *Leptospira interrogans*, *Burkholderia fungorum*, and *Bradyrhisobium japonicum*. The second include *Ralstonia matallidurans*, *Rhodopseudomonas palustris*, *Chlorobium tepidum*, *Rubrivivax gelatinosus*, and *Verrucomicrobium spinosum*. *SET* domain genes exist in organisms living at arctic temperatures (*Polaromonas*) and in hot springs (*Chlorobium*). Thus, it sounds unlikely that bacteria living on its own and under extreme conditions have acquired the *SET* domain gene through a eukaryote. Third, among the SET domain–containing bacteria, we identified 6 species that contain more than 1 copy of a *SET* gene. This finding is particularly important because it illustrates duplication events and evolution of bacteria-specific SET domain paralogs. We note also the highly variable presence of *SET* domain genes in closely related species. Of the available sequenced genomes of the δ- and ε-subdivisions, only 1 representative of the δ-subdivision (*Myxoccocus xanthus*) carries a *SET* domain gene, whereas none of the ε-subdivision has been found yet. In contrast, all reported species from the γ-subdivision (the *Xylella* and *Xanthomonas* species) have *SET* domain genes. In the β-subdivision, all members of

Table 1
Distribution of *SET* Domain Genes in Bacterial and Archaebacterial Species

Species	Genomes Sequenced (390)	Containing SET Domain Genes (83)	
Archaea			
Crenarchaeota	**5**	**0**	
Euryarchaeota	**20**	**3**	
Methanosarcinales	**4**	**3**	
Methanococcoides burtonii DSM 6242	{1}	[0]	Env./free
Methanosarcina acetivorans C2A	{1}	[1]	Env./free
Methanosarcina barkeri strain fusaro	{1}	[1]	Env./free
Methanosarcina mazei Go1	{1}	[1]	Env./free
Nanoarchaeota	**1**	**0**	
Bacteria			
Actinobacteria	**27**	**0**	
Chlamydiae	**10**	**9**	
Chlamydia muridarum Nigg	{1}	[1]	Intracell. parasite
Chlamydia trachomatis D/UW-3/CX	{1}	[1]	Intracell. parasite
Chlamydophila abortus S26/3	{1}	[1]	Intracell. parasite
Chlamydophila caviae GPIC	{1}	[1]	Intracell. parasite
Chlamydophila pneumoniae AR39	{4}	[1]	Intracell. parasite
Parachlamydia sp. UWE25	{1}	[0]	Env./free
Verrucomicrobium spinosum DSM 4136	{1}	[1]	Eutrophic ponds/free
Cyanobacteria	**18**	**2**	
Nostoc punctiforme PCC 73102	{2}	[1]	Env./free
Firmicutes	**102**	**3**	
Bacillales	**38**	**3**	
Bacillus anthracis strain ''Ames Ancestor''	{10}	[1]	Opport. path.
Bacillus cereus ATCC 10987	{4}	[1]	Opport. path.
Bacillus thuringiensis serovar konkukian	{1}	[1]	Opport. path.
Other	**37**	**6**	
Chlorobium limicola DSM 245	{1}	[1]	Obligate photolithotrop
Chlorobium phaeobacteroides DSM 266	{1}	[1]	Phototrop. sulfur/env.
Chlorobium tepidum TLS	{1}	[1]	Green sulfur/free
Cytophaga hutchinsonii	{1}	[2]	Env./free
Rhodopirellula baltica SH 1	{1}	[1]	Free/opport. path
Proteobacteria			
Alpha subdivision	**21**	**7**	
Other	**4**	**7**	
Bradyrhizobium japonicum USDA 110	{1}	[2]	Soil/symbiont
Mesorhizobium loti MAFF303099	{2}	[2]	Soil/symbiont
Rhodopseudomonas palustris CGA009	{1}	[1]	Soil/free
Beta subdivision	**38**	**29**	
Bordetella	**3**	**3**	
Bordetella bronchiseptica RB50	{1}	[1]	Opport. path.
Bordetella parapertussis 12822	{1}	[1]	Opport. path.
Bordetella pertussis Tohama I	{1}	[1]	Opport. path.
Burkholderiaceae	**23**	**23**	
Burkholderia cenocepacia AU 1054	{3}	[1]	Opport. path.
Burkholderia fungorum LB400	{1}	[1]	Soil/opport. path.
Burkholderia mallei 10229	{7}	[1]	Opport. path.
Burkholderia pseudomallei S13	{7}	[1]	Opport. path.
Burkholderia sp. 383	{1}	[1]	Opport. path.
Burkholderia vietnamiensis G4	{1}	[1]	Opport. path.
Ralstonia eutropha JMP134	{1}	[1]	Env/phenoldegrading
Ralstonia matallidurans CH34	{1}	[1]	Env/free
Ralstonia solanacearum GMI1000	{1}	[1]	Soil/opport. path.
Neisseriaceae	**5**	**0**	Pathogens
Other	**7**	**3**	
Polaromonas sp. JS666	{1}	[1]	Env./low temp.
Rubrivivax gelatinosus PM1	{1}	[2]	Env./photosynthetic
Delta subdivision	**8**	**1**	
Myxococcus xanthus DK 1622	{1}	[1]	Env./free
Epsilon subdivision	**8**	**0**	
Gamma subdivision	**89**	**7**	
Xanthomonadaceae	**7**	**7**	
Xanthomonas axonopodis pv. citri strain 306	{1}	[1]	Pathogen
Xanthomonas campestris pv. campestris strain 8004	{2}	[1]	Pathogen
Xanthomonas oryzae pv. oryzae KACC10331	{1}	[1]	Pathogen
Xylella fastidiosa and *Dixon fastidiosa*	{3}	[1]	Pathogen
Spirochaetales	**6**	**6**	
Leptospira interrogans	{2}	[3]	Opport. path.

NOTE.—The numbers of sequenced genomes of very closely related species (i.e., serovars and strains) is shown in brackets {}; The numbers in [] brackets, show the number of *SET* domain genes/per genome found in individual species. Total numbers for the respective entire groups are shown in bold. Env., environmental; free, free living; intracell. parasite, intracellular parasite; opport. path., opportunistic pathogen; temp., temperature, phototrop, phototropic; photolithotrop, photolithotropic.

the Burkholderiacea and the Bordetella families have *SET* genes, whereas no member from the Neisseriaceae family carries any. *Burkholderia cepacia* and *Ru. gelatinosus* (from an undefined β-subdivision group) have 2 *SET* domain genes in each genome.

Among 29 genomes from the α-proteobacterial subdivision, available at the time of this study, only 3 species (*Br. japonicum*, *Rh. palustris*, and *Mesorhizobium loti*) carried *SET* domain genes. It is remarkable that both *Br. japonicum* and *Me. loti* have 2 *SET* domain genes each, whereas closely related species (symbiotic Rhizobiaceae and pathogenic Rickettsiales) may have none.

Outside the proteobacterial domain, the distribution of the *SET* domain–containing genomes is both broad and sporadic. Chlamydial species (with the exception of Parachlamydia, UWE25) carry *SET* domain genes, whereas none of the 23 Actinobacteria genomes has it. Species from Cyanobacteria (all reported *Nostoc* species), Planctomycetaes (*Pirellula*), photosynthetic high-temperature inhabiting (*Chlorobium*), Flexobacter–Bacteroides (*Cytophaga*), Verrucomicrobia (*Verrucomicrobium*), and Spirochaetales (*Leptospira*) have SET domain genes. Some have more than 1 copy, indicating paralogy. Interestingly, among more than 60 sequenced low G + C Gram-positive bacteria, only 3 Bacilli (*Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*) have *SET* domain genes. Lastly, only 3 mesophylic Archaea species carry *SET* domain genes, intriguingly similar to the SETs of the Bacilli.

Collectively, these observations argue that *SET* genes have existed in the bacterial domain of life before the occurrence of eukaryotes. Absence of *SET* domain genes from many bacterial genomes could reflect gene loss.

The Eukaryotic–Bacterial SET Domain Tree

Currently, the 2 major arguments supporting a lateral gene transfer between the eukaryotic and bacterial kingdoms are: 1) the presence of *SET* domain genes in the obligatory intracellular parasites from the orders Chlamydophila/Chlamydia and 2) the high levels of similarity of the SET domain proteins from the pathogenic γ-Proteobacteria (*Xanthomonas* and *Xylella*) and eukaryotes. The first argument is based on the assumption that intracellular *Chlamydiae*, having diverged from eubacteria some 2 billion years ago (Horn et al. 2004 and references therein), constitute an isolated niche sheltered from exchanges with other bacteria. However, living within eukaryotic hosts is a plausible condition enabling a horizontal acquisition of a *SET* domain gene (Stephens et al. 1998; Aravind and Iyer 2003). The second argument, based on the high similarity (2×10^{-20}, 38% identical) of the SET domain protein of the plant pathogen *Xylella fastidiosa* to a rice SET protein, is taken as evidence for a recent transition from a host genome to the genome of the invading bacteria (Aravind and Iyer 2003).

To examine the relationships between eukaryotic and bacterial *SET* domain genes, we reconstructed phylogenetic trees by several different approaches. Phylogenetic analysis is an objective approach for determining the occurrence and the directionality of HGT, despite some recognized limitations (Stanhope et al. 2001). Thereby, we carried out additional analyses to provide independent support employing different (genome-based) approaches.

The reconstructed maximum parsimony (MP) tree (fig. 1) shows all eukaryotic entries clustered on 2 related branches (69% bootstrap). The support is not very strong, but 2 observations relevant for our further discussion are that eukaryotic proteins do not intermix with proteins of bacterial origin and that a similar distribution pattern is consistently reproduced by trees built by different techniques (ME tree; supplementary fig. SF2, Supplementary Material online) and with different combinations of eukaryotic entries (data not shown). According to the criteria of Stanhope et al. (2001), these results did not support an HGT from eukaryotes to prokaryotes. Interestingly, however, they illustrated complex relationships among the bacterial proteins including a possible HGT among *Bacilli* and the archaeal species (see further below).

The Bacterial SET Domain Tree

To explore SET domain protein relationships among bacteria, we reconstructed trees using only the bacterial SET domain sequences (fig. 2; supplementary fig. SF3, Supplementary Material online). As controls, we built trees for bacterial genes less likely to be subjected to HGT. rRNA-based trees are considered "immune" to HGT and are largely supported by genome trees (van Berkum et al. 2003). Other genes, postulated to be less prone to horizontal shuffling, code for highly integrated elements tightly coupled with a functioning integral system, like individual ribosomal proteins (Woese 1998). MP trees for the 23S rRNA genes and for the conserved 50S ribosomal subunit protein, L3, of the bacterial species carrying *SET* domain genes (supplementary figs. SF4 and SF5, Supplementary Material online) are in general agreement with the SET domain protein tree, although some relationships are not well supported.

Closer examination of the bacterial SET domain tree revealed that the bacterial proteins could be separated into 2 distinct domains, arbitrarily called here Domain 1 and Domain 2 (fig. 2). The positioning of each protein within a Domain reflects a characteristic structural feature of its SET domain. Structural studies of eukaryotic SET domain proteins have led to the discovery of an unusual fold, the "knot" (Jacobs et al. 2002). Two conserved peptides, known as the N-terminal and the C-terminal boxes, flank a nonconserved insertion module of variable length and composition (fig. 3). The N- and the C-terminal boxes form the knot and carry the 2 most conserved amino acid motifs involved in the formation of the "loop" and the "thread." Overlapping with these motifs are the NHXC and the GEELXXXY consensus sequences involved in the cofactor-binding and the enzyme active sites, respectively (reviewed in Marmostein 2003; fig. 3). Relevant for our analysis is the Cys residue in the NHXC box. While Asn and his amino acids are conserved in all known SET domain proteins, the Cys is conserved only in a subset (Marmostein 2003). As a rule, proteins with a Cys in the box carry also a conserved motif (CXCXXXXC) downstream of the SET domain, known as the post-SET domain. Structural studies have shown that the C from the NHXC box and the CXCXXXXC motif may coordinate a zinc atom to form a Zn finger. It plays a role in the substrate

specificity of the SET domain methylases (Xiao et al. 2003). A Cys in the NHXC box predicts presence of the post-SET motif, whereas absent C is always accompanied by absent post-SET domain. No violation of this structural rule has been found so far and the bacterial SET domain peptides make no exception (fig. 3). From hereon, presence or absence of post-SET domain motifs in the bacterial proteins is denoted as (+)pSET and (−)pSET, respectively. All proteins in Domain 1 carry the (+)pSET version, whereas those on Domain 2 lack the post-SET motif. This structural feature reflects their distribution into distinct phylogenetic domains. We suggest that the absence of a post-SET domain represents a secondary event in the evolution of the bacterial SET domain function resulting from a single mutation of the C in the NHXC box. In the absence of evolutionary pressure to keep the Zn finger, a loss of the post-SET domain subsequent to this mutation is a likely outcome.

Domain 1: The (+)pSET Domain Bacterial Proteins

Grouped in Domain 1 are proteins of the (+)pSET type (fig. 3). The SET domain proteins of the β-, γ-, and δ-Proteobacteria display common origins, in agreement with the respective 23S RNA and S50-L3 protein trees (supplementary figs. SF4 and SF5, Supplementary Material online). Incongruent is the clustering of *V. spinosum* with Proteobacteria in all SET domain trees (figs. 1 and 2; supplementary figs. SF2 and SF3, Supplementary Material online). *Verrucomicrobium spinosum* is a free-living environmental species from the Chlamidiae/Verrucomicrobia group (Schlesner 2004) and the clustering of its SET domain protein with Proteobacteria, but not Chlamydia, might indicate HGT for the *V. spinosum* gene. However, it is possible also that the *SET* gene has an ancestral origin in *V. spinosum*, a possibility discussed later in more detail.

Representatives of the γ- and β-subdivisions segregate into the most populous and best-supported clades. Two species, *B. cepacia* and *Ru. gelatinosus*, have 2 *(+)pSET* domain copies in their genomes each (*ZP_00425417* and *YP_771959* and *ZP_00241588* and *ZP_00243950*, respectively). These genes form sister groups with proteins from other species and, thus, represent paralogs. The α-group is represented by 2 highly related species, *Br. japonicum* and *Rh. palustris*, with 80% identical SET domain proteins. We note that the *(+)pSET* version of these species is slightly diverged (fig. 3). Two free-living species, *Pirellula* sp. and *Nostoc punctiforme*, have related (+)pSET proteins. *Pirellula*, a marine bacterium from the order Planctomycetales, is considered to be of an independent monophyletic origin with no clear ancestral relationships to the other bacteria (Glockner et al. 2003). However, the SET domain protein of *Pirellula* sp. always clusters with the SET domain protein of the *Cyanobacterium* (*Nostoc*), pointing to a common, albeit distant, ancestry.

Domain 2: The (−)pSET Domain Bacterial Proteins

A major distinction of Domain 2 members is that all proteins are of the (−)pSET type, with the exception of one *L. interrogans* copy. Species of broadly diverse origins, including 3 members of the archaebacterial, Methanosarcinae family are grouped in Domain 2. The *SET* domain genes from 2 rhizobial species are the only representatives of Proteobacteria in this phylogenetic domain.

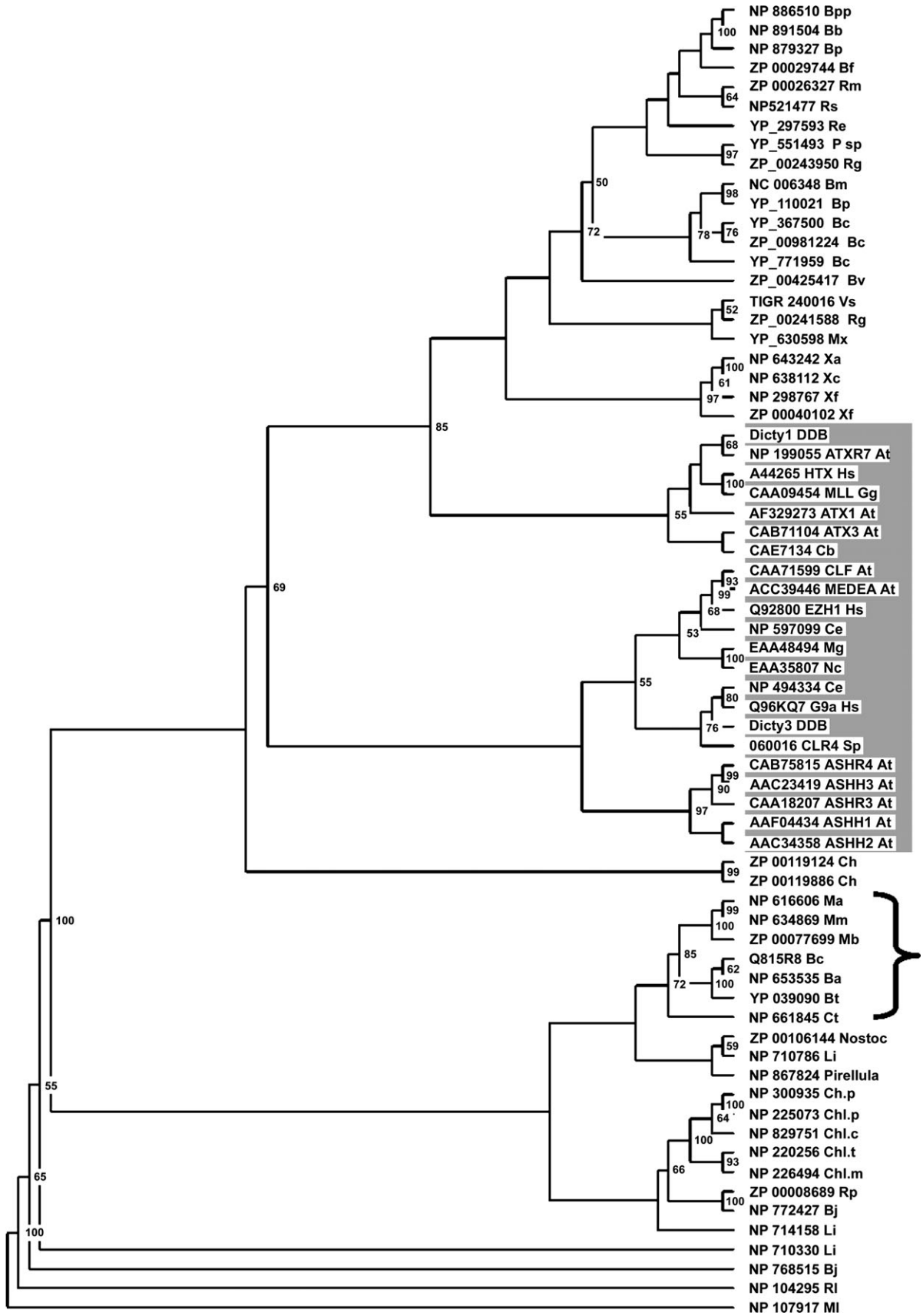
The Origin and the Evolution of the Bacterial SET Domain Genes May Be Traced in the Spirochaetae L. interrogans

Leptospira interrogans carries 3 *SET* domain copies: 1 *(+)pSET* and 2 *(−)pSET* types only weakly related to each other. The *(+)pSET* (NP_71078) is weakly (6×10^{-6}) similar to one of the *(−)pSET* (NP_710330), and there is no significant similarity to the other *(−)pSET* (NP_714158). The 2 (−)pSET copies are distantly related (8×10^{-3}), suggesting that each of the 3 *L. interrogans* genes represents an ancient paralog descending from a distinct ancestral gene line (fig. 5).

Both a facultative parasite and a saprophyte that can strive on its own, *L. interrogans* is related to the strictly parasitic spirochaetae, *Borrelia burgdorferei* and *Treponema pallidum*, but the latter do not have *SET* domain genes. Only 315 genes are shared between the 3 species and it is thought that species-specific genes provide *L. interrogans* with opportunities not required for the obligatory parasitic spirochaetae (Ren et al. 2003). Clearly, the 3 *SET* domain genes belong to the category distinguishing the free-living organism from the strictly parasitic relatives. Among eubacteria, spirochaete are evolutionarily primitive and their origins are not clear. On the 23S RNA tree, *L. interrogans* was remotely related to the Proteobacteria and *Chlamydiae* species, whereas its 50S L3 protein related it to the Bacteroides–Flexobacter group (*Cytophaga hutchinsonii*) and to the green-sulfur cyanobacterium, *Ch. tepidum* (supplementary figs. SF4 and SF5, Supplementary Material online). Through its 3 *SET* domain genes, *L. interrogans* related phylogenetically to all these bacterial groups and, thus, may be defined as a bearer of gene copies descending from ancient paralogs (see also fig. 5).

→

FIG. 1.—MP tree of bacterial and selected eukaryotic representative SET domain proteins. Figures indicate bootstrap values (100 = 100%, 500 replicates). Support higher than 50% is shown. The 3 *Bacilli*, the 3 *Methanosarcinae* species, as well as *Chlorobium*, segregate in a well-supported clade consistent with HGT (bracketed clade). The distribution of the eukaryotic proteins is in the shaded area. The following abbreviations were used: Bpp, *Bordetella parapertusis*; Bb, *Bordetella bronchiseptica*; Bp, *Bordetella pertusis*; Bf, *Burkholderia fungorum*; Bm, *Burkholderia malei*; Bpm, *Burkholderia paramalei*; Bc, *Burkholderia cepacia*; Burk, *Burkholderia* environmental sample; Rm, *Ralstonia matallidurans*; Rs, *Ralstonia solanacearum*; R. eu, *Ralstonia euthropa*; P sp, *Polaromonas* sp. *JS666*; Rg, *Rubrivivax gelatinosus*; Vs, *Verrucomicrobium spinosum*; Mx, *Myxococcus xanthus*; Xa, *Xanthomonas axonopodis*; Xc, *Xanthomonas campestris*; Xf, *Xylella fastidiosa*; DDB, *Dictyostelium discoideum*; At, *Arabidopsis thaliana*; Hs, *Homo sapiens*; Gg, *Gallus gallus*; Cb, *Caenorhabditis briggsae*; Ce, *Caenorhabditis elegans*; Mg, *Magnaporthe griseae*; Nc, *Neurospora crassa*; Sp, *Schizosaccharomyses pombe*; Ch, *Cytophaga hutchinsonii*; Li, *Leptospira interrogans*; Ma, *Methanosarcina acetivorans*; Mm, *Methanosarcina mazei*; Mb, *Methanosarcina barkerii*; Bc, *Bacillus cereus*; Ba, *Bacillus anthracis*; Bt, *Bacillus thuringiensis*; Ct, *Chlorobium tepidum*; Chl. p. *Chlamydophila pneumoniae*; Chl. c, *Chlamydophila caviae*; Chl. t, *Chlamydia trachomatis*; Chl. m, *Chlamydia muridarum*; Rp, *Rhodopseudomonas palustris*; Bj, *Bradyrhisobium japonicum*; Rl, *Rhizobium loti*; Ml, *Mesorhizobium loti*.


NP 886510 Bpp
NP 891504 Bb
NP 879327 Bp
ZP 00029744 Bf
ZP 00026327 Rm
NP521477 Rs
YP_297593 Re
YP_551493 P sp
ZP_00243950 Rg
NC 006348 Bm
YP_110021 Bp
YP_367500 Bc
ZP_00981224 Bc
YP_771959 Bc
ZP_00425417 Bv
TIGR 240016 Vs
ZP_00241588 Rg
YP_630598 Mx
NP 643242 Xa
NP 638112 Xc
NP 298767 Xf
ZP 00040102 Xf
Dicty1 DDB
NP 199055 ATXR7 At
A44265 HTX Hs
CAA09454 MLL Gg
AF329273 ATX1 At
CAB71104 ATX3 At
CAE7134 Cb
CAA71599 CLF At
ACC39446 MEDEA At
Q92800 EZH1 Hs
NP 597099 Ce
EAA48494 Mg
EAA35807 Nc
NP 494334 Ce
Q96KQ7 G9a Hs
Dicty3 DDB
060016 CLR4 Sp
CAB75815 ASHR4 At
AAC23419 ASHH3 At
CAA18207 ASHR3 At
AAF04434 ASHH1 At
AAC34358 ASHH2 At
ZP 00119124 Ch
ZP 00119886 Ch
NP 616606 Ma
NP 634869 Mm
ZP 00077699 Mb
Q815R8 Bc
NP 653535 Ba
YP 039090 Bt
NP 661845 Ct
ZP 00106144 Nostoc
NP 710786 Li
NP 867824 Pirellula
NP 300935 Ch.p
NP 225073 Chl.p
NP 829751 Chl.c
NP 220256 Chl.t
NP 226494 Chl.m
ZP 00008689 Rp
NP 772427 Bj
NP 714158 Li
NP 710330 Li
NP 768515 Bj
NP 104295 Rl
NP 107917 Ml

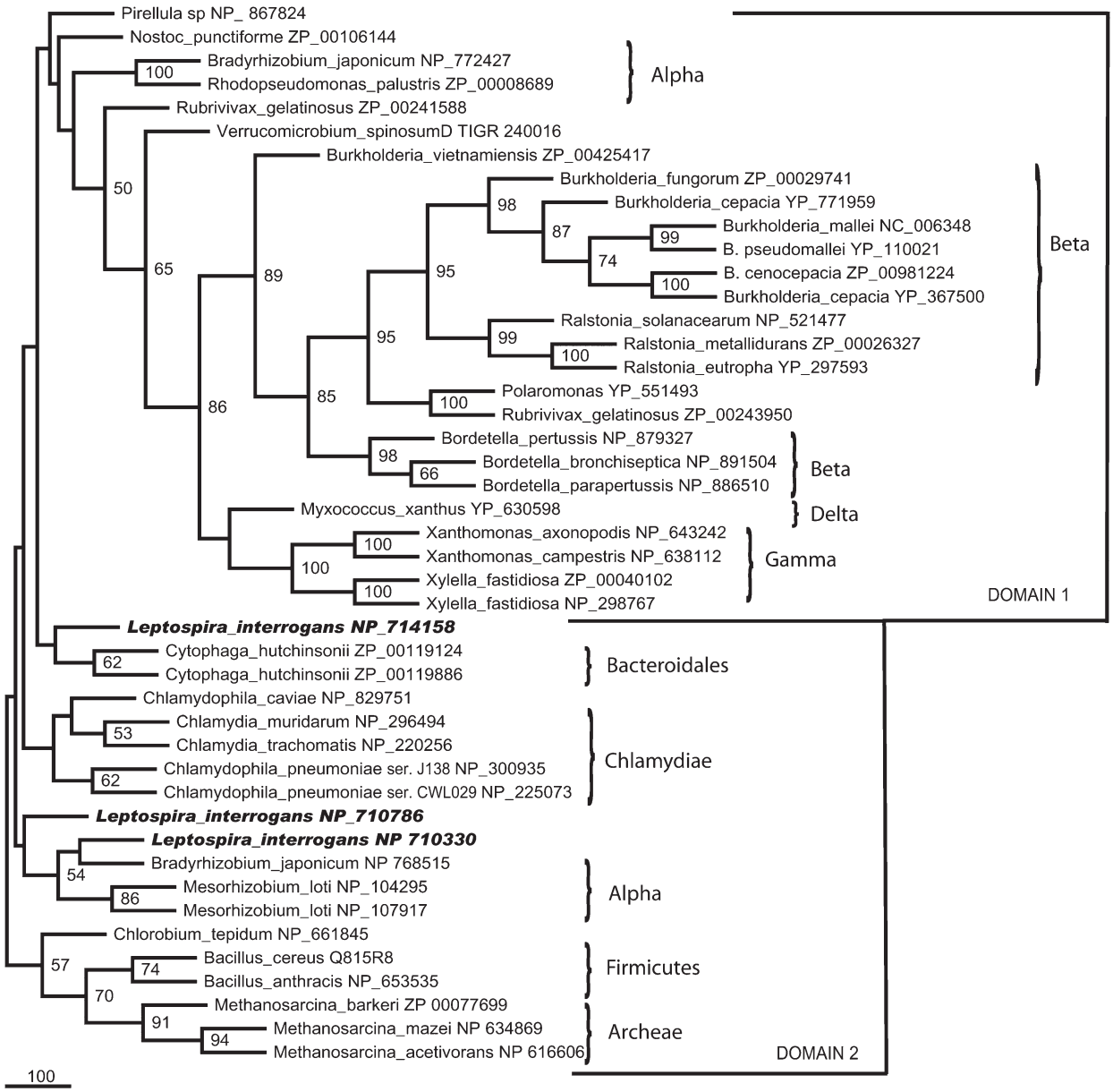


FIG. 2.—MP tree for bacterial SET domain proteins only. See figure 1 legend for descriptions and abbreviations. Domain 1 and Domain 2 cluster (+)pSET and (−)pSET-type proteins, respectively. SET domain proteins from the α-, β-, γ-, and δ-subdivisions cluster in Domain 1, congruent with the monophyletic origin of Proteobacteria.

The (+)pSET *L. interrogans* protein, NP_710786, lacks a significant portion of the N-box sequences that accounts for its positioning in Domain 2, despite the presence of a post-SET region (fig. 3). The other 2 *L. interrogans* SET domain copies represent 2 different versions of the (−)pSET. Each copy is related to SET domain proteins present in unrelated bacterial lineages: one (NP_710330) is most similar to the SET proteins found in 2 α-proteobacterial rhizobial species, *Me. loti* and *Br. japonicum*; the other (−)pSET protein (NP_714158) is related to the *Chlamydiacea* proteins, to a member of the large Cytophaga–Flavobacterium–Bacteroides subphylum, *Cy. hutchinsonii*, to *Chlorobium*, and to 3 proteins from *Bacillales*. Thereby, the 3 genes of *L. interrogans* represent bacterial *SET* domain gene paralogs. The relationships between the SET domain copies of *L. interrogans* and the other bacteria may provide important clues on the relationship and the evolution of these genes in bacteria and are analyzed in more detail.

The L. interrogans (−)pSET NP_710330 Protein and the Rhizobial Symbionts

The spirochaetal gene (*NP_710330*) is much closer to the 2 genes found in *Me. loti* (7×10^{-17} and 2×10^{-12}) and to 1 of the 2 genes found in *Br. japonicum* (7×10^{-15}) than to any of the copies in its own genome.

The 2 rhizobial symbionts *Me. loti* and *Br. japonicum* are closely related and share common ancestry (Kaneko et al. 2000, 2002). Their (−)pSET domain proteins are related between themselves (5×10^{-18}) and with the *L. interrogans* (NP_710330, 7×10^{-17} and 7×10^{-15}, re-

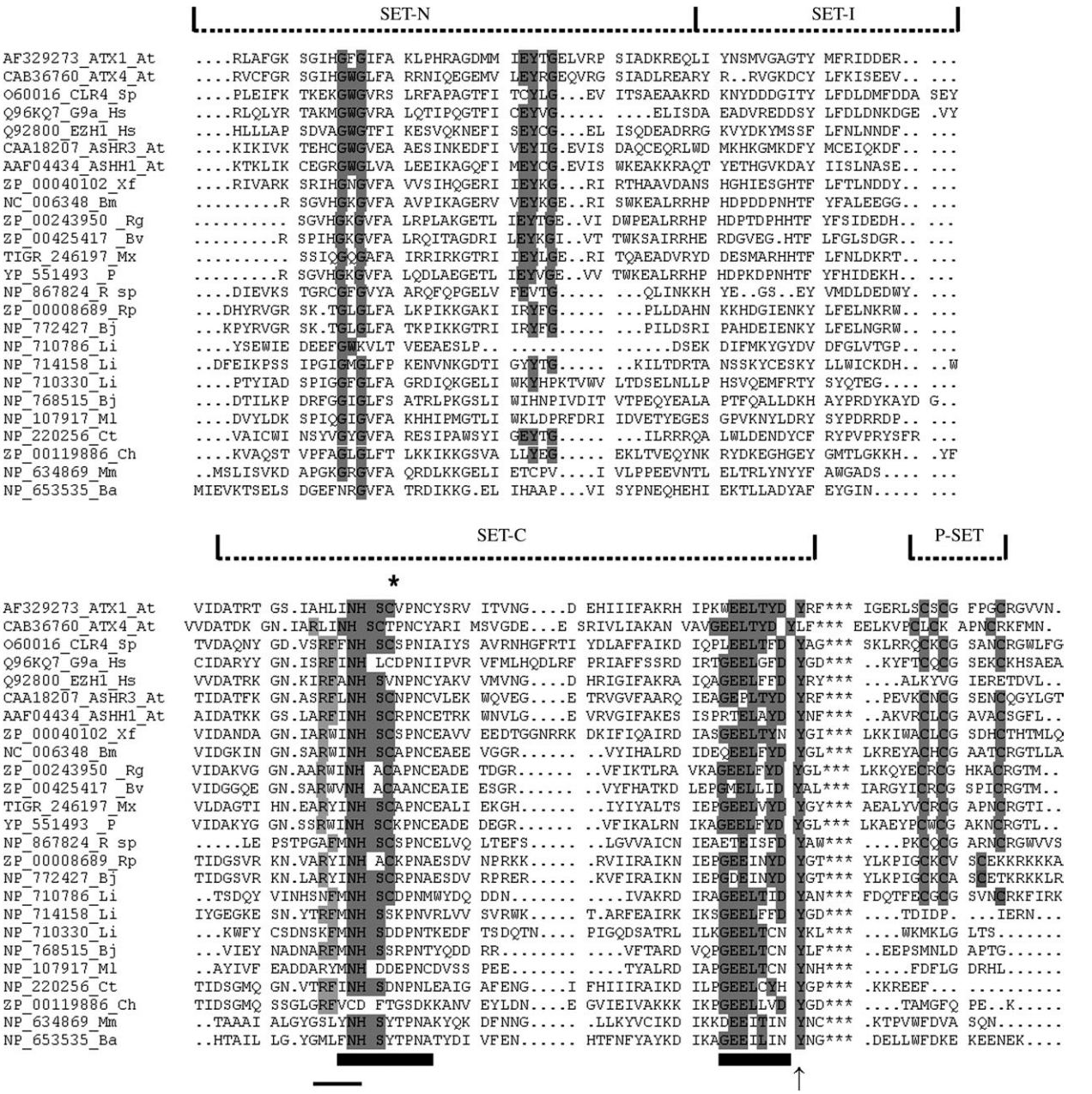


FIG. 3.—Sequence alignment of bacterial and selected eukaryotic SET domain peptides. Sequences of the SET-N, SET-I, and SET-C boxes are aligned. Identical residues are shaded in black, conserved in gray. The star marks the cystein residue from the conserved NHXXCXPN box. The conserved C-residues from the post-SET domain are shown for the proteins of the (+)pSET type (see text; Domain 1). Regions involved in the characteristic "knot" structure are underlined with thick black bars and those involved in the cofactor binding—with a thin bar. The invariant Tyr residues implicated in the catalytic methyltransferase activity is indicated with an arrow (indications are according to Marmostein 2003).

spectively), and all 4 proteins form a common clade. We may propose that the gene has been present in an ancestor before the separation of the *L. interrogans* from the *Me. loti–Br. japonicum* branch but later, after the separation of *Me. loti* and *Br. japonicum*, the gene has been duplicated in *Me. loti*. The fact that the 2 *Me. loti* proteins share higher homology between themselves (2×10^{-28}) than to any other SET domain protein supports this model.

The relationships between the SET domain proteins of the α-proteobacterial rhizobial species, however, are more complex due to facts that none of the other available sequenced genomes of rhizobial symbionts carry *SET* domain genes and that in *Br. japonicum*, there is a second copy of a *SET* domain gene less similar to the other *Br. japonicum* and *Me. loti* genes. This *Br. japonicum* NP_772427 protein is of the (+)pSET type and is phylogenetically related and highly similar to the protein found in the free-living α-proteobacterium, *Rh. palustris* ZP_0000868 (8×10^{-67}; 80% identical). The post-SET domain motif (CXCXXC) in the proteins of these species slightly differs from the (CXCXXXXC) motif found in all other known SET domain proteins (fig. 3). Compared with other bacterial and eukaryotic SET domain proteins, the *Br. japonicum* and *Rh. palustris* proteins showed similarity only to one of

the *L. interrogans* proteins, NP_714158 (7×10^{-11} and 2×10^{-11}, respectively). The 2 copies of *Br. japonicum* were less similar to each other (1×10^{-5}, 27% identical) than to proteins found in other species, indicating gene paralogy.

Mesorhizobium loti, *Br. japonicum*, and *Rh. palustris* are of a common ancestral origin (Larimer et al. 2004) and, thus, their *SET* domain genes may reflect the evolutionary history of the genes. We may suggest that their common ancestor has carried 2 paralogous genes, a *(+)pSET* and a*(−)pSET*: *Br. japonicum* has inherited both, *Rh. palustris* has inherited the *(+)pSET*, whereas *Me. loti* has inherited the *(−)pSET*. Subsequently, in *Me. loti*, this gene has undergone species-specific duplication (Model, fig. 5).

The L. interrogans Protein NP_714158 and the Cytophaga Subphylum

On the phylogenetic trees, the *L. interrogans* protein, NP_714158, clusters with the 2 copies of *Cy. hutchinsonii*, a member of the Cytophaga–Flavobacterium–Bacteroides subphylum (Gherna and Woese 1992) (figs. 1 and 2; supplementary figs. SF2 and SF3, Supplementary Material online). The 2 (−)pSET *Cytophaga* proteins (most similar to each other, $3e^{-09}$, than to any other SET domain protein) apparently represent a gene duplication event that had occurred after the separation of the species from its closest SET domain relatives. Thereby, the *L. interrogans* protein NP_714158 represents a version of the (−)pSET domain protein existing in bacteria before the separation of *Spirochatales*, *Chlamydiae*, *Cytophaga*, *Chlorobia*, and *Bacillae* (fig. 5).

The SET Domain Genes of Chlamydia

A distinctive Gram-negative family of uncertain evolutionary origin deeply separated from other eubacteria, *Chlamydiae* is thought to represent one of the kingdom-level branches on the phylogentic tree (Pace 1997). A distant relationship to Planctomycetales (Amann et al. 1997) and Cyanobacteria (Brinkman et al. 2002) was suggested. On the 23S RNA and the S50-L3 protein trees, *Chlamydiacea* did not show supported relationship to any particular bacterial group (supplementary figs. SF4 and SF5, Supplementary Material online).

Phylogenetic analysis of the SET domain proteins positioned *Chlamydiacea* in a clade of its own, among the (−)pSET domain–type bacteria, with no significant relationship to the eukaryotic cluster (figs. 1 and 2). Compared with eukaryotic and bacterial SET domain proteins available in the database, the chlamydial SET domain proteins showed significant similarity only to each other (1×10^{-86}–1×10^{-65}) and to one of the *L. interrogans* proteins, NP_714158 (6×10^{-11}). This gene version is related also to the proteins found in *Br. japonicum*, *Cytophaga*, *Chlorobium*, and *Bacillus* (see also fig. 5), suggesting ancestral bacterial origins for the chlamydial genes. Interestingly, the SET domain protein of the environmental *V. spinosum* (Chlamydiae/Verrucomicrobia group) belongs to the (+)pSET domain type. A plausible scenario is that a common ancestor has carried both the *(+)pSET* and the *(−)pSET* ancestral forms and that *Chlamydiae* have inherited one of the ancestral copies, the Verrucomicrobia lineage has kept the other, whereas the parachlamydial environmental relative, UWE25, has lost both.

Thereby, the phylogenetic relationships of the SET domain proteins of the Chlamydiae–Verrucomicrobia group do not support a horizontal gene transfer from a mammalian genome, as earlier suggested (Stephens et al. 1998; Aravind and Iyer 2003). The SET domain proteins of *V. spinosum* and *Chlamydiae* cluster into separate clades representing 2 paralogs bared by a common ancestor.

The SET Domain Proteins of Species Unrelated by Origin Cluster Together

Unrelated by the species origin, the SET domain proteins of the free-living green-sulfur photosynthetic bacterium *Ch. tepidum*, of 3 archaeal (*Methanosarcina*) species, *Methanosarcina acetivorans*, *Methanosarcina barkeii*, and *Methanosarcina mazei*, and of 3 *Bacillus* species, share a well-supported clade. It is remarkable that the proteins from these 3 groups always segregate together, independent of the method of tree construction (figs. 1 and 2; supplementary figs. SF2 and SF3, Supplementary Material online). The *SET* domain genes of the *Bacilli*, *Chlorobium*, and *Methanosarcina* are closer than relationship between their genomes, in general. *Chlorobium tepidum* has many genes for metabolic processes that are more similar to archaeal species than to other bacteria, and it was suggested that extensive HGT between *Archaea* and hot-spring bacteria had occurred (Nelson et al. 1999). *Archaea* has been defined as a separate domain of life (Woese and Fox 1977), although many of its metabolic pathways resemble more bacterial than eukaryotic counterparts, suggesting exchange between bacteria and *Archaea* (Doolittle and Logsdon 1998; Koonin et al. 2001). *Methanosarcinae* have large genomes and numerous COGs that no other *Archaea* members have, but, notably, most of these functions are present in various bacteria. *Methanosarcinae* are considered a "sink" for horizontally acquired bacterial genes. About one-third of *M. mazei* ORFs have significant hits in the bacterial genomes referred to as "bacteria-like" (Deppenmeier et al. 2002). It is thought that HGT from bacteria, Gram-positive in particular, has played an important evolutionary role in shaping its physiology (Brown 2003).

Within the large group of sequenced Gram-positive bacteria, only the 3 closely related *Bacillus* species, *Ba. anthracis*, *Ba. cereus*, and *Ba. thuringiensis*, have *SET* domain genes. They are 100% identical among themselves and surprisingly similar to the SET domain proteins from the 3 *Methanosarcinal* species (2×10^{-25}, 51% identical). Because genome-tree analysis has unequivocally supported the monophyly of Archaea, the bacteria-like metabolic genes found in *Archaea* members might have been transferred from bacteria (Galagan et al. 2002). The segregation of the *Methanosarcinae* and the *Bacillus* SET domain proteins in the same clade is also consistent with an HGT (Stanhope et al. 2001).

The relationships are more complex due to the clustering in the same clade of the SET domain protein of *Chlorobium*. Whole-genome analysis has suggested that HGT might have occurred between thermophylic eubacteria and archaea (Nelson et al. 1999; Eisen et al. 2002). However,

Methanosarcinae do not live under extreme conditions and do not share habitats with *Ch. tepidum*, but they may with *Ba. anthracis*, *Ba. cereus*, and *Ba. thuringiensis* (Bintrim et al. 1997; Radnedge et al. 2003). To explore possibilities for a lateral exchange of genetic material between bacteria, in general, and between the *Bacillae* and *Methanosarcinae*, in particular, we examined the genomic regions surrounding the *SET* domains in all bacterial genomes for molecular signs of HGT at the chromosomal level.

Analysis of the SET Domain–Containing Genomic Regions for Signs of Insertion

A major argument in support of a horizontal transfer of *SET* domain genes to bacteria has been the fact that the majority of sequenced bacterial genomes do not carry the gene. The patchy distribution of genes and gene clusters on prokaryote chromosomes reveals an underlying process of recurrent loss and gain (Doolittle 1999; Ochman et al. 2000; Makarova and Koonin 2003).

Although our phylogenetic data did not support eukaryotic origins of the bacterial *SET* domain genes, they left open a possibility that *SET* domain genes might have been exchanged between genomes of bacteria and/or Archaebacteria. To explore this possibility, we analyzed the chromosomal regions around the *SET* domain genes for recognizable signs. Hallmarks of HGT are presence of specific elements affecting integration, such as remnants of mobile elements, rearrangements bordered by identical copies of insertion sequences (IS), clusters of transposases, and interrupted synteny. Extra chromosomes, plasmid vectors, or pathogenicity- and symbiosis islands on the main chromosomes of some pathogenic and symbiont species are considered potential vehicles of foreign genes (Salanoubat et al. 2002; Parkhill et al. 2003). Indirect evidence for ancestral versus horizontal origin of bacterial genes is their GC content. The average GC ratio is considered characteristic for a particular microbial genome, and regions in the DNA with changed ratios are thought to reflect recent horizontal transfer. Consequently, we analyzed the genome arrangements and the GC contents of the *SET* domain genes in the genomes of parasitic and symbiont species as well as in the *Chlorobium*, *Bacilli*, and *Methanosarcinae* clade.

The SET Domain Genes in All Pathogenic and Symbiont Species Are Chromosomally Located

γ- and β-Proteobacterial Genomes. In γ- and β-proteobacterial genomes, genes involved in pathogenicity or other kinds of host–bacterial interaction systems are located on specific "islands" bordered by phage integrases. Whole-genome differences between the *Xylella* and *Xanthomonas* families are limited to phage-associated chromosomal rearrangements and deletions (Bhattacharyya et al. 2002; Van Sluys et al. 2003). The *SET* domain genes in *Xylella*/*Xanthomonas* genomes are not associated with the pathogenicity islands (Alfano and Collmer 1997; Galan and Collmer 1999) and, according to consensus criteria, the *SET* domain genes are not involved in interactions with the host. The GC content of the *SET* domain genes of *Xanthomonas* and *Xylella* are 49–51%, and their overall genomic contents are 51–53%, respectively.

Phylogenetically, *X. fastidiosa* is placed at the base of the γ-Proteobacteria (Bhattacharyya et al. 2002), implying that the remaining members of the subfilum have inherited the SET domain gene. The chromosomal location of the *SET* domain gene is consistent with a role in bacterial functions not necessarily related to interactions with the host.

β-Proteobacterial Genomes. Among β-proteobacterial genomes, *SET* domain genes are found in virulent species from the Bordetella family and in the large Burkholderia family including bacteria with broad ecological habitats: environmental biodegradative bacteria (*B. fungorum*), plant pathogens (*Ralstonia solanacearum*), and free-living soil bacteria (*Ralstonia matallidurans*). The *SET* domain genes are among the conserved core genes shared by the 3 *Bordetella* (*Bordetella bronchiseptica*, *Bordetella pertusis*, and *Bordetella parapertusis*) genomes. The 3 *Bordetella* genomes are highly rearranged, each rearrangement bordered by identical copies of IS 1001 or IS 1002 elements (Parkhill et al. 2003). Despite an overall lack of colinearity of the 3 chromosomes, the *SET* genes are found in remarkably conserved syntenic regions away from the virulence systems. This implies that the *SET* domain genes, together with linked sequences come from a common ancestor. The *SET* domain genes of the 3 *Bordetella* species have GC content (64–66%) comparable to the reported respective genomic contents (67–69%).

The *SET* domain genes in *R. solanacearum*, *B. fungorum*, and *R. matallidurans* are on the main chromosomes, away from regions associated with effectors of host interactions or the ability to colonize (Coenye and Vandamme 2003). The GC content for the *SET* genes of *B. fungorum*, *R. matallidurans*, and *R. solanacearum* (61–65%) are comparable to the GC ratios of overall chromosomal contents (62–66%) supporting ancestral origins for the *SET* domain genes. Furthermore, in *B. fungorum* and *R. matallidurans*, the *SET* domain genes are in largely syntenic regions, whereas in *R. solanacearum*, the synteny is "broken" due to the abundant presence of integrases and transposases in this genome (Salanoubat et al. 2002). This fact is relevant because the conserved synteny in the *SET* domain regions of *B. fungorum* and *R. matallidurans* indicates that the gene has been present before the separation of these species.

α-Proteobacteria. The α−Proteobacteria, *Me. loti* and *Br. japonicum*, provide examples of species involved in symbiotic relationships with eukaryotes. It is thought that the rhizobial lineages have diverged well before the evolution of the legumes and that the genes for the symbiosis were subsequently acquired by lateral transfer (Mergaert et al. 1997; Broughton and Perret 1999). *Mesorhizobium loti* carries a mobile symbiotic island that can convert a soil saprophyte into a symbiont (Sullivan et al. 2002). The pair of the *SET* genes, however, is on the chromosome flanking the island, making it unlikely that they belong to the mobile category. In the closely related *Br. japonicum*, the putative island is not known to be mobile providing no basis for possible mobility of its *SET* domain genes (Goettfert et al. 2001; Kaneko et al. 2002). The GC contents are 60.4% and 61.2% for the *Me. loti* and 62% for the *Br. japonicum (−)pSET* domain genes, whereas the genomic GC content is 64% for *Me. loti* and 66% for *Br. japonicum*.

The second, *(+)pSET* domain, copy of *Br. japonicum* is in a 90-kb region colinear with *Rh. palustris*. This remarkable synteny is interrupted by a block of genes, with best hits to genes found in other β- (*B. fungorum*, *R. matallidurans*, and *R. solanacearum*) and γ- (*Xanthomonas axonopodis* and *Xanthomonas campestris*) bacterial chromosomes. These findings are relevant because they reflect the close relationships between the α-, β-, and γ-proteobacterial species, supporting a common origin for the *(+)pSET* proteobacterial genes.

The Genomes of Chlamydiae. The genomes of *Chlamydiae* are small, preserving only the minimum of genes required for their exclusive lifestyle as obligatory intracellular parasites. It was suggested that parasitic *Chlamidiae* have recruited *SET* domain genes from eukaryote hosts and are using them as an invading tool (Stephens et al. 1998; Aravind and Iyer 2003). However, our phylogenetic analysis did not support a relationship between the chlamydial and eukaryotic *SET* domain genes (fig. 1; supplementary fig. SF2, Supplementary Material online). Whole-genome analysis of *Chlamydiacea* revealed absence of genes typical for chromosomal rearrangements, indicating lack of mechanisms for entry of foreign DNA (Read et al. 2000, 2003; Horn et al. 2004). It was suggested that *Chlamydia* and *Chlamydophila* could not exchange genes with their hosts, or other bacteria, and that the eukaryotic-similar genes found in their genomes reflected not lateral, but ancient evolutionary relationships (Brinkman et al. 2002). The GC content for all genes (or ORFs) of chlamydial species infecting only humans is ~41%. This is much lower than the GC contents of their mammalian hosts (~52%) and may be taken as evidence of a lack of exchange with eukaryotic DNA. The GC content of the *Chlamydia*/*Chlamydophila SET* domain genes is 41.7–43.1%, in good agreement with the other ORFs (Brinkman et al. 2002). In addition, the significant synteny around the *SET* domain genes on all chlamydial chromosomes (data not shown) is inconsistent with a recent transfer from a host genome.

A Possible HGT of SET Domain Genes among Bacterial and Archaebacterial Species

Our phylogenetic analysis did not exclude possible horizontal transfers involving the free-living hot-spring bacterium *Ch. tepidum*, the *Bacilli*, and the *Methanosarcinae* species (figs. 1 and 2; supplementary figs. SF2 and SF3, Supplementary Material online). It is thought that *Methanosarcinae* owe much of its ecological success to ancient HGT. Nearly 30% of its genes are considered to be of bacterial origin (Deppenmeier et al. 2002). Among *Archaea*, only the 3 mesophylic *Methanosarcinae* species have a *SET* domain gene. Examination of the DNA regions flanking the *SET* domain genes in the genomes of *Chlorobium*, *Bacilli*, and the *Methanosarcinae* support an exchange that might have involved members of the 3 groups.

The *Chlorobium* SET domain protein (NP_661845) is most similar to the *Methanosarcinal* NP_634869 (5×10^{-11}) and to the *Bacillus* NP_834732 (1×10^{-07}) SET domain proteins. The genes flanking the *Chlorobium SET* domain gene (shaded area in fig. 4) have best hits in the genomes of the 3 *Methanosarcinae* and the 3 *Bacilli*, although these highly similar genes are not in the immediate vicinity of the *Methanosarcinal* and *Bacillal SET* domain genes (fig. 4). *Chlorobium* genes outside the shaded box are found in a broad spectrum of bacteria (marked as "bacterial genes"). An exception is NP_661849, having its most similar counterpart in *Met. acetivorans* (1×10^{-0}, 62% identical).

In the 3 *Bacilli*, the *SET* domain genes are in syntenic regions on the main chromosomes. Immediately upstream of the *SET* domain gene is a gene encoding a methyl-accepting chemotaxis protein that has most similar counterparts in the genomes of *Methanosarcina barkerii* and *Met. acetivorans* (1×10^{-23}). This gene is absent from *M. mazei* altogether. At the other flank of the SET domain in *Bacillus* are genes with best hits in the genome of *Chlorobium*.

The genes around the *SET* domain genes in *Met. barkerii* and in *M. mazei* (shaded regions) are closest to genes from the 3 *Bacilli* chromosomes forming short "*Bacillus*-like islands" (fig. 4). We note that the colinearity around the *SET* domain genes is preserved in *M. mazei* and *Met. barkerii*, but not in *Met. acetivorans*, despite the fact that *M. mazei* and *Met. acetivorans* are closer (Galagan et al. 2002). In agreement, the SET domain proteins of *M. mazei* and *Met. acetivorans* are 92% identical, whereas the *M. mazei* and *Met. barkerii* are only 76% identical. Because the genomes of *M. mazei* and *Met. barkerii* are more distantly related, the preserved synteny around the *SET* domain genes indicates that this arrangement has existed in the common ancestor and that in *Met. acetivorans* the gene has been internally relocated after the separation of *Met. acetivorans* from *M. mazei*.

The GC contents of the *SET* domain genes of all species are similar to the overall genome contents and, thus, do not offer a clue as to the origin of the genes. In *Chlorobium*, the established contents for the gene is 57.1% versus the genomic 57%, in the 3 *Methanosarcinae*, 40.6–43.0% versus 41.5–42.7%, and in the 3 *Bacilli*, 33.8–34.3% versus 35.1–35.3% for the *SET* domain genes and reported genomes, respectively. The high similarity between the SET domain proteins of *M. mazei* and *Ba. cereus* (3×10^{-25}, 51% identical), together with their phylogenetic relationship incongruent with the origin of the species, support a lateral exchange.

Evolution of the Bacterial *SET* Domain Genes: A Model

Based on our data, we propose a model for an ancestral origin of the bacterial *SET* domain genes (fig. 5). A central postulate is that an ancient *(+)pSET* gene version has been duplicated and diverged into 2 (*(−)pSET* and *(−)*pSET*) copies and that all 3 gene versions were present in a common bacterial ancestor (CBA). The *SET* domain genes of extant bacterial genomes, accordingly, represent individual copies, or combination of copies, descending from the 3 ancestral genes. The spirochaetae (*L. interrogans*) has inherited (and subsequently diverged) all 3 gene copies; other bacteria carry 1, or combinations of 2, of the primordial lineages. This assumption may account for the relationship between extant bacterial *SET* domain genes and the genes found in *L. interrogans*. For example, the predecessors of *Nostoc* and *Pirellula* lines have inherited the *(+)pSET* gene version and have modified it in a species-specific mode; the

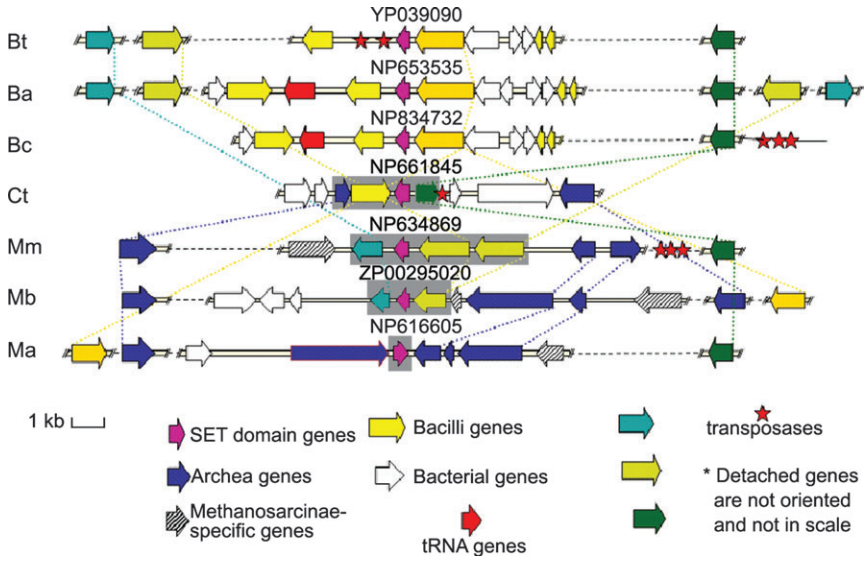


FIG. 4.—Localization of *SET* domain genes on the chromosomes of *Chlorobium*, *Methanosarcinae*, and *Bacillae* species. The chromosome of *Chlorobium tepidum* (Ct) is shown in the middle between the *Methanosarcinae* and the *Bacilli* chromosomes. The shaded region carries the SET domain gene (NP_661845, purple) and the immediately flanking genes showing highest similarity to genes found in *Methanosarcinae* and *Bacillae*. Note that one of the *SET*-flanking genes (NP_661844, yellow) is present only in the *Bacilli*, whereas its upstream neighbor (NP_661843, navy blue) is found only in *Archaea*. The *Chlorobium* gene on the other flank of the *SET* (NP_661846, dark green) showed best hits with genes present in the 3 *Bacilli* and 2 *Methanosarcinae* (*Methanosarcina acetivorans* and *Methanosarcina mazei*) genomes. Outside the boxed area, the *Chlorobium* genes are found broadly distributed among bacterial genomes (white arrows mark "common" bacterial genes). On *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* chromosomes, the *SET* domain genes are in colinear regions. The *SET* domain adjacent genes (orange arrows) code for methyl-accepting chemotaxis protein. These sequences have best hits in the genomes of *M. acetivorans* (1×10^{-23}) and *Methanosarcina barkerii* (4×10^{-23}); no homolog is present in *M. mazei*. On the other flank of *SET* are genes specific for *Bacillus* (yellow arrows). It is remarkable that these genes are not found outside the Bacilli group except in *Chlorobium*. In *M. mazei* and *Met. barkeri*, the chromosomal regions immediately surrounding the *SET* domain genes are collinear (shaded areas). The genes form a short "*Bacillus*-like" island because their best hits are found in the genomes of *Ba. anthracis*, *Ba. cereus*, and *Ba. thuringiensis*. However, in *Bacillae*, these genes are elsewhere in the genome, unlinked to SET (pistachio and turquoise arrows). Notably, the *M. acetivorans SET* domain gene location is in a different, not syntenic, region with those of its close relatives *M. mazei* and *Met. barkeri*: it is surrounded by genes found in *Archaebacteria*, in general (navy blue arrows). Despite the fact that the genomes of *M. mazei* and *M. acetivorans* are more closely related and that their *SET* domain genes are 92% identical, lack of colinearity indicates that *M. acetivorans* has undergone rearrangements in this region relocating the *SET* domain gene from the syntenic regions preserved in *M. mazei* and *Met. barkerii*. Similar genes found on all 7 chromosomes are color coded and connected by broken lines. Numbers on top of the *SET* domain genes (purple arrows) correspond to their database IDs. The genes from *Ba. anthracis*, corresponding to the genes from the "*Bacillus*-like" island in *M. mazei* and *Met. barkerii*, are in inverted positions relative to the *SET* domain loci in *Ba. cereus* and *Ba. thuringiensis*. Genes unlinked to SET are separated by broken lines, not drawn to scale. Red arrows indicate tRNA genes. Red stars indicate transposase genes.

putative ancestor of the *Chlamydia*/Verrucomicrobia group has had 2 genes: 1 *(+)pSET* and 1 *(−)pSET* version. The Verrucomicrobia lineage has preserved the *(+)pSET* gene, whereas extant *Chlamydiae* have kept the *(−)pSET* copy. The same *(−)pSET* version has been inherited also by the *Cytophaga*, *Chlorobium*, and *Bacillus* lineages where it has undergone an evolution of its own and was, perhaps, transferred to the *Methanosarcinae*.

According to our model, the common ancestor of Proteobacteria has retained 2 ancestral copies: the *(+)pSET* and a *(−)*pSET* different from the version retained by the ancestors of *Cytophaga*, *Chlorobium*, and *Bacillus* described above. Furthermore, during the proteobacterial evolution, the ancestor of the α-lineage has retained both copies, whereas the progenitor of the β-, γ-, and δ-groups has kept only the ancestral *(+)pSET* copy. Subsequently, in the α-subdivision, *Br. japonicum* has inherited both ancestral *(+)pSET* and *(−)*pSET* copies, and *Rh. palustris* has kept only the *(+)pSET* copy, whereas *Me. loti* has retained only the *(−)*pSET* version. In *Me. loti*, the gene has been subsequently duplicated for species-specific functions, absent from *Br. japonicum* and *Rh. palustris*.

The *(+)pSET* gene version has been largely conserved in Proteobacteria, has been duplicated, and has evolved for bacteria-specific roles. The high similarity between the (+)pSET domain proteins of Proteobacteria and the Verrucomicrobial member *V. spinosum* might reflect the striking conservation of a gene inherited from the CBA. The monophyletic origin of the β- and γ-subdivisions is congruent with ancestral relationships among their *SET* domain genes. Furthermore, the similarity between the eukaryotic SET domain proteins and the proteins of *Xanthomonas*/*Xylella* may have another interesting implication *vis a vis* a recent hypothesis that the eukaryotic genome has arisen from a fusion of a proteobacterium (Pγ) and an archaeal eocyte (Rivera and Lake 2004). This links, at the deepest levels, prokaryotes with eukaryotes and suggests possible common origins for the eukaryotic and the protobacterial *SET* domain genes.

Why Do Bacteria Have *SET* Domain Genes?

Unambiguous answers to this question would be provided by "bench studies," none of which, surprisingly, have been carried out to this date. Based on our analyses, we can make a few predictions. The facts that bacteria do not carry histones, the established substrate for the SET domain activity, and that *SET* domain genes were found predominantly in pathogenic and symbiotic bacteria have suggested

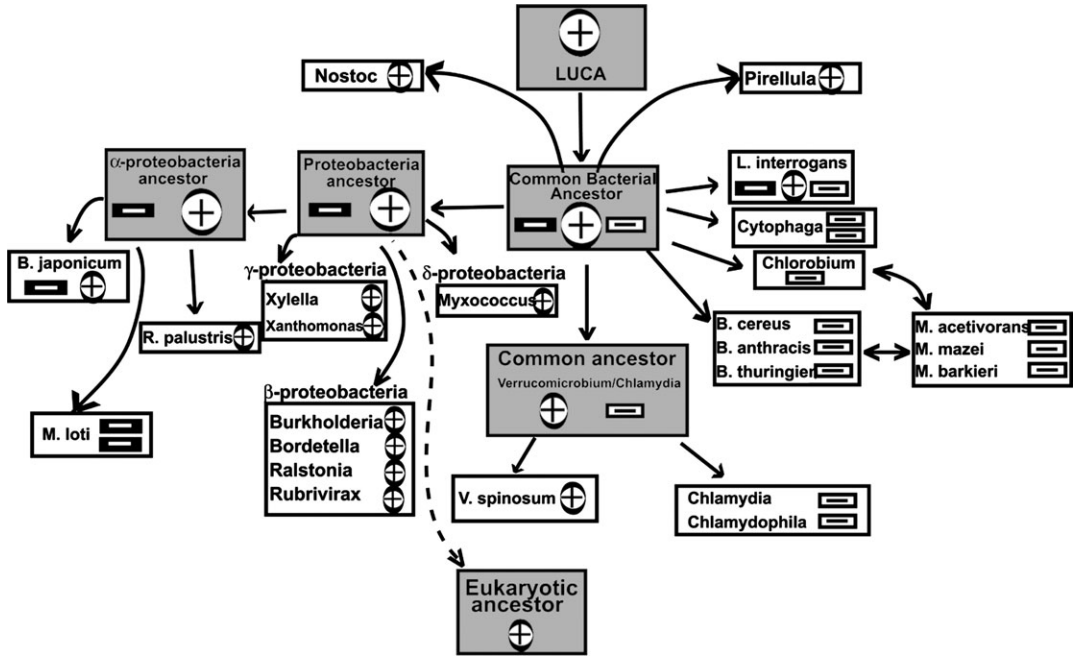


FIG. 5.—A Model for the evolution/distribution of the bacterial *SET* domain genes. A CBA descending from last universal common ancestor carries the 3 SET domain gene versions found in the extant spirochaetae *L. interrogans*. Putative ancestors are shown in shaded boxes, and extant species are in white boxes. The *(+)pSET* gene version is marked by a (+) sign in a circle. The 2 *(−)pSET* versions are marked by a (−) sign in a black or white box, respectively. *Leptospira interrogans* is shown with the there paralogs; the cyanobacterial lineage (*Nostoc*) and the planctomycetum (*Pirellula*) have kept the *(+)pSET* copy from the CBA, whereas extant *Chlamidiae*, *Cytophaga*, *Chlorobium*, and *Bacillus* carry 1 ancestral *(−)pSET* gene copy. In *Cytophaga*, the *(−)pSET* copy has been subsequently duplicated. The putative ancestor of Verrucomicrobia/*Chlamidiae* has carried 2 copies in its genome: 1 *(+)pSET* and 1 *(−)pSET*. The *(−)pSET* lineage is represented in extant pathogenic *Chlamydia*, whereas the *(+)pSET* version is retained in the Verrucomicrobia group. The model illustrates a possible HGT between *Methanosarcinae* and *Bacillus* but does not fix the direction of the transfer, neither does it exclude a possible HGT from *Chlorobium* to *Bacillus* or from *Chlorobium* to *Methanosarcinae*. Double-headed arrows show possible HGT between these groups. The common ancestor of Proteobacteria is presumed to carry the *(+)pSET* and a *(−)pSET* copy different from the *(−)pSET* version found in *Chlamidiae*, *Cytophaga*, *Chlorobium*, and *Bacillus*. The β-, γ-, and δ-lines have inherited the same *(+)pSET* version as the one found in *V. spinosum*. The eukaryotic *SET* domain gene may be ancestrally related to the proteobacterial *(+)pSET* version, accounting for the high similarity found between the *Xylella*/*Xanthomonas* and eukaryotic *SET* domain genes. Lastly, a putative ancestor of the rhizobial α-bacteria has retained 2 of the ancestral copies (*(−)pSET* and *(+)pSET*); *Bradyrhisobium japonicum* has inherited both, whereas *Rhodopseudomonas palustris* has inherited only the *(+)pSET* copy. *Mesorhizobium loti* has inherited only the *(−)pSET* gene and it was duplicated after its separation from *Br. japonicum*.

that bacterial SET domain proteins might be involved in interactions with the host. Extra chromosomes, plasmid vectors, secretion regions, pathogenicity and symbiosis islands in many pathogenic and symbiont species, are potential vehicles for "interacting factors" with the host cells. Our genomic analysis, however, showed that in none of the examined genomes were the *SET* domain genes located at these mobile structures. Because all bacteria have their *SET* domain genes on the main chromosomes, it is unlikely that the gene products are involved in direct host–bacterial interactions.

Most likely, the bacterial *SET* domain genes are involved in bacterial cell-specific functions. For example, the spirochaetae *L. interrogans*, related to the strictly parasitic *B. burgdorferei* and *T. pallidum*, carries 3 *SET* domain genes paralogs, whereas the other 2 genomes carry none. Parasitic bacteria thriving in a more homeostatic niche tend to delete expendable sequences from their genomes (Brinkman et al. 2002). We suggest that the *SET* domain genes provide *L. interrogans* with survival opportunities not needed by its parasitic relatives.

An interesting observation is that bacterial species that undergo unique types of developmental cycles carry *SET* domain genes: *M. xanthus* undergoes developmental regulation to produce multicellular fruiting bodies (Julien et al. 2000) and *Chlamydia* has biphasic developmental cycle controlled by a set of specific genes (Belland et al. 2003). In *C. trachomatis*, the SET domain gene is among the late expressing genes. Should proximity of genes be a reflection of their coordinated function, it might be informative that on all *Chlamydia* chromosomes, the *SET* gene is immediately preceded by a gene encoding a protein involved in cell division (FtsK). On the other flank is a histone-like gene, possibly involved in the compaction of the bacterial chromosome during the formation of the metabolically inactive compact elementary body. It is tempting to suggest that the *SET* domain genes may be involved in bacterial processes that are distant predecessors of eukaryotic developmental mechanisms. In agreement, the only archaebacterial group carrying *SET* domain genes, *Methanosarcinae*, are unique among *Archaea* with their ability to form complex multicellular structures suggestive of differentiation (Galagan et al. 2002).

An often-made reference for a relationship between archaeal and eukaryotic systems is the archaeal chromatin. Many euryarcheota carry homologs of eukaryotic histones that can compact DNA (Reeve et al. 1997). Some carry 1 copy

(*M. mazei*, *Met. acetivorans*, and *Met. barkerii*) and some have more than one that form dimers and tetramers (White and Bell 2002). However, because none of the *Archaea* (except *Methanosarcinae*) carries SET domain genes, it is clear that archaeal histones are not targets of the SET domain protein activity. Moreover, the archaeal histones have the characteristic histone fold, but no tails, indicating that histone-tail modifications do not take place even in species that have the modifying activity. These facts argue that unlike eukaryotes, *Methanosarcinae*, apparently do not use a "histone code" to regulate chromosome activity and gene expression (Strahl and Allis 2000). However, the *M. mazei* SET domain protein carries specific methyltransferase activity for a protein, MC1-α, associating with DNA (Manzur and Zhu 2005). It provided the first example of a SET domain function outside the eukaryotic domain of life.

A major result of our analyses, thereby, is the conclusion that the *SET* domain genes found in extant bacteria are, most likely, of bacterial origin. The presence of *SET* domain genes in members of all main clades of the bacterial domain of life indicated that ancestral versions of the gene have existed before the separation of the known bacterial divisions. The increased sample of sequenced bacterial genomes allowed us to establish presence of *SET* domain genes in free-living and environmental species that are unlikely to have acquired their *SET* domain genes from a eukaryotic donor. Apparently, the initial finding of *SET* domain genes in pathogenic and symbiont species was a result of a biased sample.

The apparently monophyletic origin of the SET domain proteins of the β- and γ-bacteria, congruent with the monophyletic origin of the species, supports ancestral relationships among the bacterial *SET* domain genes. This conclusion is important because it implies that the high similarity of the *Xanthomonas/Xylella* and the eukaryotic SET domain proteins might reflect not horizontal gene transger from eukaryotes but a common origin from distant proteobacterial genomes (Rivera and Lake 2004). The absence of *SET* domain genes in the majority of currently sequenced bacterial genomes may be attributed to gene loss. Gene loss has played a significant role in bacterial genome evolution; unusual bacteria–eukaryotic gene similarity are thought to reflect gene loss in a related lineage (Mira et al. 2001; Salzber and Eisen 2001). The small genomes of obligatory parasites suggested that host adaptations are a consequence of gene loss, not gain of function (Brinkman et al. 2002). This argues against a "gain" of a *SET* gene from the host in parasitic bacteria. Dynamic genome reorganizations, considered typical for bacteria, may account for the overall lack of synteny and the loss of *SET* domain genes. The mosaic of short patchy synteny in genomes of closely related species is evidence of internal genome activity. In genomes where *SET* domain genes were within regions flanked by transposable elements and clusters of transposases, they were associated with internal genome rearrangements rather than accommodation of foreign DNAs. In most cases, the GC contents established for individual bacterial *SET* domain genes did not differ significantly from the GC contents of the host genomes but in the case of *Me. loti* and *Br. japonicum*, the *SET* domain genes had lower GC (60.4% and 61.2%) than their respective genomes (64% for *Me. loti* and 66%). Additional evidence did not support a lateral acquisition of these genes and, thus, it is unclear how these differences in the GC contents of the *SET* domain genes and the genomes carrying them might correlate, or reflect, the history of their origin.

A most compelling argument for a bacterial versus eukaryotic origin of the *SET* domain genes found in bacteria, however, is the evidence that the bacterial *SET* domain genes have undergone an evolution of their own. The segregation of the bacterial SET domain proteins into 2 phylogenetically related Domains (1 and 2) and the presence of multiple *SET* domain gene copies within a single genome are the 2 main lines of support. We consider the loss of the post-SET sequence (appearance of Domain 2 proteins) as a later event in the evolution of the bacterial SET domain function. A few eukaryotic families (E(z), SET7/9, SET8, and RuBisCo) also lack the post-SET domain. However, the evolution of the bacterial *(−)pSET* sequences is apparently independent of the evolution of eukaryotic *(−)pSET*: phylogenetic analysis did not reveal a relationship between the bacterial (−)pSET and the eukaryotic (−)pSET types (not shown).

Presence of multiple *SET* domain genes within a genome reflects duplication events that have taken place at different times. The duplication of the *Me. loti* gene has occurred after its speciation within the α-proteobacterial subdivision, and duplication of the *SET* domain genes in *B. cepacia* and *Ru. gelatinosus* has occurred after the separation of the β- from the other proteobacterial groups, whereas the 2 *Cy. hutchinsonii* genes are species-specific duplication of an ancient *(−)pSET* version after the separation of the spirochaetal, the Bacteroides–Flexobacter, and the chlamydial lineages. Existence of more than 1 copy of *SET* domain genes in the genomes of several species and their segregation into different clades indicates evolution of species-specific and paralogous functions in bacteria.

The phylogenetic and chromosome analyses of *Chlorobium*, *Bacillus*, and *Methanosarcinal* SET domain–containing species supported an HGT between bacteria and *Archeae*. Genomic analyses in the vicinity of the *SET* domain genes and phylogenetic trees are consistent with a lateral exchange between bacterial and *Methanosarcinal* genomes. The *SET* domain genes and their immediate neighbors in *Chlorobium* and in *Ba. thuringiensis*, *Ba. cereus*, and *Ba. anthracis* have best hits in the genomes of the *Methanosarcinal* species and vice versa. It is plausible that a *Methanosarcinal* ancestor has received the *SET* domain gene (together with a few nearby genes) from bacterial donors although the donor (*Chlorobium* or *Bacillus*) cannot be unambiguously defined. *Methanosarcinae* may occupy similar habitats with *Ba. cereus* (Bintrim et al. 1997), suggesting that an ancestor of *Methanosarcinae* might have acquired it from a *Bacillus* ancestor. The GC contents of the *Methanosarcinal SET* domain genes, however, indicate that if a transfer has taken place, it has not been recent as the gene has been successfully "blended" with the genomic sequences.

Supplementary Material

Supplementary figures SF1–SF5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are indebted to F. Doolittle and Camilla Nesbo for critically reading this manuscript and for providing helpful suggestions and encouragement. This work was partially supported by a University of Nebraska Lincoln-Research Cluster Grant Progect grant award and by Molecular and Cellular Biology-0343934 National Science Foundation grant to Z.A.

Literature Cited

Alfano JR, Collmer A. 1997. The type III secretion pathway of plant pathogenic bacteria: trafficking harpins avr proteins, and death. J Bacteriol. 179:5655–5662.

Alvarez-Venegas R, Avramova Z. 2002. SET-domain proteins of the Su(var), E(z) and *Trithorax* families. Gene. 285:25–37.

Amann R, Springer N, Schonhuber W, Ludwig W, Schmid EN, Muller KD, Michel R. 1997. Obligate intracellular bacterial parasites of acanthamoebae related to Chlamydia spp. Appl Environ Microbiol. 63:115–121.

Aravind L, Iyer LM. 2003. Provenance of SET-domain histone methyltransferases through duplication of a simple structural unit. Cell Cycle. 2:369–376.

Baumbusch LO, Thorstensen T, Krauss V, Fischer A, Naumann K, Assalkhou R, Schiltz I, Reuter G, Aalen RB. 2001. The Arabidopsis thaliana genome contains at least 29 active genes encoding SET-domain proteins that can be assigned to four evolutionary conserved classes. Nucleic Acids Res. 29: 4319–4333.

Belland RJ, Zhong GM, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD. 2003. Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. Proc Natl Acad Sci USA. 100:8478–8483.

Bhattacharyya A, Stilwagen S, Ivanova N, et al. (22 co-authors). 2002. Whole-genome comparative analysis of 3 phytopathogenic Xylella fastidiosa strains. Proc Natl Acad Sci USA. 99:12403–12408.

Bintrim SB, Donohue T, Handelsman J, Roberts GP, Goodman RM. 1997. Molecular phylogeny of Archaea from soil. Proc Natl Acad Sci USA. 94:277–282.

Brinkman FSL, Blanchard JL, Cherkasov A, et al. (13 co-authors). 2002. Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Genome Res. 12:1159–1167.

Broughton WJ, Perret X. 1999. Genealogy of legume-Rhizobium symbioses. Curr Opin Plant Biol. 2:305–311.

Brown JR. 2003. Ancient horizontal gene transfer. Nat Rev Genet. 4:121–132.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Coenye T, Vandamme P. 2003. Diversity and significance of Burkholderia. Environ Microbiol. 5:719–729.

Deppenmeier U, Johan A, Hartsh T, et al. (23 co-authors). 2002. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotech. 4:453–461.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2128.

Doolittle WF, Logsdon JM. 1998. Archaeal genomics: do archaea have a mixed heritage? Curr Biol. 8:R209–R211.

Dorman CJ, Deighan P. 2003. Regulation of gene expression by histone-like proteins in bacteria. Curr Opin Genet Dev. 13:179–184.

Eisen JA, Nelson KE, Paulsen IT, et al. (35 co-authors). 2002. The complete genome sequence of Chlorobium tepidum TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci USA. 99:9509–9514.

Felsenfeld G, Groudine M. 2003. Controlling the double helix. Nature. 421:448–453.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (Version 3.2). Cladistics. 5:164–166.

Galagan JE, Nusbaum C, Roy A, et al. (55 co-authors). 2002. The genome of M-acetivorans reveals extensive metabolic and physiological diversity. Genome Res. 12:532–542.

Galan JE, Collmer A. 1999. Type III secretion machines: bacterial devices for protein delivery into host cells. Science. 284:1322–1328.

Gherna R, Woese CR. 1992. A partial phylogenetic analysis of the flavobacter-bacteroides phylum: basis for taxonomic restructuring. Syst Appl Microbiol. 15:513–521.

Glansdorff N. 2000. About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Mol Microbiol. 38:177–185.

Glockner FO, Kube M, Bauer M, et al. (14 co-authors). 2003. Complete genome sequence of the marine planctomycete *Pirellula sp*. strain 1. Proc Natl Acad Sci USA. 100:8298–8303.

Gottfert M, Rothlisberger S, Kundig C, Beck C, Marty R, Hennecke H. 2001. Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the Bradyrhizobium japonicum chromosome. J Bacteriol. 183: 1405–1412.

Horn M, Collingro A, Schmitz-Esser S, et al. (13 co-authors). 2004. Illuminating the evolutionary history of Chlamydiae. Science. 304:728–730.

Ivanova N, Sorokin A, Anderson I, et al. (23 co-authors). 2003. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature. 423:87–91.

Jacobs SA, Harp JM, Devarkonada S, Kim Y, Rastinejad F, Khorasanizadeh S. 2002. The active site of the SET domain is constructed on a knot. Nat Struct Biol. 9:833–838.

Jenuwein T, Allis CD. 2001. Translating the histone code. Science. 293:1074–1080.

Julien B, Kaiser DA, Garza A. 2000. Spatial control of cell differentiation in *Myxococcus xantus*. Proc Natl Acad Sci USA. 97:9098–9103.

Kaneko T, Nakamura Y, Sato S, et al. (24 co-authors). 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res. 7:331–338.

Kaneko T, Nakamura Y, Sato S, et al. (17 co-authors). 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110. DNA Res. 9:189–197.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microb. 55:709–742.

Larimer FW, Chain P, Hauser L, et al. (19 co-authors). 2004. Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. Nature Biotech. 22:55–61.

Makarova KS, Koonin EV. 2003. Comparative genomics of archaea: how much we learned in six years and what's next? Genome Biol. 4:115.

Manzur KL, Zhu M-M. 2005. An archaeal SET domain protein exhibits distinct methyltransferase activity towards DNA-associated protein MC1-a. FEBS Lett. 579:3859–3865.

Marmostein R. 2003. Structure of SET domain proteins: a new twist on histone methylation. Trends Biochem Sci. 28:59–62.

Mergaert P, Van Montagu M, Holsters M. 1997. Molecular mechanisms of Nod factor diversity. Mol Microbiol. 25:811–817.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:580–589.

Nelson KE, Clayton RA, Gill SR, et al. (29 co-authors). 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritime*. Nature. 399:323–329.
Ochman HM, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304.
Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science. 276:734–740.
Page RDM.1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 12:357–358.
Parkhill J, Sebaihia M, Preston A, et al. (53 co-authors). 2003. Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet. 35:32–40.
Radnedge L, Agron PG, Hill KK, Jackson PJ, Ticknor LO, Keim P, Andersen GL. 2003. Genome differences that distinguish Bacillus anthracis from Bacillus cereus and Bacillus thuringiensis. Appl Environ Microbiol. 69:2755–2764.
Rea S, Eisenhaber F, O'Carroll D, et al. (11 co-authors). 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature. 406:593–599.
Read TD, Brunham RC, Shen C, et al. (25 co-authors). 2000. Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. Nucleic Aids Res. 28:1397–1406.
Read TD, Myers GSA, Brunham RC, et al. (21 co-authors). 2003. Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Aids Res. 31:2134–2147.
Reeve JN, Sandman K, Daniels CJ. 1997. Archaeal histones, nucleosomes, and transcription initiation. Cell. 89:999–1002.
Ren S-X, Gu G, Jiang X-G, et al. (39 co-authors). 2003. Unique physiological and pathogenic features of Leptospira interrogans revealed by whole genome sequencing. Nature. 422:888–892.
Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature. 431:152–155.
Salanoubat M, Genin S, Artiguenave F, et al. (28 co-authors). 2002. Genome sequence of the plant pathogen Ralstonia solanacearum. Nature. 415:497–502.
Salzber SL, Eisen JA. 2001. Lateral gene transfer or viral colonization? Science. 293:1048.
Schlesner H. 2004. The Prokaryotes: an evolving electronic resource for the microbiological community: the genus Verrucomicrobium. New York: Springer-Verlag.
Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. 2001. Phylogenetic analysis do not support horizontal gene transfer from bacteria to vertebrates. Nature. 411:940–943.
Stephens RS, Kalman S, Lammel C, et al. (12 co-authors). 1998. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science. 282:754–759.
Strahl BD, Allis CD. 2000. The language of covalent histone modifications. Nature. 403:41–45.
Sullivan JT, Trzebiatowski JR, Cruickshank RW, et al. (14 co-authors). 2002. Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A. J Bacteriol. 184:3086–3095.
van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD. 2003. Discordant phylogenies within the rrn loci of rhizobia. J Bacteriol. 185:2988–2998.
Van Sluys MA, de Oliveira MC, Monteiro-Vitorello CB, et al. (57 co-authors). 2003. Comparative analysis of the complete genome sequences of Pierse's disease and citrus variegated chlorosis strains of Xylella fastidiosa. J Bacteriol. 185:1018–1026.
White MF, Bell SD. 2002. Holding it together: chromatin in archaea. Trends Genet. 18:621–626.
Woese CR. 1998. The universal ancestor. Proc Natl Acad Sci USA. 95:6854–6859.
Woese CR. 2000. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA. 97:8392–8396.
Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA. 74:5088–5090.
Xiao B, Wilson JR, Gamblin SJ. 2003. SET domains and histone methylation. Curr Opin Struct Biol. 13:699–705.

Jonathan Eisen, Associate Editor

Accepted November 9, 2006

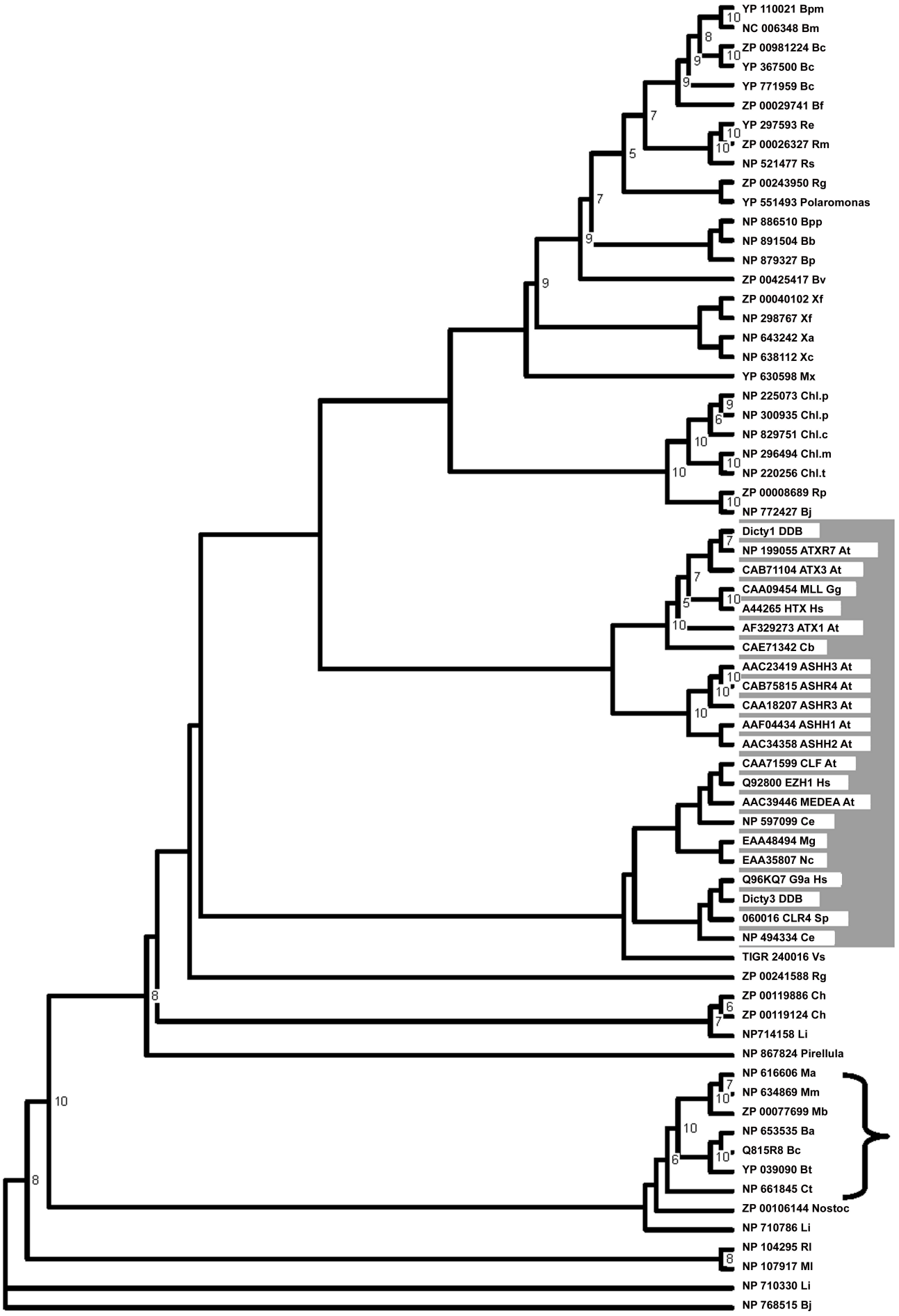
Figure SF1. Multiple Alignment of Eukaryotic and Bacterial SET Domain Sequences used for Tree-Reconstruction

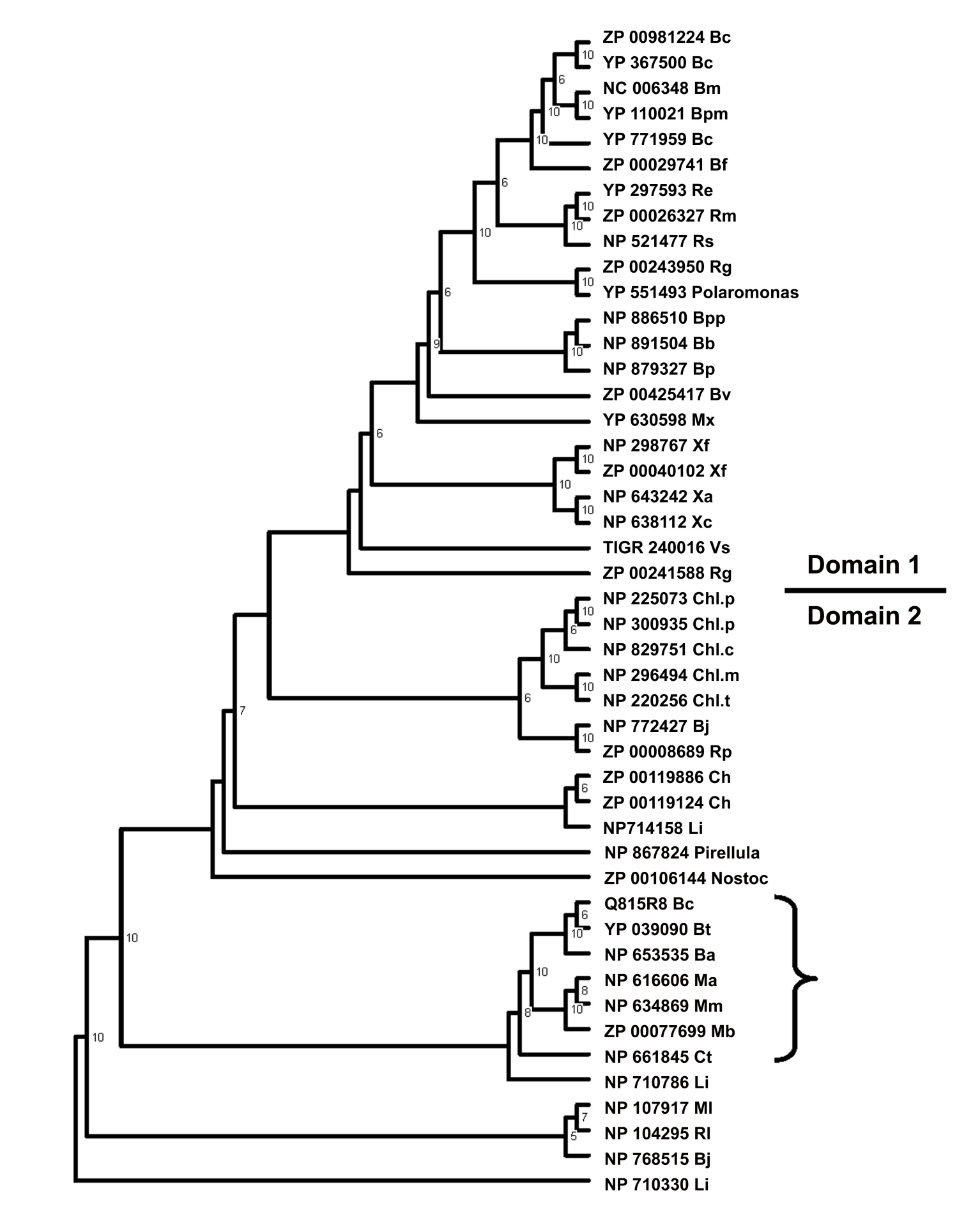
```
NP_768515    ----ILKPDR FGGIGLFSAT RLPKGSLIWI HNPIVDITVT ----PEQYEA
NP_710330    --PTYIADSP IGGFGLFAGR DIQKGELIWK YHPKTVWVLT ----DSELNL
NP_104295    --RTYVAASA IEGVGMFAAE PIRKGASIWR LNPDFDRLIP ----MDEYEA
NP_107917    --DVYLDKSP IQGIGVFAKH HIPMGTLIWK LDPRFDRIID ----VETYEG
NC_006348    -------RSG VHGKGVFAAV PIKAGERVVE YKGERISWKE ----ALRRH-
YP_110021    -------RSG VHGKGVFAAV PIKAGERVVE YKGERISWKE ----ALRRH-
ZP_00981224 -------RSG VHGKGVFAVA PIKAGERVVE YKGERISWKE ----ALRRH-
YP_367500    -------RSG VHGKGVFAVA PIKAGERVVE YKGERISWKE ----ALRRH-
YP_771959    -------RSG IHGKGVFAVE PIKAGERVVE YKGERISWKE ----ALRRH-
ZP_00029744 --RIAVRRSG VHGKGVFAVE PIAAGERLIE YKGERISWKE ----ALRRH-
ZP_00026327 --RIEVRQSG VHGKGVYAIG QIAEGERVIE YKGEHISWKE ----ALKRH-
YP_297593    -------QSG VHGKGVYAIA PIAEGERVIE YKGEHISWKK ----ALDRH-
NP_521477    --RIAVRESG VHGRGVYAVA AIAKGKKIIE YKGEHISWKE ----ALRRH-
ZP_00243950 --------SG VHGKGVFALR PLAKGETLIE YTGEVIDWPE ----ALRRH-
YP_551493    -------RSG VHGKGVFALQ DLAEGETLIE YVGEVVTWKE ----ALRRH-
NP_879327    --WHSVRRSR LHGNGVFATR KIPAGTRIIE YGGKRISAEE ----ADRRH-
NP_891504    --WHSVRRSR LHGNGVFATR KIPAGTRIIE YGGKRISAAE ----ADRRH-
NP_886510    --WHSVRRSR LHGNGVFATR KIPAGTRIIE YGGKRISAAE ----ADRRH-
ZP_00425417 -------RSP IHGKGVFALR QITAGDRILE YKGIVTTWKS ----AIRRH-
ZP_00040102 --RIVARKSR IHGNGVFAVV SIHQGERIIE YKGRIRTHAA ----VDAN--
NP_298767    ---------- -----MFAVV SIHQGERIIE YKGRIRTHAA ----VDAN--
NP_643242    ---------- -----MFALA PLRKGERIIQ YKGRLRTHAE ----VDAD--
NP_638112    ---------- -----MFAVA ALSKGERIIQ YKGRLRTHAE ----VDAD--
YP_630598    --------SS IQGQGAFAIR RIRKGTRIIE YLGERITQAE ----ADVRY-
O60016       -------KTK EKGWGVRSLR FAPAGTFITC YLGEVITSAE ----AAKRDK
Q96KQ7       -------RTA KMGWGVRALQ TIPQGTFICE YVGELISDAE ----ADVRE-
NP_494334    -------RDP WCGWGVRASV DIAFGTFIGE YAGELIDDEE ----AMDRH-
AAF04434     -------KCE GRGWGLVALE EIKAGQFIME YCGEVISWKE ----AKKRAQ
AAC34358     -------QSG KKGYGLRLLE DVREGQFLIE YVGEVLDMQS ----YETRQK
AAC23419     -------QTE KCGSGIVAEE EIEAGEFIIE YVGEVIDDKT ----CEERLW
CAB75815     -------QTE KCGYGIVADE DINSGEFIIE YVGEVIDDKI ----CEERLW
CAA18207     -------KTE HCGWGVEAAE SINKEDFIVE YIGEVISDAQ ----CEQRLW
EAA35807     -------QLE GCGYGLFTAE DISQDEFVIE YTGELITHDE GVRREARRGE
EAA48494     -------GIE GCGYGLFTAV DIAADEFIIE YVGELIQHDE GVRREARRGN
Q92800       -------PSD VAGWGTFIKE SVQKNEFISE YCGELISQDE ----ADRRGK
CAA71599     -------ISD ISGWGAFLKN SVSKHEYLGE YTGELISHKE ----ADKRGK
AAC39446     -------KSD VHGWGAFTWD SLKKNEYLGE YTGELITHDE ----ANERGR
TIGR_240016 --------SQ IHGRGLYARK AIPKDTWIVE YVGERVDKDE ----SDRRAN
ZP_00241588 --------SR IDGQGAFAAE AIPARRKIGE IRGESISVRE ----ARRRA-
A44265       -------RSP IHGRGLFCKR NIDAGEMVIE YAGNVIRSIQ ----TDKREK
CAA09454     -------RSP IHGRGLFCKR NIDAGEMVIE YSGNVIRSIL ----TDKREK
AF329273     -------KSG IHGFGIFAKL PHRAGDMMIE YTGELVRPSI ----ADKREQ
NP_199055    -------QSK IHDWGLVALE PIEAEDFVIE YVGELIRSSI ----SEIRER
NP_587812    -------PSR IHTLGLFAME NIDKNDMVIE YIGEIIRQRV ----ADNREK
CAB71104     ---------- ---------- ---DG--IIE YRGVKVRRSV ----ADLREA
NP_194512    -------RSG IHGWGLFARR NIQEGEMVLE YRGEQVRGSI ----ADLREA
CAE71342     -------RSR IAGLGLYAKT DIPMGEYIIE YKGEIIRSEL ----CEVREK
ZP_00008689 --DHYRVGRS KTGLGLFALK PIKKGAKIIR YFGPLLDAHN ----------
```

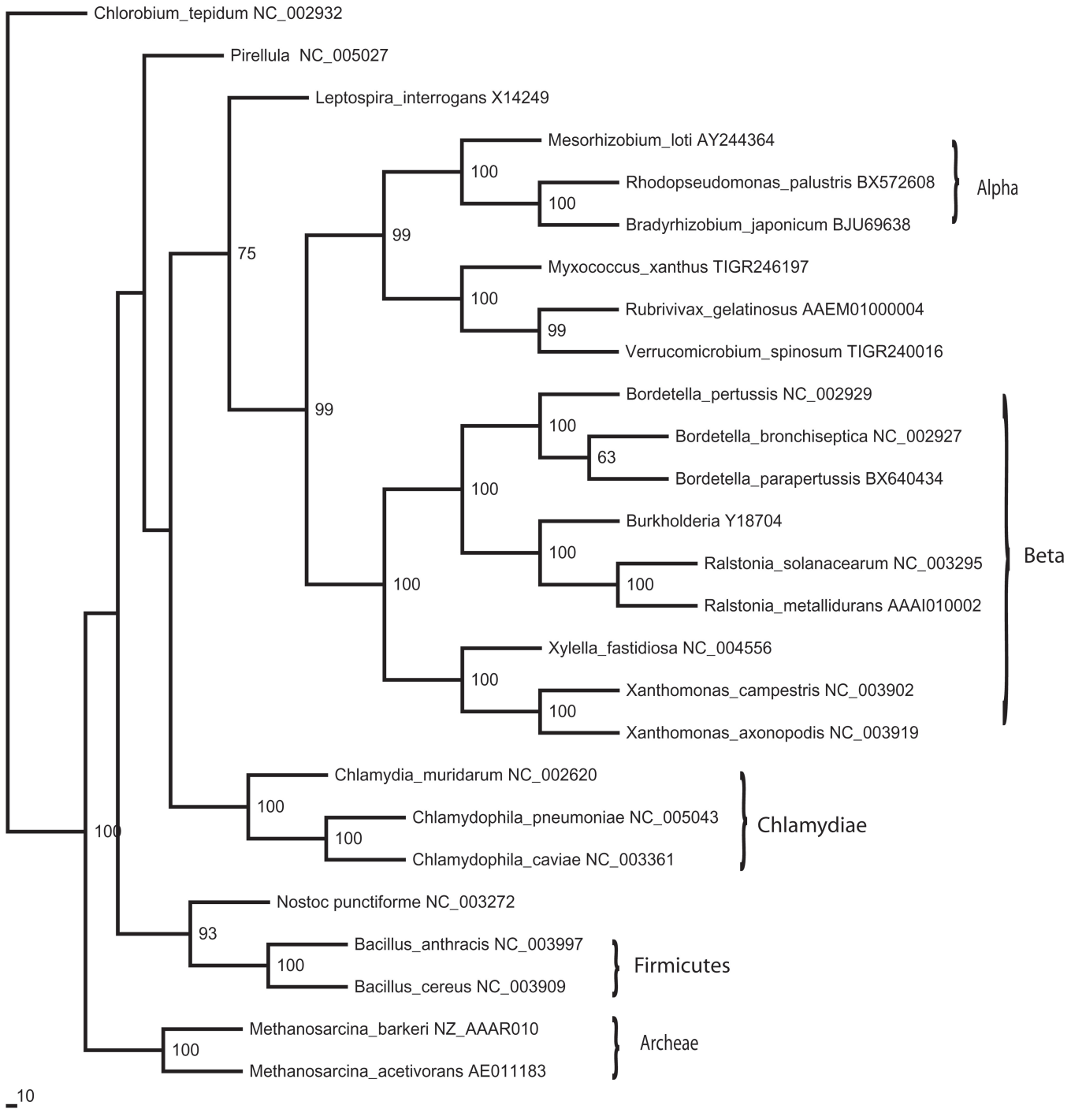
```
NP_772427   --KPYRVGRS KTGLGLFATK PIKKGTRIIR YFGPILDSRI ----------
NP_220256   --VAICWINS YVGYGVFARE SIPAWSYIGE YTGILRRRQA ----------
NP_296494   --VAICWIDA LIGYGVFARE FIPAWSYIGE YTGILRRRQA ----------
NP_225073   --VSVCWINA HVGYGVFARD EIAPWTYIGE YTGILRHRQA ----------
NP_300935   --VSVCWINA HVGYGVFARD EIAPWTYIGE YTGILRHRQA ----------
NP_829751   --VAVCWVSS YIGYGVFARE RIPAWTYIGE YTGILRRRQA ----------
ZP_00119886 --KVAQSTVP FAGLGLFTLK KIKKGSVALL YEGEKLTVEQ ----YNKRY-
ZP_00119124 --YTKESQLP DAGKGLYTSI DIFKDEVISI FKGEVLTDKE ----AARR--
NP_714158   --EIKPSSIP GIGMGLFPKE NVNKGDTIGY YTGKILTDRT ----ANSS--
NP_634869   --LISVKDAP GKGRGVFAQR DLKKGELIET CPVIVLPPEE ----------
NP_616606   MSLISVKDAP GKGRGVFAQR NLKKGEVIET CPVIVLPPEE ----------
ZP_00077699 ---------- -----MFAQR NFKRGEVIET CPVIVLPTEE ----------
NP_653535   --KTSELSDG EFNRGVFATR DIKKGELIHA APVISYPNEQ ----------
Q815R8      --KTSELSDG EFNRGVFATR DIKKGELIHA APVISYPNEQ ----------
YP_039090   -------SDG EFNRGVFATR DIKKGELIHA APVISYPNEQ ----------
NP_661845   --SVRIGPST VAGRGAFALT PIKEGDIIER CPALEVTDKD ----------
NP_710786   ----YSEWIE DEEFG-WKVL TVEEAESLPD SEKDIFMKYG ----------
ZP_00106144 --FHVTIQET AKGRGVFATK KFAKGETVVV GIPIEEVPQR ----------
NP_867824   --DIEVKSTG RCGFGVYAAR QFQPGELVFE VTGQLINKKH ----------

          L---APTFQA LLDKHAYPRD ---YKAYDGV IEYNADNARF MNHSSRP---
          L---PHSVQE MFRTYSYQTE ---GKWF--- --YCSDNSKF MNHSDDP---
          A---PQPLKE LLDRYAYPSP ---DRPG--F MVYEVDNGRF MNHSATP---
          E---SGPVKN YLDRYSYPDR ---RDPA--Y IVFEADDARY MNHDDEP---
          ----PHDPDD PNHTFYFALE ----EGGVID GKINGNSARW INHSCAP---
          ----PHDPDD PNHTFYFALE ----EGGVID GKINGNSARW INHSCAP---
          ----PHDPSE PNHTFYFALD ----EGGVID GKIDGNSARW INHSCAP---
          ----PHDPSE PNHTFYFALD ----EGGVID GKIDGNSARW INHSCAP---
          ----PHDPND PNHTFYFALE ----DGGVID GKVNGNSARW INHSCTP---
          ----PHNPAE PNHTFYFALD ----SGKVID GKVNGNSARW INHSCAP---
          ----PHDPND PNHTFYFSLD ----DGDVID AKFGGNRARW INHACDP---
          ----PHDPND PNHTFYFSLD ----DGSVID AKFGGNRARW INHACTP---
          ----PHDPDD PNHTFYFSLE ----DGSVID AKYGGNRARW INHACKP---
          ----PHDPTD PHHTFYFSID ----EDHVID AKVGGNAARW INHACAP---
          ----PHDPKD PNHTFYFHID ----EKHVID AKYGGNSSRW INHSCKP---
          ----PTNPDD PFHTFFFSLS ----SGRVID GGDEGNDARW INHSCDP---
          ----PTNPDD PFHTFFFSLS ----SGRVID GGDEGNDARW INHSCDP---
          ----PTNPDD PFHTFFFSLS ----SGRVID GGDEGNDARW INHSCDP---
          ----ERDGVE G-HTFLFGLS ----DGRVID GGQEGNSARW VNHACAA---
          ----SHGHIE SGHTFLFTLN ----DDYVID ANDAGNIARW INHSCSP---
          ----SHGHIE SGHTFLFTLN ----DDYVID ANDAGNIARW INHSCSP---
          ----DTGDVE SGHTFLFTLS ----DDYVLD ANYEGNVARW INHSCDP---
          ----DTGDVE SGHTFLFTLS ----DDYVLD ANYEGNIARW INHSCNP---
          ----DDESMA RHHTFLFNLD ----KRTVLD AGTIHNEARY INHSCAP---
          ----NYDDDG ITYLFDLDMF D-DASEYTVD AQNYGDVSRF FNHSCSP---
          -------DD- -SYLFDLDN- K-DGEVYCID ARYYGNISRF INHLCDP---
          -------DS- -TFLFETKV- --GSETLTID AKYSGNYTRF INHSCAP---
          ----TYETHG VKDAYIISL- ---NASEAID ATKKGSLARF INHSCRP---
          ----EYAFKG QKHFYFMTL- ---NGNEVID AGAKGNLGRF INHSCEP---
          ----KMKHRG ETNFYLCEI- ---TRDMVID ATHKGNKSRY INHSCNP---
          ----KLNHKV ETNFYLCQI- ---NWNMVID ATHKGNKSRY INHSCSP---
          ----DMKHKG MKDFYMCEI- ---QKDFTID ATFKGNASRF LNHSCNP---
          ----GFGSQG TSSYLFTLLE ---HEGIWVD AAMYGNLSRY INHASENDKK
          ----VFDEES NVSYLFTLLE ---DDGIWVD AAVYGNLSRY MNHASESDRN
```

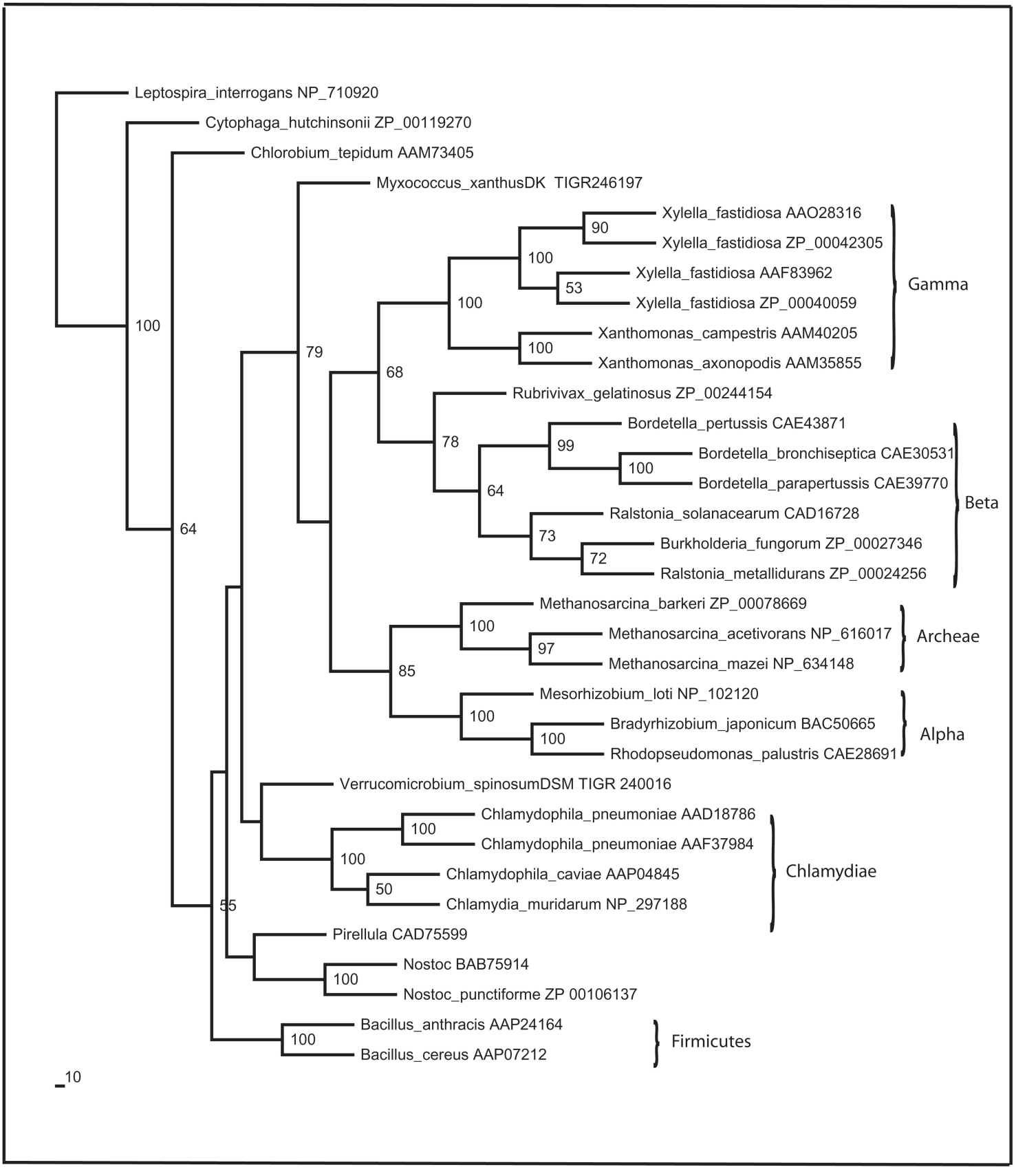
```
----VYDKYM -SSFLFNLN- ----NDFVVD ATRKGNKIRF ANHSVNP---
----IYDREN -CSFLFNLN- ----DQFVLD AYRKGDKLKF ANHSPEP---
----IEDRIG -SSYLFTLN- ----DQLEID ARRKGNEFKF LNHSARP---
ALLDEAKESG GARVYMFIL- ---NDEWDID GNVSWNTARL MNHSCEP---
--------RG QARIMIVEV- ---SERRAID ASQSADPLRF TNHSCAP---
----YYDSKG IG-CYMFRI- ---DDSEVVD ATMHGNRARF INHSCEP---
----YYDSKG IG-CYMFRI- ---DDSEVVD ATMHG-AARF INHSCEP---
LI--YNSMVG AG-TYMFRI- ---DDERVID ATRTGSIAHL INHSCVP---
----QYEKMG IGSSYLFRL- ---DDGYVLD ATKRGGIARF INHSCEP---
----NYVREG IGDSYLFRI- ---DEDVIVD ATKKGNIARF INHSCAP---
----NYRSQG KD-CY----- ---------- ---------- ----CMP---
----RYRRVG KD-CYLFKI- ---SEEVVVD ATDKGNIARL INHSCTP---
----RYNAQN RG-VYMFRL- ---DEEWVID ATMSGGPARY VNHSCDP---
-----KKHDG IENKYLFELN ----KRWTID GSVRKNVARY INHACKP---
-----PAHDE IENKYLFELN ----GRWTID GSVRKNLARY INHSCRP---
-----LWLDE NDYCFRYPVP RYSFRYFTID SGMQGNVTRF INHSDNP---
-----LWLDE NDYCFRYPVP RHSFRYFTID SGKLGNITRF INHSDNP---
-----IWMDE NDYCFRYPMP LFTLRYFTID SGKQGNVTRF INHSEQP---
-----IWMDE NDYCFRYPMP LFTLRYFTID SGKQGNVTRF INHSEQP---
-----IWMDE NDYCFRYPLS SWLWRYFTID SGRQGNFTRF INHSDKP---
-------DKE GHGEYGMTLG ---KKHVIDA RKTSSGLGRF VCDFTGSDKK
-------AKA HEDGYFINML ---DGSILDS -KNVFCFARY ANDSQGLKKT
--------KY CESKYLLWIC ---KDHWIYG EGKESNYTRF MNHSSKPN--
-----VNTLE LTRLYNYYFA W---GADSTA AAIALGYGSL YNHSYTP---
-----VNTLE LTQLYNYYFA W---GTDSRA AAIALGYGSL YNHSYTP---
-----IDSLE LTQLYNYYCA W---DSNSKD AAIALGCGSL YNHSYNP---
-----HEHIE KTLLADYAFE Y---GIN--H TAILLGYGML FNHSYTP---
-----HEHIE KTLLADYAFE Y---GIN--H TAILLGYGML FNHSYTP---
-----HEHIE KTLLADYAFE Y---GIN--H TAILLGYGML FNHSYTP---
-----IG--- -GELLNYVFY G---SAED-R RLIAMGYGMM FNHSSNP---
---------- YDVDFGLVTG -------PTS DQYVINHSNF MNHSCDP---
---------- --TIYSFQMD -------FNL YVNLDEPAVV INHSCDP---
---------Y EGSEYVMDLD ----EDWYLE PSTPG---AF MNHSCSP---

--NTYQD--- ---------D RRVFTARD-- ---VQPGEEL TCNYLFFDPR
--NTKEDFT- --------SD QTNPIGQDSA TRLILKGEEL TCNYKLFDDN
--NTDFS--- --------QY GGATATRD-- ---IAAGEEI TCDYGEFFED
--NCDVS--- --------SP EETYALRD-- ---IAPGEEL TCNYNHFFEA
--NCEAEE-- VGG------R VYIHALRD-- ---IDEQEEL FYDYGLVIDA
--NCEAEE-- VGG------R VYIHALRD-- ---IDEQEEL FYDYGLVIDA
--NCEAEE-- VKG------R VYIHALRD-- ---IEPEEEL FYDYGLVIDA
--NCEAEE-- VKG------R VYIHALRD-- ---IGAEEEL FYDYGLVIDA
--NCEAEE-- IKG------R VFVHALRD-- ---IEPEEEL FYDYGLVIDE
--NCEAEE-- IDG------H VYVHALRD-- ---IAEGEEV FYDYGLVIDA
--NCEARE-- KKG------R VFIHALRD-- ---IEPGEEL FYDYGLVIDA
--NCEARE-- KKG------R VFIHALRD-- ---IATGEEL FYDYGLVIDA
--NCEARE-- KDG------R VFIHALRD-- ---IDAGEEL FYDYGLVIEG
--NCEADE-- TDG------R VFIKTLRA-- ---VKAGEEL FYDYGLVIDE
--NCEADE-- DEG------R VFIKALRN-- ---IKAGEEL FYDYGLIIDA
--NCEAQEGR HGK------R VYIVALRD-- ---IARGEEL FYDYGLVLDG
--NCEAQEGR HGK------R VYIVALRD-- ---IARGEEL FYDYGLVLDG
--NCEAQEDR HGK------R VYIVALRD-- ---IARGEEL FYDYGLVLDG
--NCEAIE-- ESG------R VYFHATKD-- ---LEPGMEL LIDYALELDV
--NCEAVVEE DTGGNRRKDK IFIQAIRD-- ---IASGEEL TYNYGIVLAE
```

```
--NCEAVVEE DTGGNRRKDK IFIQAIRD-- ---IASGEEL TYNYGIVLAE
--NCEAVIEE AEGDDRRKDK IFIEAKRD-- ---IKPGEEL TYNYGITLAE
--NCEAVIEE AEGDDRSKDK VFIEAKRA-- ---IKPGQEL TYNYGITLGE
--NCEALIEK G--------H IYIYALTS-- ---IEPGEEL VYDYGYERTE
--NIAIYSAV RNHGFRTIYD LAFFAIKD-- ---IQPLEEL TFDYAGAKDF
--NIIPVRVF MLHQDLRFPR IAFFSSRD-- ---IRTGEEL GFDYG-DRFW
--NVKVANIS WDYDKIQLIH MCFFTDKA-- ---IRKGEEL TIDYG-EAWW
--NCETRKWN VLGEVR---- VGIFAKES-- ---ISPRTEL AYDYNFEWYG
--NCRTEKWM VNGEIC---- VGIFSMQD-- ---LKKGQEL TFDYNYVRVF
--NTQMQKWI IDGETR---- IGIFATRG-- ---IKKGEHL TYDYQFVQFG
--NTEMQKWI IDGETR---- IGIFATRF-- ---INKGEQL TYDYQFVQFG
--NCVLEKWQ VEGETR---- VGVFAARQ-- ---IEAGEPL TYDYRFVQFG
ACNITPKIIY VNNEYR---- IKFTALRD-- ---IKAGEEL FFNYGDNFPN
SCNVVPKIVQ VNGDFR---- IRFTALRD-- ---IKAGEEL FFNYGENFPN
--NCYAKVVM VNGDHR---- IGIFAKRA-- ---IQAGEEL FFDYRYSQAD
--NCYAKVIM VAGDHR---- VGIFAKER-- ---ILAGEEL FYDYRY-EPD
--NCYAKLMI VRGDQR---- IGLFAERA-- ---IEEGEEL FFDYCY-GPE
--NVEAQTWD -------EQE IWFVALRD-- ---IKKGEEL TFNYGFDLEC
--NASLRIR- -------QGR VEFYAMRD-- ---IAVGEEL CVDYGESHHE
--NCYSRVIN IDG----QKH IVIFAMRK-- ---IYRGEEL TYDYKFP-IE
--NCYSRVIN IDG----QKH IVIFAMRK-- ---IYRGEEL TYDYKFP-IE
--NCYSRVIT VNG----DEH IIIFAKRH-- ---IPKWEEL TYDYRFFSIG
--NCYTKIIS VEG----KKK IFIYAKRH-- ---IDAGEEI SYNYKFP-LE
--NCIARIIR VEG----KRK IVIYADRD-- ---IMHGEEL TYDYKFP-EE
--NCYARIVS MGDGE--DNR IVLIAKTN-- ---VAAGEEL TYDYLFEVDE
--NCYARIMS VGD-E--ESR IVLIAKAN-- ---VAVGEEL TYDYLFDPDE
--NCSTMLFD SNSGAR-DKK ILITANRP-- ---ISANEEL TYDYQFELED
--NAESDVNP RK------KR VIIRAIKN-- ---IEPGEEI NYDYGTDYFK
--NAESDVRP RE------RK VFIRAIKN-- ---IEPGDEI NYDYGTDYFK
--NLEAIGAF ENG----IFH IIIRAIKD-- ---ILPGEEL CYHYGPLYWK
--NLEAVGAF ENG----IFH IIIRAIKD-- ---IFPGEEL CYHYGPLYWK
--NAEAIGVF SEG----LFH VIIRTVAP-- ---IYAGQEI CYHYGPLYWK
--NAEAIGVF SEG----LFH VIIRTIAP-- ---IYAGQEI CYHYGPLYWK
--NVEAIGVF QNG----LFH VIIRTIQA-- ---IEAGEEL SYHYGPLYWK
AN----VEYL DNE-----GV IEIVAKKK-- ---IKPGEEL LVDYGDEMRT
SFSYNAEIHL DDE-----ER VCLVALKK-- ---IKSGEEI FCSYGKKYWQ
----VRLVVS VRW-----KT ARFEAIRK-- ---IKSGEEL FFDYGDEYWI
--NAKYQKDF NNG------L LKYVCIKD-- ---IKKDEEI TINYNCDPED
--NAEYQKDF ING------L LKYVCIKD-- ---IRKDEEI TINYNCDPED
--NARYCKDF ENS------L LKYVCIRD-- ---IQEDEEI TINYNCDPKT
--NATYDIVF ENH------T FNFYAYKD-- ---IKAGEEI LINYNGEVDN
--NATYDIVF ENH------T FNFYAYKD-- ---IKAGEEI LINYNGEVDN
--NATYDIVF ENH------T FNFYAYKD-- ---IKAGEEI LINYNGEVDN
--NVAYYRED TPTGP----E LIIYALRN-- ---IAEGEEM YYNYGDDWWK
--NMWYDQ-- ---------D DNIVAKRD-- ---IRAGEEL TIDYANFIVN
--NTGVSNNQ FGG------- YDFVALGD-- ---IEVGEEI TWDYETTEYE
--NCELVQLT EFS------- LGVVAICN-- ---IEAETEI SFDYAWEAFD
```


YP 110021 Bpm
NC 006348 Bm
ZP 00981224 Bc
YP 367500 Bc
YP 771959 Bc
ZP 00029741 Bf
YP 297593 Re
ZP 00026327 Rm
NP 521477 Rs
ZP 00243950 Rg
YP 551493 Polaromonas
NP 886510 Bpp
NP 891504 Bb
NP 879327 Bp
ZP 00425417 Bv
ZP 00040102 Xf
NP 298767 Xf
NP 643242 Xa
NP 638112 Xc
YP 630598 Mx
NP 225073 Chl.p
NP 300935 Chl.p
NP 829751 Chl.c
NP 296494 Chl.m
NP 220256 Chl.t
ZP 00008689 Rp
NP 772427 Bj
Dicty1 DDB
NP 199055 ATXR7 At
CAB71104 ATX3 At
CAA09454 MLL Gg
A44265 HTX Hs
AF329273 ATX1 At
CAE71342 Cb
AAC23419 ASHH3 At
CAB75815 ASHR4 At
CAA18207 ASHR3 At
AAF04434 ASHH1 At
AAC34358 ASHH2 At
CAA71599 CLF At
Q92800 EZH1 Hs
AAC39446 MEDEA At
NP 597099 Ce
EAA48494 Mg
EAA35807 Nc
Q96KQ7 G9a Hs
Dicty3 DDB
060016 CLR4 Sp
NP 494334 Ce
TIGR 240016 Vs
ZP 00241588 Rg
ZP 00119886 Ch
ZP 00119124 Ch
NP714158 Li
NP 867824 Pirellula
NP 616606 Ma
NP 634869 Mm
ZP 00077699 Mb
NP 653535 Ba
Q815R8 Bc
YP 039090 Bt
NP 661845 Ct
ZP 00106144 Nostoc
NP 710786 Li
NP 104295 Rl
NP 107917 Ml
NP 710330 Li
NP 768515 Bj


ZP 00981224 Bc
YP 367500 Bc
NC 006348 Bm
YP 110021 Bpm
YP 771959 Bc
ZP 00029741 Bf
YP 297593 Re
ZP 00026327 Rm
NP 521477 Rs
ZP 00243950 Rg
YP 551493 Polaromonas
NP 886510 Bpp
NP 891504 Bb
NP 879327 Bp
ZP 00425417 Bv
YP 630598 Mx
NP 298767 Xf
ZP 00040102 Xf
NP 643242 Xa
NP 638112 Xc
TIGR 240016 Vs
ZP 00241588 Rg
NP 225073 Chl.p
NP 300935 Chl.p
NP 829751 Chl.c
NP 296494 Chl.m
NP 220256 Chl.t
NP 772427 Bj
ZP 00008689 Rp
ZP 00119886 Ch
ZP 00119124 Ch
NP714158 Li
NP 867824 Pirellula
ZP 00106144 Nostoc
Q815R8 Bc
YP 039090 Bt
NP 653535 Ba
NP 616606 Ma
NP 634869 Mm
ZP 00077699 Mb
NP 661845 Ct
NP 710786 Li
NP 107917 Ml
NP 104295 Rl
NP 768515 Bj
NP 710330 Li
Domain 1
Domain 2


Chlorobium_tepidum NC_002932
Pirellula NC_005027
Leptospira_interrogans X14249
Mesorhizobium_loti AY244364
Rhodopseudomonas_palustris BX572608
Bradyrhizobium_japonicum BJU69638
Alpha
Myxococcus_xanthus TIGR246197
Rubrivivax_gelatinosus AAEM01000004
Verrucomicrobium_spinosum TIGR240016
Bordetella_pertussis NC_002929
Bordetella_bronchiseptica NC_002927
Bordetella_parapertussis BX640434
Burkholderia Y18704
Ralstonia_solanacearum NC_003295
Ralstonia_metallidurans AAAI010002
Beta
Xylella_fastidiosa NC_004556
Xanthomonas_campestris NC_003902
Xanthomonas_axonopodis NC_003919
Chlamydia_muridarum NC_002620
Chlamydophila_pneumoniae NC_005043
Chlamydophila_caviae NC_003361
Chlamydiae
Nostoc punctiforme NC_003272
Bacillus_anthracis NC_003997
Bacillus_cereus NC_003909
Firmicutes
Methanosarcina_barkeri NZ_AAAR010
Methanosarcina_acetivorans AE011183
Archeae
100
75
99
99
100
100
100
100
100
63
100
100
100
100
100
100
100
93
100
100
_10


Leptospira_interrogans NP_710920
Cytophaga_hutchinsonii ZP_00119270
Chlorobium_tepidum AAM73405
Myxococcus_xanthusDK TIGR246197
Xylella_fastidiosa AAO28316
Xylella_fastidiosa ZP_00042305
Xylella_fastidiosa AAF83962
Xylella_fastidiosa ZP_00040059
Xanthomonas_campestris AAM40205
Xanthomonas_axonopodis AAM35855
Gamma
Rubrivivax_gelatinosus ZP_00244154
Bordetella_pertussis CAE43871
Bordetella_bronchiseptica CAE30531
Bordetella_parapertussis CAE39770
Ralstonia_solanacearum CAD16728
Burkholderia_fungorum ZP_00027346
Ralstonia_metallidurans ZP_00024256
Beta
Methanosarcina_barkeri ZP_00078669
Methanosarcina_acetivorans NP_616017
Methanosarcina_mazei NP_634148
Archeae
Mesorhizobium_loti NP_102120
Bradyrhizobium_japonicum BAC50665
Rhodopseudomonas_palustris CAE28691
Alpha
Verrucomicrobium_spinosumDSM TIGR 240016
Chlamydophila_pneumoniae AAD18786
Chlamydophila_pneumoniae AAF37984
Chlamydophila_caviae AAP04845
Chlamydia_muridarum NP_297188
Chlamydiae
Pirellula CAD75599
Nostoc BAB75914
Nostoc_punctiforme ZP 00106137
Bacillus_anthracis AAP24164
Bacillus_cereus AAP07212
Firmicutes
100
64
79
68
100
90
100
53
100
100
78
99
100
64
73
72
100
97
85
100
100
100
100
50
55
100
100
10

LEGEND TO SUPPLEMENTARY FIGURES (SF)

Figure SF1. Multiple Alignment of Eukaryotic and Bacterial SET Domain Sequences used for Tree-Reconstruction

Figure SF2. Minimal Evolution Tree of Eukaryotic-Bacterial SET domain proteins

ME tree generated with 10 global rearrangements and 10 random jumbles. Numbers show bootstrap values (10=100%). Abbreviations are as indicated in the Legend to Figure 1(main text). The three *Bacilli* and the three *Methanosarcinae* species, as well as *Chlorobium,* segregate in a well supported clade consistent with HGT among them (bracketed clade). The distribution of the eukaryotic proteins is in the shaded area.

Figure SF3. Minimal Evolution Tree of Bacterial SET domain proteins

ME tree generated with 10 global rearrangements and 10 random jumbles with bootstrap values. Abbreviations are as indicated in the Legend to Figure 1 (main text). The three *Bacilli* and the three *Methanosarcinae* species, as well as *Chlorobium,* segregate in a well supported clade consistent with HGT among them. The distribution of the species in Domains 1 and 2 is similar with the pattern of the MP tree in Figure 2 (main text), except the unsupported relocation of *Pirellula* and *Nostoc* in Domain2.

Figure SF4. Distribution of bacterial rRNA23S sequences

Bootstrap analysis with the Seqboot program (500 pseudo-replicates) was used. Figures indicate bootstrap values (100=100%). Unrooted Majority Rule Consensus Tree performed with the Consense program and plotted using the TREEVIEW program.

Figure SF5. Distribution of bacterial 50S ribosomal protein L3.

Bootstrap analysis with the Seqboot program (500 pseudo-replicates) was used. Figures indicate bootstrap values (100=100%). Unrooted Majority Rule Consensus Tree performed with the Consense program and plotted using the TREEVIEW program.