RNA Biology 10:4, 636–651; April 2013; © 2013 Landes Bioscience

Archaeal and eukaryotic homologs of Hfq
A structural and evolutionary perspective on Sm function

Cameron Mura,[1,*] Peter S. Randolph,[1] Jennifer Patterson[1] and Aaron E. Cozen[2]

[1]Department of Chemistry; University of Virginia; Charlottesville, VA USA; [2]Department of Biomolecular Engineering; University of California; Santa Cruz, CA USA

Keywords: Sm, Lsm, Hfq protein, Sm fold, oligomers, archaeal RNA, RNA chaperone, RNP assembly, RNP evolution

Abbreviations: 3D, three-dimensional; EM, electron microscopy; HMM, hidden Markov model; Lsm, Sm-like; ncRNA, noncoding RNA; RNase P, ribonuclease P; RNP, ribonucleaoprotein; SmAP, Sm-like archaeal protein; SMN, survival of motor neurons protein; snRNP, small nuclear RNP; sno-, small nucleolar; ssRNA, single-stranded RNA; sRNA, small (noncoding) RNA; SSE, secondary structural element (α-helix, β-strand, etc.)

Hfq and other Sm proteins are central in RNA metabolism, forming an evolutionarily conserved family that plays key roles in RNA processing in organisms ranging from archaea to bacteria to human. Sm-based cellular pathways vary in scope from eukaryotic mRNA splicing to bacterial quorum sensing, with at least one step in each of these pathways being mediated by an RNA-associated molecular assembly built upon Sm proteins. Though the first structures of Sm assemblies were from archaeal systems, the functions of Sm-like archaeal proteins (SmAPs) remain murky. Our ignorance about SmAP biology, particularly vis-à-vis the eukaryotic and bacterial Sm homologs, can be partly reduced by leveraging the homology between these lineages to make phylogenetic inferences about Sm functions in archaea. Nevertheless, whether SmAPs are more eukaryotic (RNP *scaffold*) or bacterial (RNA *chaperone*) in character remains unclear. Thus, the archaeal domain of life is a missing link, and an opportunity, in Sm-based RNA biology.

Introduction: The Sm Family, its Biology and an Archaeal Lineage

A history of the Sm/Lsm-SmAP-Hfq family. Human Sm proteins were discovered over 30 y ago[1] as a group of small antigens involved in the autoimmune disease systemic lupus erythematosus.[2,3] The ≈80-residue proteins were identified in association with ribonucleoprotein (RNP) complexes from eukaryotic cellular extracts.[4] Other early work uncovered vital roles for Sm proteins in forming the cores of the uracil-rich small nuclear RNPs (U snRNPs) that further assemble into spliceosomes and excise introns in eukaryotic pre-mRNAs (reviewed in ref. 5). Over the ensuing decades, great strides in elucidating the physiological and biochemical properties of Sm proteins, as well as the three-dimensional (3D) structures and assembly behavior of these RNA-associated proteins, led to our current view that eukaryotic Sm proteins function as molecular *scaffolds*[6] for RNP assembly. As depicted in **Figure 1A**, eukaryotic Sm assemblies act in a vast array of RNA-related pathways; for recent reviews of this work, see for instance references 7–10. Paralleling this work on the canonical eukaryotic Sm proteins of the spliceosomal snRNPs, early biochemical and bioinformatic analyses,[11,12] along with biophysical and crystallographic studies,[13-15] expanded our view of the phylogenetic distribution of the Sm family to include an Sm-like (Lsm) subfamily, and revealed Sm proteins in the archaeal domain[16] of life (Sm systems resembling those of eukaryotes were not necessarily expected in the archaea, given the lack of introns in their protein-coding genes and their presumably more primitive RNA-processing machineries[17]). Finally, in a third line of seemingly unrelated discoveries—in bacteria, dating to the late 1960s—an *Escherichia coli* "host factor I" (HF-I) protein was found to be necessary for replication of the bacteriophage Qβ.[18] Biochemical characterization of this host factor for bacteriophage Qβ replication, dubbed "Hfq," revealed that the protein (1) forms thermostable hexamers,[19] (2) occurs at high intracellular concentrations[20,21] and (3) preferentially binds A/U-rich single-stranded RNA (ssRNA) via multiple sites on the protein.[22-24] This Hfq was also capable of interacting with DNA,[21,25,26] such as in the *E. coli* nucleoid.[27]

These three lines of Sm research—*eukarya*, *archaea* and *bacteria*—were unified by the realization ca. 2002 that Hfq is the bacterial branch of the Sm family. The Hfq↔Sm homology was first suggested by weak sequence similarities between the N-terminal regions of the ≈80–120 residue Hfq and Sm proteins, was further corroborated by phylogenetic, biophysical, fold recognition and homology modeling studies of *E. coli* Hfq,[28-31] and was firmly established by the first crystal structure of Hfq,[32] which revealed a hexamer composed of Hfq subunits that adopt the Sm fold. A surge of biochemical, biophysical and genetic/RNomic studies of Hfq over the past decade has revealed much about the roles of this Sm protein in bacterial RNA metabolism, as well as structure/function relationships in the Hfq branch of the Sm family. Whereas eukaryotic Sm proteins serve more "passive" functions as structural scaffolds, Hfq acts as an RNA *chaperone*,[33-35] mediating antisense interactions between small regulatory, noncoding RNAs (ncRNA)[a] and their targets (**Fig. 1B**) and directly influencing the structures of some RNAs. Relatively recent reviews are available on Hfq-based RNA biology from microbiological and

*Correspondence to: Cameron Mura; Email: cmura@virginia.edu
Submitted: 03/18/13; Revised: 03/29/13; Accepted: 04/03/13
http://dx.doi.org/10.4161/rna.24538

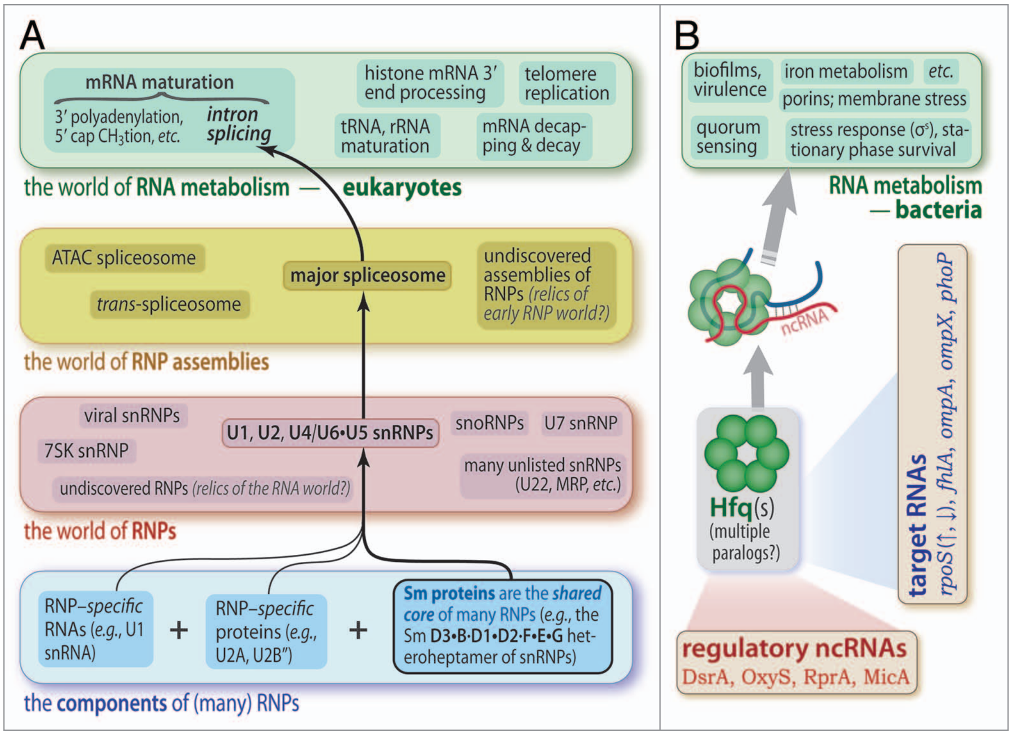
A
mRNA maturation
3′ polyadenylation, 5′ cap CH₃tion, etc.
intron splicing
histone mRNA 3′ end processing
telomere replication
tRNA, rRNA maturation
mRNA decapping & decay
the world of RNA metabolism — eukaryotes
ATAC spliceosome
trans-spliceosome
major spliceosome
undiscovered assemblies of RNPs (relics of early RNP world?)
the world of RNP assemblies
viral snRNPs
7SK snRNP
U1, U2, U4/U6•U5 snRNPs
snoRNPs
U7 snRNP
many unlisted snRNPs (U22, MRP, etc.)
undiscovered RNPs (relics of the RNA world?)
the world of RNPs
RNP–specific RNAs (e.g., U1 snRNA)
+
RNP–specific proteins (e.g., U2A, U2B″)
+
Sm proteins are the shared core of many RNPs (e.g., the Sm D3•B•D1•D2•F•E•G heteroheptamer of snRNPs)
the components of (many) RNPs
B
biofilms, virulence
iron metabolism
etc.
porins; membrane stress
quorum sensing
stress response (σˢ), stationary phase survival
RNA metabolism — bacteria
ncRNA
Hfq(s) (multiple paralogs?)
target RNAs
rpoS(↑, ↓), fhlA, ompA, ompX, phoP
regulatory ncRNAs
DsrA, OxyS, RprA, MicA

Figure 1. A bottom-up approach to Sm function in RNA metabolism. Placing the Sm protein family in a biochemical context underscores its central role in myriad RNA processing pathways in the eukaryotic (**A**) and bacterial (**B**) domains of life, highlighting the gaps in our knowledge for the archaea. The diagram indicates how RNA processing events (top layer) hierarchically build upon Sm proteins (bottom layer). One of the most extensively characterized Sm-based pathways is the excision of introns from pre-mRNA, which can be dissected (**A**) as *intron splicing←spliceosome←U1, U2, U4/U6* and *U5 snRNPs←Sm core of snRNPs*. While this eukaryotic example demonstrates a functional niche of Sm proteins as *scaffolds*, Hfq acts instead as a *chaperone* (**B**), mediating interactions between regulatory ncRNAs (red) and their targets (blue). This schematic is not comprehensive (for clarity, not all known connections are shown) and new examples of Sm function are being discovered continuously, particularly in the bacterial context of Hfq; the pace of discoveries of new Sm functions will likely increase as new interactions and functional linkages are uncovered by genome- and proteome-wide studies.

structural perspectives,[36-39] including the other reviews in this Special Focus issue.[40-42]

The in vivo functions of archaeal Sm proteins remain unknown, in contrast to the eukaryotic and bacterial homologs, and despite the fact that the first atomic-resolution structures of intact Sm rings were from archaeal systems. SmAP function can be approached by using the homology between SmAP↔Hfq and SmAP↔Sm/Lsm subfamilies to make phylogenetic inferences about likely Sm functions in the archaea. Thus, the remainder of this introductory section summarizes eukaryotic (Sm/Lsm) and bacterial (Hfq) biology (**Fig. 1**), as well as the evidence for authentic Sm proteins in the archaea. The next section reviews SmAP 3D structures at the levels of monomers and oligomers, and as regards modularity of the Sm fold; Sm sequence/structure relationships are also described, and some nomenclature issues are raised from a bioinformatic perspective. In all of this, a major question is whether the cellular roles of SmAPs are more eukaryotic (RNP *scaffold*) or bacterial (RNA *chaperone*). Thus, the final third of this review examines the possible biochemical roles of SmAPs, starting with what is already known about (potentially Sm-linked) RNA processing in the archaea; this final section also considers the genomic context of Sm genes and offers an exploratory discussion of what may be expected for archaeal Sm function. As suggested by the absence of an archaeal panel in **Figure 1**, the main motivation for this review is that SmAPs represent a significant opportunity in Sm-based RNA biology.

A synopsis of eukaryotic and bacterial Sm biology. *Eukaryotic Sm proteins serve as RNP scaffolds.* A modular approach to eukaryotic Sm-based RNA biology is shown in **Figure 1A**. Various forms of RNA processing occupy the top level of this hierarchy, including rRNA processing by small nucleolar RNPs (snoRNPs),[43] RNase P-based splicing and maturation of tRNA,[44] processing of the 3' ends of histone mRNA by U7 snRNP,[45,46] mRNA decapping and decay[47] and chromosome end maintenance by telomerase.[48] Each of these pathways employ Sm or Lsm proteins. Indeed, a central theme of **Figure 1** is that a great diversity of RNA processing events (on a cellular scale) can be traced back to the Sm proteins (on a molecular scale). Because Sm proteins were first identified in connection with RNA splicing, the most thorough biochemical and structural picture available for the molecular basis of Sm function concerns their roles in snRNP-mediated intron excision; snRNPs also provide a useful starting point in considering the potential cellular niches of Sm protein in the archaea.

To simplify our understanding of the architectural role of Sm proteins, each U snRNP can be viewed as an RNP composed of two parts: the respective U snRNAs (U1, U2, etc.) and up to dozens of proteins. The U snRNPs are dissected in the two bottom layers of **Figure 1A**. The protein components fall into two classes: snRNP-specific proteins, such as U2A' and U2B'' of the U2 snRNP, and core proteins that are common to each snRNP.[10,49] The snRNP-specific proteins mediate specific RNA···RNA, protein···RNA, and protein···protein interactions and function in ways unique to each snRNP (e.g., DEAD-/DxxH-box helicases). In contrast, the molecular functions of the shared core snRNP proteins—the Sm/Lsm proteins—are presumably more generic.

The *scaffolding* functionality of eukaryotic Sm proteins is exemplified by their roles in snRNP biogenesis. Sm proteins nucleate the early stages of snRNP assembly by binding single-stranded regions of snRNA. The consensus Sm-binding site is a short uracil-rich sequence $PuAU_{\approx 4-6}GPu$ (Pu = purine)[50] flanked by RNA stem-loops. However, the Lsm ring binds at the single-stranded 3' end of U6 snRNA, thus demonstrating the variation that is possible in local RNA 2° structures for different Sm- or Lsm-binding sites. Consistent with a shared ancestry, both eukaryotic and bacterial Sm proteins appear to bind U-rich RNAs, such as the snRNA Sm motif, in the central pore toward the same face of the ring (corresponding to the proximal face of Hfq). Also of interest as regards to SmAP function and oligomeric plasticity (detailed below), eukaryotic Sm proteins form stable sub-complexes, such as Sm D1•D2 and F•E•G heteromers, that can then associate into a pentameric "subcore;" notably, these assembly intermediates are of functional relevance.[51] The Sm-templated assembly of snRNPs is guided by interactions between specific Sm proteins and the survival of motor neurons (SMN) protein complex. These and other biochemical features of Sm function have been reviewed in great detail[7,52] and an atomic-resolution picture has begun to emerge via recent structural work.

Structural enlightenment. Decades of genetic, biochemical and electron microscopic (EM)[53] studies have now culminated in three lines of structural work that substantially advance our understanding of Sm function.[9,10] First, recent structures of the U1[54,55] and U4[56] snRNPs expose Sm rings in their final assembly state, bound to snRNAs and snRNP-specific proteins.[9,10] These new structures establish that an snRNA threads through the eukaryotic Sm pore, unlike what is thought to be the case in RNA_1•Hfq•RNA_2 ternary complexes (wherein RNAs bind to distinct regions of an Hfq ring). These structures also show the Sm surface to be a versatile platform for protein···protein and protein···RNA interactions. In a second line of work, the structure of a late intermediate in the snRNP assembly pathway, an Sm D1•D2·F•E•G pentamer bound to part of the Gemin/SMN complex,[57] elucidates the mechanistic basis for SMN-chaperoned snRNA···Sm associations. Rather than thread through a preformed ring, snRNA is sequentially bound by Sm subunits via discrete, metastable intermediates.[58] Finally, a third line of work offers structures of early intermediates in snRNP assembly and unveils a fascinating case of molecular mimicry: A β-sheet-rich assembly chaperone (pICln) that resembles the overall shape of roughly two Sm subunits "wedges" into a crescent-shaped Sm pentamer[59] and stabilizes the partially assembled Sm ring. Beyond these purely structural/scaffolding roles, Sm proteins also serve as regulatory points in RNA pathways via post-translational modifications. Sm RNP biogenesis can be modulated by dimethylation of arginines in the C-term RG dipeptides of Sm and Lsm proteins (the C-term RG in **Fig. 2A** is not one of these tandem RG methylation substrates); these and other cellular roles of Sm methylation have been reviewed.[60-62] While it remains to be seen if archaeal Sm complexes match the functional intricacy found in these recent structural studies of eukaryotic Sm RNPs, the elaborate web of Sm-mediated RNA···RNA, RNA···protein and protein···protein interactions underscores the roles of Sm proteins in RNP biogenesis and function. Because Sm/Lsm homologs occur in the archaea and likely existed in the eukaryotic ancestor,[63-65] we do not dismiss the possibility that SmAPs serve similar scaffolding functions in as yet undiscovered archaeal RNPs.

Bacterial Sm proteins act as RNA chaperones. Hfq functions as an RNA chaperone—viz., a single-stranded nucleic acid-binding protein with flexible sequence recognition capacity, such that it can facilitate base-pairing interactions between diverse ncRNAs (regulatory sRNAs) and protein-coding mRNA targets. These antisense sRNA···RNA_{target} interactions, shown schematically in **Figure 1B**, often exhibit only partial base-pairing complementarity. By binding the two RNAs independently, Hfq increases their local effective concentration, thereby enhancing their binding affinity. Structural and mechanistic aspects of the "cycling"[66] of RNA on the surface of the Hfq ring are reviewed by Sauer and by Wagner[40] in this issue. Hfq-mediated RNA···RNA interactions typically have repressive physiological effects, downregulating either mRNA stability or translational activity. However, recent studies indicate that Hfq can also guide RNA···RNA interactions that exert positive regulatory effects.[67] Hfq has also been shown to modulate mRNA stability by promoting polyadenylation,[68] which is often perceived as a eukaryotic-specific function but that also occurs in bacteria and may be intricately linked to Hfq function.[42] A rapidly growing body of work has established pleiotropic roles for Hfq in physiological processes ranging from oxidative stress response and metal homeostasis to regulation of pathogenicity.[31,35,37,69-72] The discovery that Hfq mediates a fundamental regulatory step in quorum sensing[73] further expands the scope of Sm function to include microbial cell···cell communication networks and intercellular signaling, which enables the emergence of population-wide behaviors.

Compared with the substantial progress on eukaryotic and bacterial Sm proteins, little is known about Sm-related RNA biology in the archaea. There are more questions than answers and, therefore, portions of this review should be taken as more speculative and interrogative rather than conclusive.

Sm-like archaeal proteins: Suggested by sequence, confirmed by structure. Sm sequences, often described as *Sm1* and *Sm2* signature motifs joined by a variable linker (**Fig. 2A and B** and nomenclature note below), are conserved in many species across the tree of life. Stimulated by the flood of sequences at the dawn of the genomic era, early database searches[11,12] revealed that Sm proteins are not exclusive to metazoans or other higher eukaryotes with elaborate mRNA splicing; indeed, several Sm homologs have been found in eukaryotes as divergent from humans as

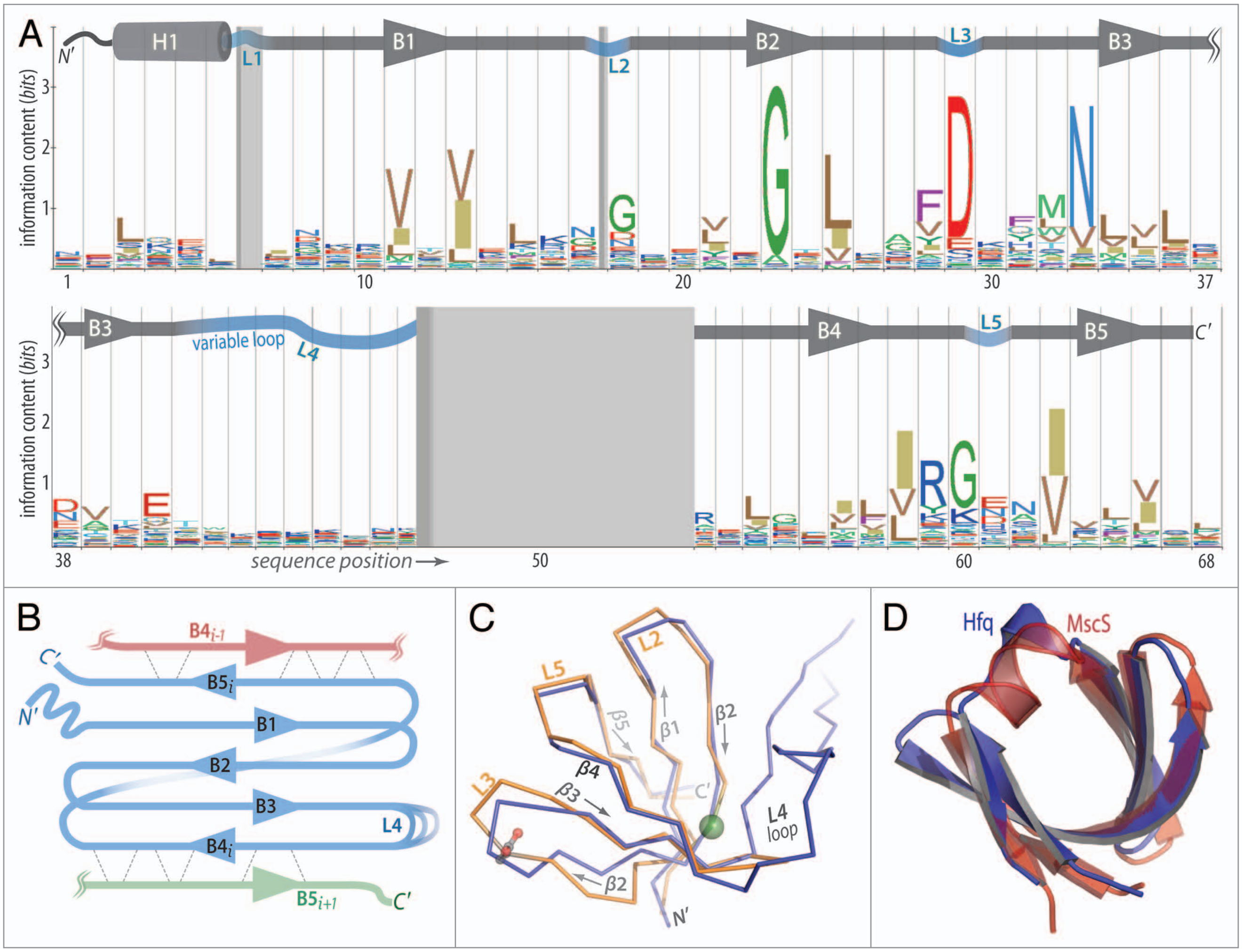

Figure 2. SmAP monomers: Sequence profiles and a 3D structure of versatile functionality. A probabilistic model of sequence variation across the Sm family is shown (**A**) as a profile hidden Markov model (pHMM). This visual display of pHMMs using logos[144] is roughly analogous to the more familiar sequence logos used in representing multiple sequence alignments. In this pHMM, the vertical axis corresponds to the *information content*, measured in bits[145] relative to the profile's background distribution; positions that contain more information correspond to higher stacks, and amino acid letter heights within a stack are scaled by that residue's relative contribution to the position. The horizontal axis can be considered as the position "*s*" along the Sm sequence profile. (Technically it is the sequential chain of HMM states, with the hitting probability of visiting state "*s*" along the HMM chain colored dark gray[144] and the contribution of a (match or insertion) state "*s*" to the overall Markov chain shown as the sum of the widths of light- + dark-gray regions). The ≈70-residue Sm core is split across two rows for clarity, and the chief SSEs are depicted near the top of each row. Note that the loop L4 variation is captured by this pHMM, as are other important sites in Sm sequences. The SSEs of the Sm fold arrange into a five-stranded antiparallel β-sheet that interacts in an antiparallel configuration with strands of flanking subunits in an oligomer (**B**). The highly bent β-sheet of the Sm fold is shown as a C_α trace in (**C**). Loops L2, L3 and L5 lie toward the lumen of the ring (L3 is nearest the proximal face); loop L4 lies toward the distal face. Important residues from the profile HMM are marked in the 3D structure (**C**): the β2 strand Gly is shown as a green sphere and the loop L3 Asp is shown in ball-and-stick (lower-left). The backbone traces in (**C**) of two Sm homologs of < 35% sequence similarity—*E. coli* Hfq (orange) and a putative cyanophage Sm (blue)—illustrate the persistence of the Sm fold at low sequence similarity. Representing even greater sequence divergence, the Sm fold of *S. aureus* Hfq (PDB 1KQ1; blue) and an Sm-like fold in the membrane channel MscS (PDB 2VV5; red) are shown superimposed in (**D**).

yeast[74] and trypanosomes,[75] and Sm proteins likely existed in the ancestor of eukaryotes.[65] Sm homologs also have been found in several archaeal species.[12,17,63] The discovery of SmAPs was not entirely expected as Sm proteins were thought to act in snRNP biogenesis and splicing, not general purpose RNA processing; an archaeal RNP complex homologous to the sophisticated eukaryotic splicing apparatus was (and remains) unknown. The discovery of SmAPs raises several implications and questions about the role of these proteins in archaeal RNA metabolism.[17] In short, what are the archaea doing with Sm proteins?

The finding that Hfq is the bacterial Sm completes our modern understanding, showing that Sm proteins occur in each domain of life and making their existence in the archaea less startling. Also fascinating from a phylogenetic perspective are the emerging links between host Sm proteins and exogenously encoded (e.g., viral) RNAs: *Herpesvirus saimiri* produces viral RNA transcripts that recruit host Sm proteins,[76] and the yeast *Brome* mosaic virus encodes two distinct RNA elements that directly interact with host Lsm1-7 rings in a manner resembling that of Hfq-RNA interactions.[77] Somewhat similarly, a novel pentameric Sm-like protein of putative cyanophage origin was recently uncovered in an ocean metagenomics sampling expedition.[78] These results suggest that the phylogenetic diversity of Sm proteins is far broader than previously thought, including virtually every known form of life, and also expand the realm of possible Sm functions well beyond splicing and other familiar forms of RNA processing.

The first putative SmAPs were detected by sequence analysis. Since then, the existence of a distinct, albeit phylogenetically disperse, Sm family has been substantiated via biophysical, biochemical, ultrastructural[79] and crystallographic studies of SmAP orthologs[13-15] and paralogs.[80-82] Such work has also uncovered several sets of Lsm paralogs in organisms already known to have homologs of the canonical Sm proteins.[11] Biochemical and structural studies of Sm homologs verify their similarity to canonical Sm proteins.[12,79] All known Sm structures, from eukaryotes, bacteria and archaea, are markedly similar to one another in terms of coordinate root-mean-squared deviation (RMSD) of monomer backbones. As shown in **Figure 2C**, the Sm fold is conserved even for highly divergent pairs lying near the "twilight zone"[83] of similarity scores for the alignment of two random sequences. (A basic premise of structural bioinformatics is that function arises from structure; function is the level at which evolutionary pressure applies and, therefore, biomolecular structure persists more strongly over deep evolutionary timescales than does sequence conservation). Thus, the existence of Sm homologs in most species across all three domains of life implies an ancient evolutionary origin for the Sm family, predating the archaeal/eukaryotic divergence.

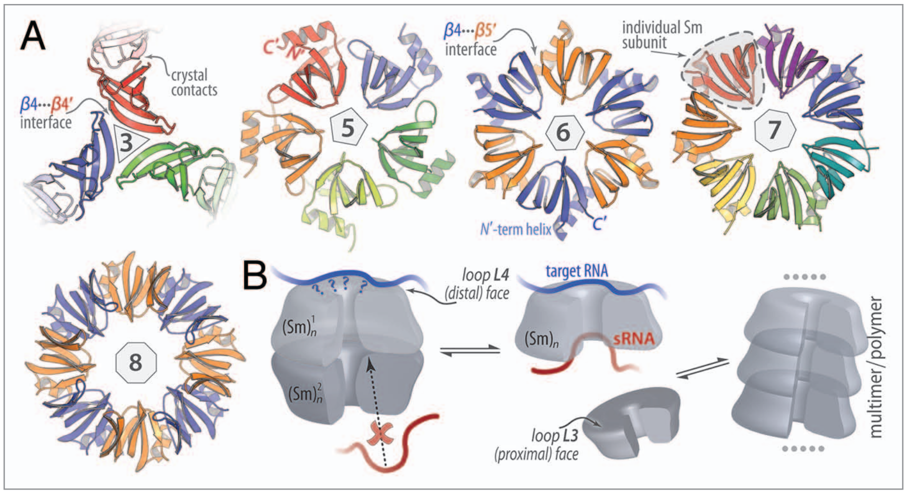

Figure 3. Oligomeric plasticity of SmAPs and other Sm assemblies. Despite substantial similarity at the level of monomers, SmAPs and their Sm homologs exhibit profound variability at the levels of single-ring (**A**), multi-ring (**B**, left), and higher-order (**B**, right) assemblies. Each subunit in these ribbon cartoons is colored individually, the *n*-fold rotational symmetry axis is indicated, and each ring is viewed onto the *L4* (*distal*) face; the N′- and C′-termini of one subunit are indicated for $n = 5$ and 6 but are not marked for each subunit so as to minimize clutter. A speculative model for the potential roles of multi-ring and higher-order assemblies is shown in (**B**).

SmAP Structure: Monomers, Assemblies, Modularity

For a protein of only ≈70 residues, Sm monomers are exceptionally multifunctional (**Fig. 2**): one part of the Sm fold mediates interatomic contacts between subunits in a ring ($\beta4_i\cdots\beta5_{i+1}$ interface), while other portions of the fold help create not one but possibly three distinct RNA-recognition regions, including (1) a U-rich ssRNA-binding region near the often cationic pore of Sm/Lsm, SmAP and Hfq rings, (2) an A-rich binding surface defined by the L4 face of Hfq and (3) a newly recognized[24] RNA-contacting region around the lateral periphery of Hfq rings. Other structural landmarks also have variation as the theme: extensive variation in loop L4 length, variation in the termini (some Sm domains are fused to other domains at the N- and/or C-termini) and variable oligomeric states. Much of our knowledge of Sm structure and assembly originated from studies of SmAPs. This section reviews the 3D structures of Sm homologs. The assembly behavior of SmAPs and related homologs is also examined, as is the possibility that the main functional/evolutionary niche of the Sm domain is a generic structural module for protein···protein and protein···RNA interactions, akin to the activity of Hfq as a generic facilitator of RNA···RNA interactions.

SmAPs and Sm monomers. The first Sm structures were of the human Sm D1•D2 and D3•B heterodimers.[84] Soon, thereafter, the crystal structures of three SmAPs were reported concurrently: a *Methanobacterium thermautotrophicum* (*Mth*) SmAP,[13] *Pyrobaculum aerophilum* (*Pae*) SmAP1[14] and an *Archaeoglobus fulgidus* (*Afu*) SmAP,[15] providing the first atomic-resolution glimpse of Sm monomers in an intact ring. All three of these SmAP orthologs assemble as homoheptamers comprised of subunits that adopt the same Sm structure found in the human D1•D2 and D3•B dimers. In the subsequent decade, dozens of Sm crystal structures have been determined for orthologs and paralogs from eukaryotic, bacterial and archaeal lineages [see refs. 10, 36, 38 and 39 and Sauer (this issue) for reviews]. All structural studies, including by solution state NMR spectroscopy of Sm[85] and Sm-like[86] domains, show that Sm monomers adopt a unique fold: a strongly bent, five-stranded antiparallel β-sheet often capped by an N-terminal α-helix. This N-terminal helix has been used as a structural marker to define the *proximal* face of Hfq rings (e.g., the Hfq hexamer in **Fig. 3A**), but the helix is an inessential feature of the Sm fold and is likely absent from many Sm sequences. Also, at least one Sm structure (the pentamer in **Fig. 3A**) features no N-term helix but rather a C-term helix that occurs on the *distal* face; thus, the presence of a particular helix (or any SSE beyond the Sm core sheet) is of limited utility as a landmark for distinguishing the faces of Hfq and other Sm rings.

As gauged by sequence analysis, the Sm core is ≈60–70 residues in length (**Fig. 2A**). The Sm β-sheet is highly curved, and the degree of curvature can be approximated as the distance between the two termini of a segment of β-strand in a given conformation (the *chord length*, l_c) vs. the corresponding distance for that segment in a fully extended conformation (the *arc length*, l_a); the l_c/l_a ratio, which is unity for a straight line, can be taken as a crude estimate of curvature. For example, the distance between the *Pae* SmAP1 β2-strand termini (C_α^{S31} and C_α^{Q43}) is 24 Å, vs. a value of 40 Å for this pair of residues in an unbent, fully extended conformation. Such curvature is a hallmark of the trough-shaped Sm fold, making Sm proteins nearly elliptical or U-shaped in cross-section (see the perspective in **Fig. 2D**). The polypeptide backbone can adopt this bent conformation because of specific glycine residues that serve as pivot points, particularly in strand β2 (**Fig. 2A** and the green sphere in **Fig. 2C**) but also in strands β3, β4 and the loops. The phylogenetically conserved glycines are among the most characteristic features of the Sm sequence family, in the information theoretic "profile" sense shown in **Figure 2A**; less strictly conserved glycines also serve structural roles, as can be found in SmAP-specific multiple sequence alignments.[14,82] In addition to the 3D conformational pliability that enables the β-sheet to bend upon itself, a hallmark of the Sm fold is its resilience to sequence variation, such that two randomly selected Sm structures typically feature backbone (C_α) RMSDs of only ≈1–2 Å (**Fig. 2C and D**).

Variation in loop L4 is another characteristic feature of Sm monomer structure. This loop links strands β3 and β4 (**Fig. 2A and B**) and varies more than other Sm loops in length and amino acid sequence—from just a few residues in bacterial homologs (Hfq) to potentially dozens of residues in eukaryotic homologs (e.g., human SmB). Within Sm rings, the geometric orientation of individual subunits positions L4 "outward," making these loops the most prominent structural feature on the L4(/distal) face of the rings. This is an important factor in considering structure/function relationships because the L4 face of Hfq is the primary region of interaction with A-rich RNAs;[23,87] amino acid variation in L4 modulates the electrostatic potential—and, therefore, the RNA-binding properties—across that face of the Sm ring (see e.g., ref. 88 for a discussion of this effect). The L4 loop can also lead one astray in purely sequence-based bioinformatics: Multiple sequence alignments of SmAPs exceeding ≈100 residues, such as *Pae* SmAP3, erroneously assign the "extra" (non-Sm) residues to two regions—some residues were flagged as L4 loop insertions while the remainder were predicted to form a C-terminal extension.[82] However, the *Pae* SmAP3 crystal structure (see below) shows that the extra ≈60 residues actually comprise a unique, autonomous C-terminal domain.

Sm sequence/structure relationships and bioinformatic nomenclature. Sm subunits have been described as consisting of "Sm1+Sm2" motifs, a view of Sm structure that dates to early sequence analyses.[89] A probabilistic model of sequence variation across the entire Sm family is shown in **Figure 2A** as a Pfam-generated[90] profile hidden Markov model (pHMM). Profile HMMs[91] can effectively capture such features of sequence variation as amino acid insertions, thus making them a potentially effective approach for quantitatively modeling Sm loop variation. The profile HMM shown in **Figure 2A** captures known features of Sm sequence/structure relationships. For instance, a particular site in the Sm sequence profile can be seen to encode more information than most other sites (site 23), and a Gly dominates the uneven distribution of letters at this site. Overlaying the structural elements of the Sm fold (**Fig. 2A**) shows that this site corresponds to the strictly conserved Gly near the middle of the highly bent strand β2 (**Fig. 2C**). Also, the pHMM recapitulates the variability known to occur between strands β3 and β4—i.e., the variable length loop L4 (**Fig. 2A**). However, to our knowledge there is no evidence for distinct Sm1 and Sm2 *motifs*, in a structural or evolutionary sense (for instance, the *Sm2* motif would resemble a partially opened β-hairpin); thus, we avoid this terminology. We also make this as a practical point to avoid confusion, as paralogous SmAP genes have been occasionally referred to as *Sm1* and *Sm2* (e.g., *Afu*,[79,80] *Solfolobus solfataricus*,[92] *Pyrococcus abyssi*[93]). Other issues of terminology also arise.

Nomenclature issues, from a structural bioinformatics perspective. Considered as a complete set of all homologs, the Sm family exhibits immense complexity—in terms of cellular pathways and functional roles (splicing, telomere maintenance, quorum sensing, etc.); in terms of sequence motifs and other sequence-level properties (e.g., domain fusions); in terms of oligomerization (homomeric and heteromeric assemblies, multiplicity of oligomeric states); in terms of structural and physicochemical properties (e.g., multiple RNA-binding regions of Sm rings) and so on. Thus, it may be unsurprising that some ambiguities may have arisen in the Sm literature with respect to nomenclature.

For clarity, the following terminological conventions are used in this review. (1) In terms of protein classification, Sm proteins comprise a *superfamily*;[94,95] nonetheless, in this review we refer to the Sm *family* for simplicity. (2) In terms of sequence and function, Sm proteins go by many names: archaeal homologs have been termed *SmAP*s,[14] the bacterial branch of this family is known for historical reasons as *Hfq*,[18,38] and eukaryotic homologs are referred to as *Sm* (the archetypal Sm core of spliceosomal snRNPs). In addition, the term *Lsm* (*Like-Sm*)[12] was introduced early on to refer to eukaryotic Sm-like proteins, such as the paralogous Lsm1-7 (cytosolic, mRNA decay) and Lsm2-8 rings (nuclear, pre-mRNA maturation).[8] Though generally used in the context of eukaryotes, "*Lsm*" also has been used to label non-eukaryotic homologs, such as those of archaeal origin.[13,96] Here, we attempt to use the labels *Sm*, *Hfq*, etc. only as precisely as is justified by our current knowledge and intended meaning. For example, an occurrence of *Sm*, rather than *SmAP*, means a statement applies to all members of the Sm family (to our knowledge), whereas usage of *Hfq* would indicate that we intend the statement to be limited in scope to the bacterial lineages of Sm. (3) For reasons described above, we avoid describing Sm proteins as consisting of "Sm1+Sm2" motifs. (4) We adopt the labeling of 2° structural elements (SSEs) shown in **Figure 2A**; note that many structurally and biochemically important regions (e.g., RNA-contacting amino acids) lie near loops L2, L3 and L4. (5) The terms *proximal* and *distal* are often used to refer to RNA-contacting surfaces of some Sm rings, such as in Hfq•RNA cocrystal structures. For reasons elaborated below, we instead refer to these surfaces as the *L4* (distal) and *L3* (proximal) faces.

The Sm domain as a module: Lessons from *Pae* SmAP3 and MscS. The post-genomic era affords new insights about Sm protein structure, archaeal and otherwise. With myriad open reading frames (ORFs) and bona fide proteins now known, and with increasingly sensitive bioinformatic methods, the Sm fold can be detected as a structural module in many multi-domain proteins.[97,98] Notably, modularity of Sm domains is consistent with a scaffolding[6] role for some eukaryotic (and perhaps archaeal?) Sm homologs. These Sm-*containing* (Sm-like?) proteins feature a wide range of pairwise similarities to one another, even below the level of significant sequence homology. Based on the properties of most known Sm proteins, Sm-containing ORFs may be expected to assemble into homo- or hetero-mers. However, Sm-containing proteins may also act as monomers, as seen with some eukaryotic Sm orthologs that exhibit highly divergent structures and functions (e.g., the enhancer of RNA decapping protein *EDC3* features an N-term Sm module that does not oligomerize in solution[99]). A recent discovery is remarkable because it links Hfq, Sm modularity and SmAPs: In sequencing studies aimed at examining plasmid-encoded mobile genetic elements in the *Thermococcus* lineage of archaea, Krupovic et al. discovered "Hfq-like" genes in four distinct archaeal plasmids.[100] In three of these plasmids the putative archaeal Hfq is fused to an N-terminal C_2H_2-type zinc-finger domain, suggesting a potential role in DNA binding.

Also striking, Sm-containing "homologs" can be found in pathways entirely unrelated to RNA or DNA metabolism. For example, an Sm domain was unexpectedly found[82] in the crystal structure of a voltage-gated mechanosensitive channel of small conductance, MscS,[101] and can be seen in other structures too (e.g., a biotin ligase; Mura, unpublished data). Least-squares structural superimposition of a SmAP and the MscS_{Sm} domain demonstrates their 3D similarity (**Fig. 2D**). Analogous to SmAPs, the MscS membrane protein also forms homoheptamers; however, that superficial resemblance seems to be the only shared feature between these otherwise unrelated, non-homologous proteins (the Sm domain in MscS does not mediate subunit···subunit contacts in the heptamer). This degree of structural conservation, yet functional divergence, challenges our grasp of Sm structure/function relationships, and may imply a heretical view: That Sm proteins do not, in fact, comprise a homologous superfamily, but rather the Sm fold arose in multiple independent instances over the course of protein structural evolution.

An "augmented" Sm protein can be defined as one that consists of an Sm module and at least one additional structural domain. All three possibilities—N-term Sm, C-term Sm and middle-Sm—have been found. MscS is an example of a middle-Sm domain, and the aforementioned thermococcal plasmid ORFs illustrate C-term Sm(/Hfq) domains. A *Pae* SmAP3 paralog provides the only known structure of an N-term Sm module fused to another domain (**Fig. 4**).[82] SmAPs with similarly augmented C-term domains (CTD) can be detected by sequence analysis, particularly for SmAP3s in the *Sulfolobus* genus of the crenarchaea. The novelty of the mixed α/β fold of the *Pae* SmAP3 CTD limited what could be inferred about its function via comparative sequence or structural analysis, though weak structural similarity was found with a CTD of yeast TATA-box binding protein. In addition to providing a structure of an Sm protein fused to a new fold, *Pae* SmAP3 illuminated (1) the assembly of stable 14-mers both in crystals and in solution, (2) a peculiar form of differential divalent cation-binding by Sm proteins, in a manner coupled to its self-assembly and (3) the large-scale conformational heterogeneity that can occur as a possible feature of augmented Sm proteins. Involvement of the SmAP3 CTD both in metal-binding and in shaping the SmAP3 heptamer interface suggests that the main purpose of this auxiliary domain could be either biochemical or structural (the CTD adds over 15,000 Å^2 of solvent-inaccessible surface area to the ≈4,300 Å^2 heptamer···heptamer interface formed by the Sm domains alone).

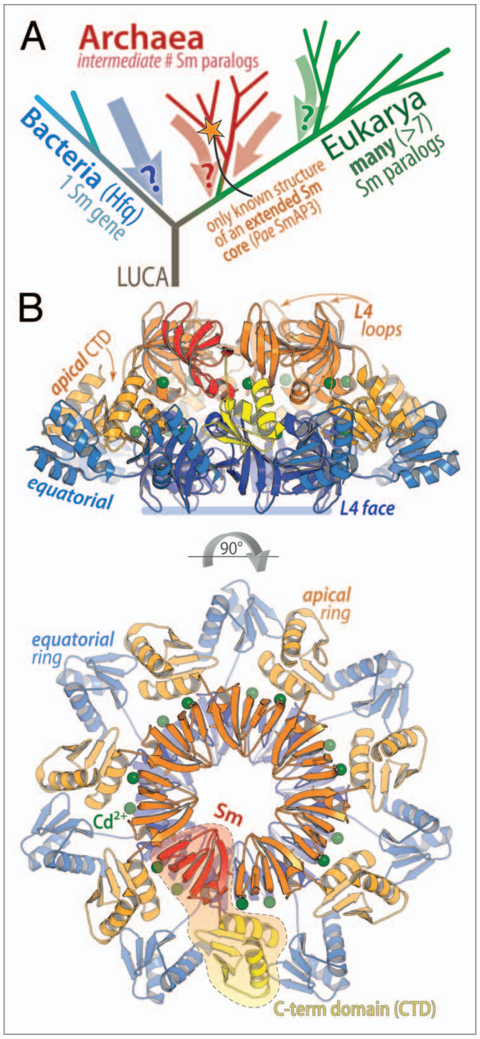

Figure 4. New structural insights from a SmAP3 paralog. A schematic tree of life (**A**) shows the approximate phylogenetic location of *P. aerophilum* SmAP3, which supplies the only known structure of an extended Sm protein. The structure of this paralog (**B**) reveals a core Sm domain (dark hues) decorated with a C-terminal domain (CTD; light hues) that adopts a novel fold; for clarity, a single chain of the tetradecamer is demarcated with a broken line and colored red (Sm domain) and yellow (CTD). This augmented SmAP forms 14-mers and higher-order assemblies both in solution and in crystals, and exhibits intriguing conformational heterogeneity: The CTDs of subunits in the apical ring (orange hues) are hinged 'down' (below the plane of the Sm ring) whereas the CTDs of the equatorial ring (blue hues) splay-out laterally, nearer the plane of the Sm ring. Assembly of the 14-mer is modulated by differential divalent cation-binding in the apical and equatorial subunits (Cd^{2+} ions are shown as green spheres).

Cyclic oligomers and higher-order assemblies. Sm proteins tend to assemble into cyclic oligomers (**Figs. 3 and 4**). Single- and double-ring assemblies occur, as do higher-order polymers. The single-ring oligomers are generally considered to be the biologically functional units. Early EM studies of eukaryotic snRNP particles suggested that the Sm and Lsm cores assemble as "doughnut-shaped heteromers."[102] The gradual realization that Sm/Lsm genes occur in groups of at least seven subtypes within eukaryotic genomes supported an oligomeric structural model; notably, a differential tagging/pull-down experiment established the stoichiometry of the yeast Sm heptamer in vivo and confirmed the sequential order of subunits in the eukaryotic ring.[103] The homo-heptameric nature of an *A. fulgidus* SmAP bound to oligo(U) RNA was established by multivariate statistical analysis of electron micrographs[79] and, concurrently, the first Sm ring structures were reported from a crenarchaeote (*Pae*[14]) and two euryarchaeotes (*Afu*,[15] *Mth*[13]). Each of these SmAPs is homoheptameric. The hetero-heptameric nature of eukaryotic Sm cores was established in a relatively native environment (intact U1 snRNPs) as part of a single-particle cryo-EM reconstruction.[104] Shortly thereafter the first non-heptameric Sm structures were discovered: a second *Afu* SmAP paralog was found to form hexamers (SmAP2),[80] and EM[31] and crystallography[32] revealed hexamers of Hfq. Many lines of genetic, biochemical, biophysical, ultrastructural, NMR and crystallographic data now provide a complex picture of homomeric and heteromeric Sm assemblies. In many cases, an interesting pattern has emerged wherein modern/high-resolution studies are presaged by earlier/lower-resolution results. For instance, an Sm(F·E·G)_2 hexamer was detected in pioneering transmission EM studies of Sm assembly intermediates,[105] and recent crystallographic and NMR studies[85] of the paralogous Lsm triplet revealed an Lsm(6·5·7)_2 hexamer at atomic resolution. The gallery of Sm oligomers in **Figure 3A** includes a trimer (*N*-terminal fragment of a *Schizosaccharomyces pombe* Lsm4[106]), pentamer (an Lsm of putative cyanophage origin[78]), hexamer (*E. coli* Hfq[32]), heptamer (*Pae* SmAP1[14]) and octamer (*S. cerevisiae* Lsm3[107]). The *Pae* SmAP3 tetradecamer is an example of a well-defined higher-order Sm assembly: this double-ring SmAP features an intricate, > 20,000 Å^2 *heptamer-heptamer* interface (**Fig. 4B**).

Despite the severe variation in Sm oligomers, the structural basis of subunit interactions in an Sm ring is fairly clear. In virtually every known structure (canonical Sm, Lsm, SmAP, Hfq), the Sm···Sm interface forms via hydrogen bonds, van der Waals interactions and other interatomic contacts between strands $\beta 4_i$ and $\beta 5_{i+1}$ of subunits *i* and *i*+1. This interface is marked for *n* = 6 in **Figure 3A**. The antiparallel association of neighboring β-strands extends the sheet of the central subunit (**Fig. 2B**) across the entire Sm ring. Consistent with this model of Sm interactions, any Sm dimer (a homodimer excised from a homomeric ring, a heterodimer from a heteromeric ring) can be structurally superimposed on any other dimer with reasonably low RMSD values, demonstrating the structural conservation of the Sm•Sm interface. The greater RMSDs for alignment of dimers vs. monomers (and heptamers vs. dimers) implies that much of the structural variation in an Sm ring is a result of rigid-body displacements of subunits.[81] The only exception that we are aware of to the general $\beta 4_i \cdots \beta 5_{i+1}$ assembly model for bona fide Sm proteins is the recent structure of a truncated construct of *S. pombe* Lsm4 (**Fig. 3A**, *n* = 3); though the atypical β•β interface in this trimer could be an artifact of truncation or crystallization, a similar $\beta 4_i \cdots \beta 4_{i+1}$ interface also occurs between Sm-like domains in a biotin ligase (Mura, unpublished results). In typical Sm, Lsm, SmAP and Hfq rings, the head-tail assembly of subunits that propagates the β-sheet across the Sm ring is enabled by the unique geometric orientation of Sm subunits: the U-shaped Sm monomers are oriented like the blades in a turbine, resulting in the $\beta 4_i$ and $\beta 5_{i+1}$ edge strands being optimally positioned for interaction. The edge strands often contain apolar amino acids that can engage in energetically favorable packing interactions; the standard hydrogen bonding pattern between the β-strand backbones from adjacent subunits can be supplemented by other contacts that further sculpt the β•β interface (e.g., sulfur···π aromatic interactions in *Pae* SmAP1[14]).

In terms of RNA binding, the most salient features of an Sm ring are the topography and physicochemical properties of its surface (binding grooves, electrostatic potential, etc.). The RNA-binding properties of SmAPs have not been thoroughly characterized, though U-rich ssRNAs are known to bind to the face of the ring that corresponds to Hfq's proximal surface.[15,81] Consistent with its RNA chaperone activity, Hfq features a more complex RNA-binding profile: U-rich RNAs primarily contact the proximal side of the ring, A-rich RNAs [e.g., poly(A) tails] bind across the distal surface and a third RNA interaction site was recently identified by Sauer et al. along the lateral rim of the disc-shaped hexamer.[24] The *N*-terminal α-helices found in many Sm structures lie on the *L3*(/proximal) face, opposite the *L4*(/distal) face. However, the sole helix in a pentameric Sm is C-terminal and, thus, is not structurally analogous to the N-term helix (**Fig. 3A**). We raise these points because the *proximal/distal* labels, which were defined relative to the N-term helix face (proximal to the helix), can be structurally ambiguous: *proximal* and *distal* are relative geometric terms that require an external reference frame (an arbitrary point is proximal to some fixed reference point). The terms *L4 face* and *L3 face*, vs. distal and proximal (respectively) avoid this difficulty, as they are referred to fixed structural features of the Sm fold/ring. The L4/L3 labeling scheme also draws attention to the most prominent structural features on the respective face of the Sm ring: the L4 loops appear as turret-like projections, particularly in Sm homologs with longer L4 loops, such as human SmD2 and B (Fig. 1 and Fig. S7 in ref. 56) and the yeast Lsm3 octamer.[107] With respect to the orientation of an Sm monomer subunit, the proximal face is toward loop L3 and the distal face is toward loop L4 in **Figure 2C**.

The spontaneous assembly of Sm monomers into functional rings in the presence or absence of RNA is another key, yet enigmatic feature of Sm oligomerization. As a case in point, consider the *Afu* SmAP2 paralog. Crystallographic and in vitro biophysical characterization of this SmAP show that it can adopt both hexameric and heptameric states, in a manner coupled to both solution pH and RNA-binding.[108] *Afu* SmAP2 hexamers occur at acidic pHs and in the absence of RNA, whereas the addition of U-rich RNA induces the formation of heptamers. Perhaps the coupling between snRNA-binding and SMN-mediated assembly of the canonical eukaryotic Sm snRNP ring, via discrete oligomeric intermediates (discussed above), is an evolutionary echo of the remarkable oligomeric plasticity exemplified by *Afu* SmAP2? Whereas snRNP Sm core assembly is chaperoned by SMN and occurs *on* an RNA site, eukaryotic Lsm complexes autonomously self-assemble into stable rings that then associate with RNA; examples include the nuclear Lsm2-8 complex that binds the 3' terminus of U6 snRNA and the cytosolic Lsm1-7, which associates with P-bodies and is involved in mRNA degradation.[8] Similarly to the eukaryotic Lsm rings, Hfq likely exists in the bacterial cell primarily as pre-formed rings;[109] this is especially likely given Hfq's high intracellular concentration. SmAPs that have been characterized thus far seem more Lsm- and Hfq-like, insofar as they spontaneously self-assemble into rings in solution and in the absence of RNA binding. This distinction between RNA-templated assembly of Sm rings, vs. Sm rings that are stable in the absence of RNA, is related to Scofield and Lynch's functional classification of Sm rings[63] as either *fixed* (specific function, such as in the snRNP Sm core) or *flexible* (generic/multi-functional, such as Hfq or Lsm).

Beyond their self-assembly into cyclic oligomers at the single- and double-ring levels, Sm homologs can also polymerize into fibrillar ultrastructures. Well-defined, finite Sm double-rings, such as Hfq 12-mers and SmAP 14-mers, are often found as head-head ($L3_{face}$-$L3_{face}$) associations of rings in crystal lattices. Though higher-order SmAP complexes (double-ring and beyond) can be detected by in vitro biophysical characterization (e.g., ref. 82), the existence and potential significance of Hfq dodecamers in solution has not been easily resolved; as discussed in reference 110, the detection of Hfq_6 and $(Hfq_6)_2$ species, and the apparent Hfq:RNA stoichiometry, are influenced by the mode of analysis (gel shifts, analytical ultracentrifugation, etc.). In addition to the single- and double-ring oligomers, SmAPs from at least two archaeal lineages (*Pae*, *Mth* SmAP1) undergo head-tail polymerization into well-ordered fibrils.[81] In an intriguing parallel to SmAPs, *E. coli* Hfq also polymerizes into well-ordered fibers with morphologies resembling those of SmAPs, albeit with a different assembly architecture.[111]

The oligomerization/RNA-binding question—The potential biological significance of SmAP and Hfq assemblies remains unclear at the double-ring level and in terms of the various fibrillar polymers. A speculative model for the potential roles of higher-order SmAP assemblies is shown in **Figure 3B**. Here, $(SmAP)_n$ single-rings are indicated as being functional with respect to RNA chaperoning activity (middle panel), while putative $(SmAP)_{2n}$ double-rings (left panel) would exhibit only a subset of interactions (e.g., putative binding of A-rich RNAs to the *L4* face, such as occurs with Hfq, is denoted by "?" marks); multi-ring polymers would be effectively RNA-silent (right panel). As suggested in this simple model, the oligomerization and RNA-binding properties of SmAPs are likely to be intricately coupled. In the model shown in **Figure 3B**, particular oligomeric states of the Sm ring can be viewed as an RNA-coupled molecular switch or as an "RNA-o-stat" (functionally analogous to a thermostat or rheostat, facilitating the cellular pool of RNA···RNA interactions).

The oligomeric plasticity challenge—Viewed across the entire family, Sm complexes exhibit a degree of oligomeric plasticity that outstrips many protein families, despite conservation at the levels of amino acid sequence and 3D fold. Unlike the availability of a geometric theory accounting for the 7-fold symmetry of β-propellers,[112] there has not emerged any general principle relating the order of an Sm oligomeric state (n = 3, 5, …) to whether it is homo- or heteromeric, whether the Sm serves a generic or specific functional niche, and so on. Sm subunits assemble into homo-heptamers (often archaeal), hetero-heptamers (generally eukaryotic) and homo-hexamers (often bacterial, Hfq), though all four possible combinations of ring types—{homomeric, heteromeric} × {hexamers, heptamers}—have been found. Beyond the common heptamer and hexamer states, trimers, pentamers and octamers also exist (**Fig. 3A**). The closest analog to such large-scale quaternary structural variability may be the quasi-equivalent n = 5/6 states adopted by coat proteins in icosahedral virus capsids. What is the physicochemical and stereochemical basis of such immense plasticity? Are Sm ring assembly/disassembly and RNA binding coupled to Sm protein dynamics and allostery? (If so, how?) Pursuit of these and related questions would advance our understanding of the molecular basis of Sm structure and function.

Sm Functional Roles in the Archaea: Scaffolds or Chaperones (or Both, or Neither)?

Despite the availability of data on SmAP 3D structures, oligomerization, ligand-binding and other biophysical and biochemical properties, little is known about the physiological functions of Sm proteins in the archaea (**Fig. 5**). This dearth of knowledge stands in stark contrast to the well-characterized eukaryotic Sm proteins and the recently amassed knowledge of bacterial Hfq function (reviewed in refs. 36, 38 and 113). SmAP function remains opaque, both in terms of broad functional niches/cellular contexts (splicing, telomere maintenance, etc.) as well as specific biochemical properties and detailed molecular interactions in vivo. Do SmAPs act as sRNA chaperones, like Hfq, or do they function primarily as scaffolds for the assembly of complex RNPs, akin to the molecular activities of the eukaryotic Sm proteins? One plausible scenario is that the single Sm ortholog present in essentially all archaeal species serves as a single-stranded nucleic acid-binding chaperone (like Hfq), while the paralogous SmAPs found in some (though not all) archaea serve other, more specific, functional roles. One cannot exclude a fourth possibility: that SmAPs act via altogether different sets of mechanisms, which resemble neither Hfq nor eukaryotic Sm proteins.

What is (definitively) known about RNA processing in the archaea? Like bacteria, and unlike eukarya, archaea generally lack introns in protein coding genes. However, many introns do occur in archaeal tRNA and rRNA genes.[114] Archaeal tRNA introns are typically in the anticodon loop, while rRNA introns occur at diverse locations. Whereas bacterial introns are usually self-splicing (e.g., group I introns), several forms of archaeal intron removal resemble their eukaryotic counterparts in terms

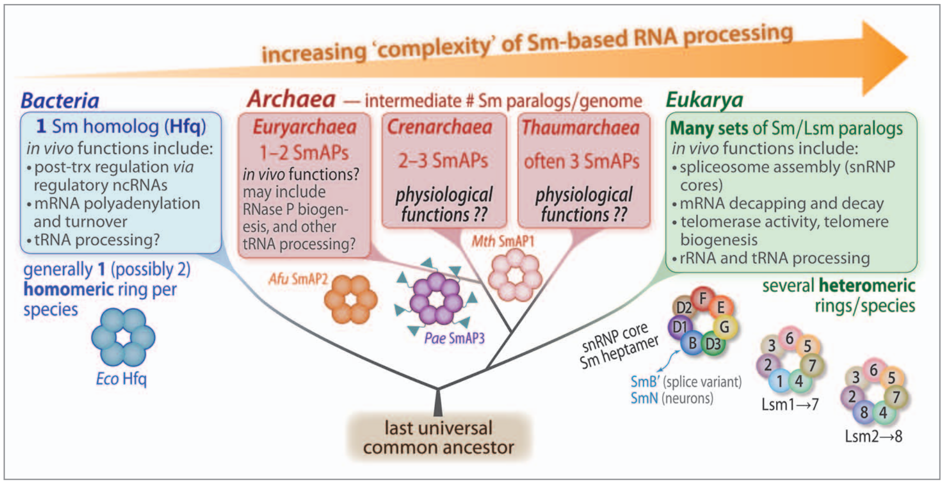
increasing 'complexity' of Sm-based RNA processing
Bacteria
1 Sm homolog (Hfq)
in vivo functions include:
• post-trx regulation via regulatory ncRNAs
• mRNA polyadenylation and turnover
• tRNA processing?
generally 1 (possibly 2) homomeric ring per species
Eco Hfq
Archaea — intermediate # Sm paralogs/genome
Euryarchaea
1–2 SmAPs
in vivo functions? may include RNase P biogenesis, and other tRNA processing?
Crenarchaea
2–3 SmAPs
physiological functions ??
Thaumarchaea
often 3 SmAPs
physiological functions ??
Afu SmAP2
Pae SmAP3
Mth SmAP1
Eukarya
Many sets of Sm/Lsm paralogs
in vivo functions include:
• spliceosome assembly (snRNP cores)
• mRNA decapping and decay
• telomerase activity, telomere biogenesis
• rRNA and tRNA processing
several heteromeric rings/species
snRNP core Sm heptamer
D2 F E D1 G B D3
SmB' (splice variant)
SmN (neurons)
3 6 5 2 7 1 4
Lsm1→7
3 6 5 2 7 8 4
Lsm2→8
last universal common ancestor

Figure 5. Functional repertoire of the Sm fold, from a phylogenomic perspective. This phylogenetic tree shows Sm protein functional roles mapped onto the three domains of life (boxes). The typical number of Sm paralogs/species is indicated for each domain: one Sm per bacterial genome (i.e., Hfq), many Sm per eukaryotic genome, and an intermediate number (1→3) per archaeal genome. Sm oligomerization properties are also indicated. Note that the eukaryotic ring schematics are drawn in correct rotational "register" – *i.e.*, SmF↔Lsm6, SmE↔Lsm5, etc. are the most closely matching pairs of sequences, and are presumably paralogous.

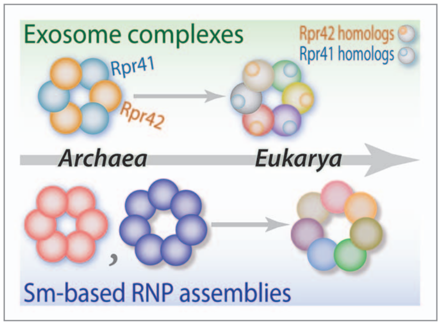

Figure 6. An evolutionary parallel in another RNA-associated system. The archaeal exosome core ring has a relatively simple subunit composition, consisting of a 3 × 2 arrangement of Rpr41 (blue) and Rpr42 (orange) homologs. The eukaryotic ring is a more elaborate heterohexamer of Rpr41, Rpr46 and Mtr3 subunits (all three of which are Rpr41 homologs) along with Rpr42, Rpr43 and Rpr45 (all of which are Rpr42 homologs). The transition from a more primitive (archaeal) to a more sophisticated (eukaryotic) architecture presumably occurred via gene duplication, neutral drift and subsequent subfunctionalization among the paralogs that comprise this RNA-processing machine. This trend is mirrored in the evolution of Sm-based systems from homomeric rings with relatively generic functions (single-stranded nucleic acid-binding) to more sophisticated/specialized heteromeric assemblies.

of a protein requirement, e.g., endonuclease-mediated splicing of archaeal tRNA introns[115] or rRNA processing.[116] The occurrence of archaeal homologs of U3 snoRNP proteins suggests that snoRNP-based rRNA processing may be a shared feature between archaea and eukaryotes.[117] Archaeal RNA processing other than intron removal is also beginning to be characterized, e.g., tRNA 5'- and 3'-end processing.[118] Another RNA processing pathway that appears to be conserved between archaea and eukaryotes is the exosome, a large complex of RNA exonucleases, RNA-binding proteins and RNA helicases that mediates the 3'→5' degradation of mRNA and other RNAs.[119] Intriguingly, exosome evolution mirrors that of Sm assemblies insofar as eukaryotic exosomes feature greater compositional complexity (greater number of heteromer subunits) than their archaeal counterparts (**Fig. 6**).

It is generally assumed that archaea do not have spliceosomal U snRNP-like particles, as their pre-mRNAs are not generally viewed as containing introns. However, there is some precedent for archaeal mRNA introns: the gene for a tRNA- and rRNA-modifying pseudouridine synthase (an archaeal homolog of eukaryotic centromere-binding factor 5, Cbf5p) was found to contain an intron that is spliced in vivo.[120] The intron/exon boundaries in this gene are predicted to adopt bulge-helix-bulge (BHB) motifs, which are the motifs recognized by the splicing endonucleases involved in processing of archaeal pre-tRNAs and rRNAs. It is striking that the intron-containing protein targets tRNAs and rRNAs as substrates, suggesting the potential for co-regulation via modulation of the BHB splicing and ligation apparatus. Although the regulation and diversity of RNA metabolism in archaea may not be as sophisticated as in eukaryotes, these examples suggest that many intricate features of archaeal RNA processing remain to be discovered. The central role of the highly conserved Sm proteins in eukaryotic mRNA processing suggests that archaeal RNA processing may utilize SmAPs in similar RNP assemblies (snRNP-like or otherwise).

Driven by the need for more specific and concrete experimental data about SmAP function in vivo, the past 3 y have seen progress on the cellular functions of these proteins, chiefly via proteomic and RNomic detection of interactions between SmAPs and other proteins or RNAs.[121] The basic strategy has been a "guilt by association" approach (e.g., the CLIP-Seq method),[122] wherein the relevant cellular pathways for a protein or RNA of unknown function are inferred based on the co-precipitated binding partners, a subset of which are presumed to have been functionally characterized. Much of this work has been pioneered by Marchfelder and coworkers in the euryarchaeote *Haloferax volcanii*.[96,121,123]

A bacteria-like (Hfq-like) function in the archaea? Most, if not all, archaeal genomes encode one SmAP, many encode two and some species (primarily among the crenarchaea) encode three Sm paralogs (**Fig. 5**).[82] In terms of sequence similarity, assembly mode (homomeric, heteromeric) and oligomeric states (hexamer, heptamers, etc.), the SmAPs more closely resemble their eukaryotic Lsm counterparts than the bacterial (Hfq) branch of the family.[39,124] Thus, the discovery of an "Hfq-like" protein in *Methanococcus* *jannaschii* (*Mja*), a euryarchaeal methanogen, came as a surprise.[88] Sequence comparisons of this *Mja* Hfq with homologs from *E. coli* (*Eco*) and *Staphylococcus aureus* (*Sau*) suggested conservation of the Sm core of these proteins; the C-terminal tail of *Mja* is quite abbreviated relative to *Eco* Hfq and somewhat shorter than the *Sau* ortholog. Crystallographic work revealed that differences between *Mja* Hfq and the bacterial Hfqs localize near the N-terminal α-helix, the loop L4 variable region and the C-termini. These differences include a shorter N-terminal α-helix in *Mja* Hfq, which correlates with a smaller diameter of the hexamer ring in *Mja* (~54 Å) vs. *Eco* Hfq (~62 Å).[88] The charge distribution on the L4 (distal) face also differs between *Mja* and *Eco* Hfq, which are predominately negative and positive, respectively. Though some bacterial Hfqs also feature an acidic L4 face, the predominately negative charge on the L4 face of *Mja* Hfq suggests that this archaeal Hfq may deviate from the poly(A)-binding site that is characteristic of many bacterial Hfq rings.

Despite these structural and biophysical differences, in vivo studies show that *Mja* Hfq can partially complement the pleiotropic phenotypes of Hfq-knockout mutants in both *E. coli*[88] and *Salmonella enterica*.[125] Specifically, *Mja* Hfq was shown to interact with and stabilize sRNAs[88,125] and participate in sRNA-mediated mRNA turnover.[88] Furthermore, *Mja* Hfq can form a ternary complex with an mRNA (*sucC*) and the sRNA (Spot42) in vitro; interestingly, gel-shift assays suggest that *sucC* may compete with Spot42 at the Hfq-binding site.[88] Competitive binding is not typically observed between sRNAs and their mRNA targets, suggesting that the detailed molecular interactions that underlie the formation of ternary RNA_1-Hfq-RNA_2 complexes may fundamentally differ between this *Mja* Hfq and more extensively studied bacterial homologs such as *Eco* Hfq. Regardless, the *Mja* Hfq work suggests some degree of functional interchangeability between archaeal and bacterial Hfq orthologs.

In addition to the genomically encoded *Mja* Hfq, archaeal "Hfq-like" proteins were recently discovered in four *Thermococcus* plasmids and three unrelated *Methanococcal* plasmids.[100] As described above, these presumptive Hfq homologs contain an N-term C_2H_2-type zinc finger domain fused to a C-term Hfq domain. These novel homologs represent an exciting new group of augmented Sm proteins that may be directed specifically to DNA; intriguingly, both Hfq[26] and SmAPs[81] interact fairly non-specifically with DNA. Functional and structural studies of these new zinc finger-Hfq fusion proteins could greatly illuminate our understanding of both Hfq function and the Hfq/SmAP relationship.

A eukaryote-like (Sm- or Lsm-like) function in the archaea? Although archaea, like bacteria, are unicellular organisms that lack nuclei and other well-defined organelles, many key features of RNA-based cellular metabolism in archaea are more similar to those of eukarya than bacteria. Homologies between rRNAs helped establish that the archaea and eukarya have a shared ancestor that diverged from early bacterial lineages.[16] Other important similarities include archaeal and eukaryal RNA polymerases,[126] and the usage of a specific class of ncRNAs (small nucleolar RNAs, snoRNAs) in both archaea and eukarya to direct modifications to other RNA molecules.[127,128] However, archaea lack many of the sophisticated RNA-processing pathways in which eukaryotic Sm proteins play central and essential roles, including the major and minor spliceosomes and telomere maintenance.[48] Deciphering SmAP function may shed light on the evolution of these key RNA processing features in eukaryotes.

One plausible role for archaeal Sm proteins is in the general biogenesis of abundant and often essential ncRNAs, including tRNAs, rRNAs and snoRNAs; these are all pathways in which eukaryotic Sm proteins are known to play key roles.[12,44,129,130] However, this functional theme of "ncRNA biogenesis" (where "ncRNA" is a placeholder for t/r/sno/etc-RNAs) shows no clearly continuous line extending to the bacteria. In bacteria, the best established Sm function is as a general purpose chaperone for antisense-mediated hybridization of regulatory ncRNAs and their targets, with the Hfq-mediated ncRNA···RNA_{target} interaction typically encoded in *trans*. Hfq also has high affinity for tRNAs, suggesting a direct, but as yet unresolved, role in bacterial tRNA processing or maturation.[131] Intriguingly, Sm···tRNA interactions can also occur in eukaryotes: In studies of SMN-mediated snRNP assembly, Pellizzoni et al. noted an association between the canonical eukaryotic Sm proteins and tRNA,[132] suggesting that Sm···tRNA interactions may not be limited to Hfq. The pleiotropic effects of Hfq inactivation in many bacteria also suggest a potential role in biogenesis of housekeeping RNAs, perhaps independent of its role as a chaperone for regulatory RNAs. However, the fact that Hfq is not strictly required for growth in many species is inconsistent with a vital function in the biogenesis of essential ncRNAs. As described below, an *H. volcanii* SmAP deletion strain was found to be viable, and exhibited a similarly permissive/pleiotropic phenotype as for Hfq-knockout strains in some bacteria.[96]

A potential twist on SmAP function is provided by the eukaryotic "Tudor" domain, which is a five-stranded antiparallel β-sheet[86] that bears a striking resemblance to the Sm fold. Tudor domains occur in many proteins involved in RNA metabolism,[133,134] including the SMN complex that chaperones the assembly of Sm proteins onto snRNA. Tudor domains bind methylated residues on substrate proteins, such as the dimethylated arginines of eukaryotic Sm proteins. The functional linkage and physical interactions between Tudor domains and Sm heteromers occurs in the early stages of snRNP biogenesis. Intriguingly, the Tudor domain is not found in archaeal sequences in the standard protein family databases (Pfam, Superfamily, InterPro, etc.; Mura, unpublished). Thus, the likely absence of a Tudor/SMN system in the archaea implies that SmAPs differ from eukaryotic Sm proteins in not being methylated (archaeal methyltransferase homologs can be detected by sequence analysis); or, if SmAPs are methylated, then such modifications may occur via alternative (non-Tudor) pathways.

Recent investigations using high-throughput sequencing methods have yielded new knowledge about the diversity and abundance of archaeal sRNAs. Many of these sRNAs represent promising partners for functional associations with SmAPs. These include *cis*- and *trans*-encoded antisense RNAs that may modulate post-transcriptional processing of target mRNAs,[135-137] as well as tRNA-derived fragments that may modulate translational efficiency in response to stress[138] (the latter resembles the functional role of tRNA-derived fragments in eukaryotes[139]). Among the first *trans*-acting regulatory RNAs discovered in archaea, there appeared to be a particularly promising candidate for interactions with SmAPs in the euryarchaeal methanogen *Methanosarcina mazei*, but the sRNA showed no particular affinity for either *M. mazei* SmAP paralog in vitro.[137] Though the myriad roles for Hfq and Sm/Lsm proteins suggest highly general functions as RNA chaperones and RNP scaffolds, respectively, eukaryotic Sm proteins have not yet been found to play a role in RNA interference. Similarly, neither Hfq nor SmAPs have been implicated in the processing or targeting of CRISPR-derived RNAs, which function analogously in antisense-mediated defense against phage and other infectious genetic elements and account for a substantial fraction of archaeal sRNAs discovered via high-throughput sequencing studies.[123] The C/D box and H/ACA snoRNAs are frequently among the most abundant sRNAs in archaeal and eukaryotic cells, but SmAPs have not been linked to snoRNA-guided modification of target RNAs.

The functions of SmAPs are unlikely to emerge from obscurity without studies specifically directed at experimental discovery of new interactions in vivo. To date, few such studies have been reported. In a key study of SmAP structure and function, co-immunoprecipitation (co-IP) experiments found both SmAP paralogs in the euryarchaeote *Afu* associated with RNase P RNA (which trims the 5' ends of pre-tRNAs) and a longer precursor, suggesting a role in the maturation of this ubiquitous and essential ribozyme.[15] That work also found that antibodies specific for one *Afu* SmAP could co-precipitate the other paralog; a similar result was found in preliminary co-IP experiments with *Pae* SmAP paralogs (Mura, unpublished data), suggesting the potential interaction of SmAP paralogs in vivo. With respect to **Figure 5**, such an association would represent a small step away from the homomeric complexes of SmAPs and Hfqs, toward the heteromeric Sm complexes of eukaryotes. A recent co-IP study with the SmAP in the euryarchaeal halophile *H. volcanii* recovered a diverse array of RNA and protein-binding partners, but no particularly clear functional themes emerged from the population of sRNAs.[96] Intriguingly, this work found the single *H. volcanii* SmAP ortholog to be inessential for growth; similarly to the bacterial Hfq, genetic inactivation of the SmAP yielded pleiotropic phenotypes and growth defects that were more pronounced under some growth conditions than others. Whether paralogous Sm genes are similarly dispensable in species encoding more than one SmAP remains to be determined. As suggested by **Figure 5**, it is possible that SmAP paralogs became more ingrained in essential cellular pathways as they increased in copy number, and biochemical diversification, along individual lineages of euryarchaea, crenarchaea and thaumarchaea.

What can be inferred from genomic context? Patterns of conservation among gene neighbors provide a way to infer phylogenetic and functional relationships among SmAPs, and between SmAPs and other gene families. Such an approach is potentially useful because, despite their conserved β-barrel 3D structures, the short length and great sequence variation across most Sm proteins limits the utility of sequence-based analysis as a means of function inference.[64] Nearly all sequenced archaeal genomes contain at least one Sm homolog, situated directly adjacent to a gene for ribosomal protein L37e (*rpl37*); this association was first documented when only a few complete archaeal genomes were available.[13] L37e is a zinc finger motif protein. In the euryarchaeote *Haloarcula marismortui*, L37e contacts conserved A-rich patches in 50S rRNA via long N- and C-term extensions. A SmAP gene is virtually always located immediately upstream of L37e and transcribed in the same direction, suggesting co-transcription as part of a conserved operon (and possibly association of the encoded proteins following translation?). In the euryarchaeal halophile *H. volcanii*, L37e was shown to be co-transcribed with the upstream SmAP gene, but was not found to be associated among the proteins co-immunoprecipitated using anti-SmAP antibodies.[96] Nevertheless, the near universality of the genomic association between SmAP and L37e genes in all major archaeal clades suggests a conserved role in processing or stabilization of rRNA; such a function would make SmAPs most homologous to the eukaryotic Lsm proteins, some of which are known to be involved in pre-rRNA maturation.[129] It is also possible that SmAPs and L37e associate in evolutionarily conserved processing of other well-structured ncRNAs, such as tRNAs (tRNA genes often occur adjacent to the Sm-L37e pair) or the RNA component of the tRNA-processing RNase P complex, which was shown to associate with both SmAP paralogs in the euryarchaeon *Afu*.[15,44] *Nanoarchaeum equitans* is among the few exceptions to this Sm-L37e genomic association; instead, this archaeon's SmAP gene is adjacent to (and convergently transcribed, relative to) the gene for an alternative ribosomal zinc-finger motif protein known as "L37ae." *N. equitans* is an obligate endosymbiont with a highly reduced genome that is notable for the absence of a detectable RNase P RNA gene; a corresponding biochemical activity has not been found in *N. equitans*,[140] which may support a role for the Sm-L37e gene tandem in maturation of the RNA component of RNase P.

Whereas most euryarchaea have one or two Sm genes, other archaeal phyla typically encode at least two, and often three, SmAP paralogs. We are unaware of species with four, five or six Sm genes. This pattern in the paralog count—both within archaeal clades and between the archaea and eukarya (**Fig. 5**)—implies that Sm proteins evolved via gene duplication and neutral drift, subject to the geometric constraint that the paralogs assemble into functional homo/heteromeric rings.[17] A gene duplication model, along with gene dosage effects, accounts for the pattern of Sm diversification/subfunctionalization across the tree of life (**Fig. 5**); an analogous evolutionary path is thought to have led to the modern exosome (**Fig. 6**). Eukaryotic Sm/Lsm genes likely underwent two waves of duplication,[64] although lateral gene transfer, which pervades the microbial world,[141,142] has not been excluded as a possible source of multiple Sm genes/species. The conserved genomic context of the second Sm paralog in the euryarchaeal *Archaeoglobaceae*, and a number of methanogens, suggests co-transcription with a homolog of the RNA polymerase III subunit RPC34; in eukaryotes this zinc-finger protein is involved in transcription of ncRNAs, including tRNAs and 5S rRNA.[143] We also note that a SmAP2 paralog in most crenarchaea and thaumarchaea is directly upstream and transcribed in the same direction as a methionine adenosyl transferase (MAT), which is potentially involved in methylation of DNA or RNA.

Other gene context relationships also exist but with some variation among the archaeal clades. Irrespective of this variation, the genomic neighborhood of each SmAP typically includes multiple genes predicted to operate in specific RNA processing pathways. For example, crenarchaeal species in the family *Thermoproteaceae* (which includes *P. aerophilum*) are notable for an abundance and diversity of tRNA introns;[115] in these species, we find that the Sm-L37e gene tandem is often adjacent to a divergently transcribed tRNA splicing endonuclease, again suggesting a role in tRNA splicing and maturation. In contrast, the Sm-MAT gene pair in the *Thermoproteaceae* clade is downstream of a large, well-conserved cluster of genes that includes RNA polymerase subunits and ribosomal proteins—a general contextual feature for at least one SmAP gene in archaeal species with more than one Sm homolog (AE Cozen, unpublished). Other genes that co-occur in the same regions as many SmAP genes include (1) cdc6-type genes possibly involved in cell cycle regulation (in *Sulfolobaceae*), (2) type II/IV secretion genes that may be linked to conjugation (in *Thermoproteaceae*), (3) RecA/RadA homologs potentially involved in DNA recombination (in thaumarchaea) and (4) β-lactamase-type nucleases potentially involved in 3' polyadenlyation of mRNAs (these can be found in most crenarchaea). Again, involvement in RNA-related pathways is a recurring theme from these genomic inferences of SmAP functional roles.

Conclusion, Outlook

Sm proteins exhibit a phenomenal range of RNA-related functionality, from Hfq's activity as an RNA chaperone to the scaffolding roles of eukaryotic Sm proteins. In contrast, the functions of SmAPs remain unknown. Further motivation for studying *archaeal* Sm systems is at least 2-fold: (1) practically, Sm RNPs from thermophilic archaea may prove to be more amenable to structural analysis, such as was the case for the ribosome and (2) conceptually, SmAP-based systems may offer a window into the evolution of modern RNP assemblies (e.g., snRNPs), as well as the origins of Hfq-mediated riboregulation.

Hfq and other Sm proteins seem to achieve their great functional breadth by virtue of their ability to interact with myriad proteins and nucleic acids—either alone, in complex with other Sm proteins, or as a structural domain in a larger polypeptide. This versatility can be attributed to at least four factors: (1) Though small, the Sm fold is a flexible platform (evolutionarily, physiologically) for displaying amino acid side-chains that can interact with proteins (e.g., the two Sm neighbors in a ring, other proteins) as well as nucleic acids (e.g., the multiple RNA-binding sites of Hfq). (2) In higher-order complexes built upon Sm rings, a "complex" function (such as splicing) can be regulated by a "simpler" upstream function (such as the assembly state of the ring); that is, Sm protein activity can be toggled by modulating the assembly state (**Fig. 3B**). Finally, a toroidal architecture enables two types of flexibility: (3) *Biochemical modularity*, wherein exchange of a single subunit within a heteromeric ring can alter the cellular role (e.g., Lsm1-7 and Lsm2-8, which differ by a single subunit). (4) *Oligomeric plasticity* at the level of single rings (e.g., a SmAP that forms homohexamers at low pHs without RNA, but heptamers when bound to U-rich RNA) and higher-order ring assemblies. As the missing link between all that we know about bacterial Hfq function and eukaryotic Sm function (**Fig. 5**), Sm-like archaeal proteins occupy a unique and promising evolutionary niche in RNA biology.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We gratefully acknowledge financial support from the University of Virginia, the Jeffress Memorial Trust (J-971; to C.M.), and NIH training grant T32GM080186 (to P.S.R.). We thank the editor for several helpful suggestions. Note that this review provides only a glimpse of the Sm family, reflecting our own familiarity with a subset of many potential topics; absence of specific citations reflects only space constraints and not opinions about the significance or merit of earlier work by others.

End Notes

[a] Many ncRNAs are referred to as small RNAs (sRNAs), which operationally can be considered to be ssRNA species of less than ≈70-80 nucleotides; however, these are not strictly synonymous—for instance, "long ncRNAs" (lncRNAs) are being discovered at an increasing pace and recognized as an important class of regulatory RNAs.

References

1. Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. Proc Natl Acad Sci USA 1979; 76:5495-9; PMID:316537; http://dx.doi.org/10.1073/pnas.76.11.5495.
2. Tan EM, Kunkel HG. Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. J Immunol 1966; 96:464-71; PMID:5932578.
3. Tsokos GC. In the beginning was Sm. J Immunol 2006; 176:1295-6; PMID:16424152.
4. Lerner MR, Boyle JA, Hardin JA, Steitz JA. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. Science 1981; 211:400-2; PMID:6164096; http://dx.doi.org/10.1126/science.6164096.
5. Kazimierz TT, Nikolay GK, Nicholas KC, Victor F, Steitz JA. (2006). The Ever-Growing World of Small Nuclear Ribonucleoproteins. In *The RNA World* 3rd edit., Vol. 43. Cold Spring Harbor Monographs.
6. Good MC, Zalatan JG, Lim WA. Scaffold proteins: hubs for controlling the flow of cellular information. Science 2011; 332:680-6; PMID:21551057; http://dx.doi.org/10.1126/science.1198701.
7. Patel SB, Bellini M. The assembly of a spliceosomal small nuclear ribonucleoprotein particle. Nucleic Acids Res 2008; 36:6482-93; PMID:18854356; http://dx.doi.org/10.1093/nar/gkn658.
8. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. Int Rev Cell Mol Biol 2009; 272:149-89; PMID:19121818; http://dx.doi.org/10.1016/S1937-6448(08)01604-3.
9. Newman AJ, Nagai K. Structural studies of the spliceosome: blind men and an elephant. Curr Opin Struct Biol 2010; 20:82-9; PMID:20089394; http://dx.doi.org/10.1016/j.sbi.2009.12.003.
10. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol 2011; 3; PMID:21441581; http://dx.doi.org/10.1101/cshperspect.a003707.
11. Séraphin B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. EMBO J 1995; 14:2089-98; PMID:7744014.
12. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. EMBO J 1999; 18:3451-62; PMID:10369684; http://dx.doi.org/10.1093/emboj/18.12.3451.

13. Collins BM, Harrop SJ, Kornfeld GD, Dawes IW, Curmi PM, Mabbutt BC. Crystal structure of a heptameric Sm-like protein complex from archaea: implications for the structure and evolution of snRNPs. J Mol Biol 2001; 309:915-23; PMID:11399068; http://dx.doi.org/10.1006/jmbi.2001.4693.

14. Mura C, Cascio D, Sawaya MR, Eisenberg DS. The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. Proc Natl Acad Sci USA 2001; 98:5532-7; PMID:11331747; http://dx.doi.org/10.1073/pnas.091102298.

15. Törö I, Thore S, Mayer C, Basquin J, Séraphin B, Suck D. RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. EMBO J 2001; 20:2293-303; PMID:11331594; http://dx.doi.org/10.1093/emboj/20.9.2293.

16. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576-9; PMID:2112744; http://dx.doi.org/10.1073/pnas.87.12.4576.

17. Mura C. (2002). The Structures, Functions, and Evolution of Sm-like Archaeal Proteins (SmAPs), University of California, Los Angeles.

18. Franze de Fernandez MT, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. Nature 1968; 219:588-90; PMID:4874917; http://dx.doi.org/10.1038/219588a0.

19. Franze de Fernandez MT, Hayward WS, August JT. Bacterial proteins required for replication of phage Q ribonucleic acid. Pruification and properties of host factor I, a ribonucleic acid-binding protein. J Biol Chem 1972; 247:824-31; PMID:4550762.

20. Carmichael GG, Weber K, Niveleau A, Wahba AJ. The host factor required for RNA phage Qbeta RNA replication in vitro. Intracellular location, quantitation, and purification by polyadenylate-cellulose chromatography. J Biol Chem 1975; 250:3607-12; PMID:805130.

21. Kajitani M, Kato A, Wada A, Inokuchi Y, Ishihama A. Regulation of the *Escherichia coli hfq* gene encoding the host factor for phage Q β. J Bacteriol 1994; 176:531-4; PMID:8288550.

22. Brescia CC, Mikulecky PJ, Feig AL, Sledjeski DD. Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure. RNA 2003; 9:33-43; PMID:12554874; http://dx.doi.org/10.1261/rna.2570803.

23. Mikulecky PJ, Kaw MK, Brescia CC, Takach JC, Sledjeski DD, Feig AL. *Escherichia coli* Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. Nat Struct Mol Biol 2004; 11:1206-14; PMID:15531892; http://dx.doi.org/10.1038/nsmb858.

24. Sauer E, Schmidt S, Weichenrieder O. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. Proc Natl Acad Sci USA 2012; 109:9396-401; PMID:22645344; http://dx.doi.org/10.1073/pnas.1202521109.

25. Takada A, Wachi M, Kaidow A, Takamura M, Nagai K. DNA binding properties of the *hfq* gene product of *Escherichia coli*. Biochem Biophys Res Commun 1997; 236:576-9; PMID:9245691; http://dx.doi.org/10.1006/bbrc.1997.7013.

26. Updegrove TB, Correia JJ, Galletto R, Bujalowski W, Wartell RM. E. coli DNA associated with isolated Hfq interacts with Hfq's distal surface and C-terminal domain. Biochim Biophys Acta 2010; 1799:588-96; PMID:20619373; http://dx.doi.org/10.1016/j.bbagrm.2010.06.007.

27. Azam TA, Hiraga S, Ishihama A. Two types of localization of the DNA-binding proteins within the *Escherichia coli* nucleoid. Genes Cells 2000; 5:613-26; PMID:10947847; http://dx.doi.org/10.1046/j.1365-2443.2000.00350.x.

28. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from *Escherichia coli*. J Mol Biol 2002; 320:705-12; PMID:12095248; http://dx.doi.org/10.1016/S0022-2836(02)00548-X.

29. Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG, et al. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Mol Cell 2002; 9:23-30; PMID:11804583; http://dx.doi.org/10.1016/S1097-2765(01)00436-1.

30. Sun X, Zhulin I, Wartell RM. Predicted structure and phyletic distribution of the RNA-binding protein Hfq. Nucleic Acids Res 2002; 30:3662-71; PMID:12202750; http://dx.doi.org/10.1093/nar/gkf508.

31. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. Mol Cell 2002; 9:11-22; PMID:11804582; http://dx.doi.org/10.1016/S1097-2765(01)00437-3.

32. Schumacher MA, Pearson RF, Møller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. EMBO J 2002; 21:3546-56; PMID:12093755; http://dx.doi.org/10.1093/emboj/cdf322.

33. Moll I, Leitsch D, Steinhauser T, Bläsi U. RNA chaperone activity of the Sm-like Hfq protein. EMBO Rep 2003; 4:284-9; PMID:12634847; http://dx.doi.org/10.1038/sj.embor.embor772.

34. Rajkowitsch L, Schroeder R. Dissecting RNA chaperone activity. RNA 2007; 13:2053-60; PMID:17901153; http://dx.doi.org/10.1261/rna.671807.

35. Valentin-Hansen P, Eriksen M, Udesen C. The bacterial Sm-like protein Hfq: a key player in RNA transactions. Mol Microbiol 2004; 51:1525-33; PMID:15009882; http://dx.doi.org/10.1111/j.1365-2958.2003.03935.x.

36. Brennan RG, Link TM. Hfq structure, function and ligand binding. Curr Opin Microbiol 2007; 10:125-33; PMID:17395525; http://dx.doi.org/10.1016/j.mib.2007.03.015.

37. Chao Y, Vogel J. The role of Hfq in bacterial pathogens. Curr Opin Microbiol 2010; 13:24-33; PMID:20080057; http://dx.doi.org/10.1016/j.mib.2010.01.001.

38. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol 2011; 9:578-89; PMID:21760622; http://dx.doi.org/10.1038/nrmicro2615.

39. Murina VN, Nikulin AD. RNA-binding Sm-like proteins of bacteria and archaea. similarity and difference in structure and function. Biochemistry (Mosc) 2011; 76:1434-49; PMID:22339597; http://dx.doi.org/10.1134/S0006297911130050.

40. Wagner EG. Cycling of RNAs on Hfq. RNA Biol 2013; 10: In press; PMID:23466677; http://dx.doi.org/10.4161/rna.24044.

41. Wilusz CJ, Wilusz J. Lsm proteins and Hfq: Life at the 3' end. RNA Biol 2013; 10: In press; PMID:23392247; http://dx.doi.org/10.4161/rna.23695.

42. Régnier P, Hajnsdorf E. The interplay of Hfq, poly(A) polymerase I and exoribonucleases at the 3' ends of RNAs resulting from Rho-independent termination: A tentative model. RNA Biol 2013; 10: In press; PMID:23392248; http://dx.doi.org/10.4161/rna.23664.

43. Tomasevic N, Peculis BA. Xenopus LSm proteins bind U8 snoRNA via an internal evolutionarily conserved octamer sequence. Mol Cell Biol 2002; 22:4101-12; PMID:12024024; http://dx.doi.org/10.1128/MCB.22.12.4101-4112.2002.

44. Kufel J, Allmang C, Verdone L, Beggs JD, Tollervey D. Lsm proteins are required for normal processing of pre-tRNAs and their efficient association with La-homologous protein Lhp1p. Mol Cell Biol 2002; 22:5248-56; PMID:12077351; http://dx.doi.org/10.1128/MCB.22.14.5248-5256.2002.

45. Schümperli D, Pillai RS. The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein. Cell Mol Life Sci 2004; 61:2560-70; PMID:15526162; http://dx.doi.org/10.1007/s00018-004-4190-0.

46. Dominski Z, Marzluff WF. Formation of the 3' end of histone mRNA: getting closer to the end. Gene 2007; 396:373-90; PMID:17531405; http://dx.doi.org/10.1016/j.gene.2007.04.021.

47. Tharun S, He W, Mayes AE, Lennertz P, Beggs JD, Parker R. Yeast Sm-like proteins function in mRNA decapping and decay. Nature 2000; 404:515-8; PMID:10761922; http://dx.doi.org/10.1038/35006676.

48. Seto AG, Zaug AJ, Sobel SG, Wolin SL, Cech TR. *Saccharomyces cerevisiae* telomerase is an Sm small nuclear ribonucleoprotein particle. Nature 1999; 401:177-80; PMID:10490028; http://dx.doi.org/10.1038/43694.

49. Jurica MS, Moore MJ. Pre-mRNA splicing: awash in a sea of proteins. Mol Cell 2003; 12:5-14; PMID:12887888; http://dx.doi.org/10.1016/S1097-2765(03)00270-3.

50. Jones MH, Guthrie C. Unexpected flexibility in an evolutionarily conserved protein-RNA interaction: genetic analysis of the Sm binding site. EMBO J 1990; 9:2555-61; PMID:2142451.

51. Will CL, Lührmann R. Spliceosomal UsnRNP biogenesis, structure and function. Curr Opin Cell Biol 2001; 13:290-301; PMID:11343899; http://dx.doi.org/10.1016/S0955-0674(00)00211-8.

52. Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA 2011; 2:718-31; PMID:21823231; http://dx.doi.org/10.1002/wrna.87.

53. Stark H, Lührmann R. Cryo-electron microscopy of spliceosomal components. Annu Rev Biophys Biomol Struct 2006; 35:435-57; PMID:16689644; http://dx.doi.org/10.1146/annurev.biophys.35.040405.101953.

54. Pomeranz Krummel DA, Oubridge C, Leung AK, Li J, Nagai K. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. Nature 2009; 458:475-80; PMID:19325628; http://dx.doi.org/10.1038/nature07851.

55. Weber G, Trowitzsch S, Kastner B, Lührmann R, Wahl MC. Functional organization of the Sm core in the crystal structure of human U1 snRNP. EMBO J 2010; 29:4172-84; PMID:21113136; http://dx.doi.org/10.1038/emboj.2010.295.

56. Leung AK, Nagai K, Li J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. Nature 2011; 473:536-9; PMID:21516107; http://dx.doi.org/10.1038/nature09956.

57. Zhang R, So BR, Li P, Yong J, Glisovic T, Wan L, et al. Structure of a key intermediate of the SMN complex reveals Gemin2's crucial function in snRNP assembly. Cell 2011; 146:384-95; PMID:21816274; http://dx.doi.org/10.1016/j.cell.2011.06.043.

58. Raker VA, Plessel G, Lührmann R. The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro. EMBO J 1996; 15:2256-69; PMID:8641291.

59. Grimm C, Chari A, Pelz JP, Kuper J, Kisker C, Diederichs K, et al. Structural Basis of Assembly Chaperone- Mediated snRNP Formation. Mol Cell 2013; 49:692-703; PMID:23333303; http://dx.doi.org/10.1016/j.molcel.2012.12.009.

60. Paushkin S, Gubitz AK, Massenet S, Dreyfuss G. The SMN complex, an assemblyosome of ribonucleoproteins. Curr Opin Cell Biol 2002; 14:305-12; PMID:12067652; http://dx.doi.org/10.1016/S0955-0674(02)00332-0.

61. Yu MC. The Role of Protein Arginine Methylation in mRNP Dynamics. Mol Biol Int 2011; 2011:163827; PMID:22091396; http://dx.doi.org/10.4061/2011/163827.

62. Blackwell E, Ceman S. Arginine methylation of RNA-binding proteins regulates cell function and differentiation. Mol Reprod Dev 2012; 79:163-75; PMID:22345066; http://dx.doi.org/10.1002/mrd.22024.

63. Scofield DG, Lynch M. Evolutionary diversification of the Sm family of RNA-associated proteins. Mol Biol Evol 2008; 25:2255-67; PMID:18687770; http://dx.doi.org/10.1093/molbev/msn175.

64. Veretnik S, Wills C, Youkharibache P, Valas RE, Bourne PE. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. PLoS Comput Biol 2009; 5:e1000315; PMID:19282982; http://dx.doi.org/10.1371/journal.pcbi.1000315.

65. Collins L, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol 2005; 22:1053-66; PMID:15659557; http://dx.doi.org/10.1093/molbev/msi091.

66. Lease RA, Woodson SA. Cycling of the Sm-like protein Hfq on the DsrA small regulatory RNA. J Mol Biol 2004; 344:1211-23; PMID:15561140; http://dx.doi.org/10.1016/j.jmb.2004.10.006.

67. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci USA 2010; 107:9602-7; PMID:20457943; http://dx.doi.org/10.1073/pnas.1004435107.

68. Mohanty BK, Maples VF, Kushner SR. The Sm-like protein Hfq regulates polyadenylation dependent mRNA decay in Escherichia coli. Mol Microbiol 2004; 54:905-20; PMID:15522076; http://dx.doi.org/10.1111/j.1365-2958.2004.04337.x.

69. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proc Natl Acad Sci USA 2002; 99:4620-5; PMID:11917098; http://dx.doi.org/10.1073/pnas.032066599.

70. Zhang A, Wassarman KM, Rosenow C, Tjaden BC, Storz G, Gottesman S. Global analysis of small RNA and mRNA targets of Hfq. Mol Microbiol 2003; 50:1111-24; PMID:14622403; http://dx.doi.org/10.1046/j.1365-2958.2003.03734.x.

71. Christiansen JK, Larsen MH, Ingmer H, Søgaard-Andersen L, Kallipolitis BH. The RNA-binding protein Hfq of Listeria monocytogenes: role in stress tolerance and virulence. J Bacteriol 2004; 186:3355-62; PMID:15150220; http://dx.doi.org/10.1128/JB.186.11.3355-3362.2004.

72. Sittka A, Pfeiffer V, Tedin K, Vogel J. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. Mol Microbiol 2007; 63:193-217; PMID:17163975; http://dx.doi.org/10.1111/j.1365-2958.2006.05489.x.

73. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. Cell 2004; 118:69-82; PMID:15242645; http://dx.doi.org/10.1016/j.cell.2004.06.009.

74. Fromont-Racine M, Mayes AE, Brunet-Simon A, Rain JC, Colley A, Dix I, et al. Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. Yeast 2000; 17:95-110; PMID:10900456; http://dx.doi.org/10.1002/1097-0061(20000630)17:2<95::AID-YEA16>3.0.CO;2-H.

75. Palfi Z, Lücke S, Lahm HW, Lane WS, Kruft V, Bragado-Nilsson E, et al. The spliceosomal snRNP core complex of *Trypanosoma brucei*: cloning and functional analysis reveals seven Sm protein constituents. Proc Natl Acad Sci USA 2000; 97:8967-72; PMID:10900267; http://dx.doi.org/10.1073/pnas.150236097.

76. Lee SI, Steitz JA. *Herpesvirus saimiri* U RNAs are expressed and assembled into ribonucleoprotein particles in the absence of other viral genes. J Virol 1990; 64:3905-15; PMID:2164602.

77. Galão RP, Chari A, Alves-Rodrigues I, Lobão D, Mas A, Kambach C, et al. LSm1-7 complexes bind to specific sites in viral RNA genomes and regulate their translation and replication. RNA 2010; 16:817-27; PMID:20181739; http://dx.doi.org/10.1261/rna.1712910.

78. Das D, Kozbial P, Axelrod HL, Miller MD, McMullan D, Krishna SS, et al. Crystal structure of a novel Sm-like protein of putative cyanophage origin at 2.60 A resolution. *Proteins: Structure, Function*. Bioinformatics 2009; 75:296-307.

79. Achsel T, Stark H, Lührmann R. The Sm domain is an ancient RNA-binding motif with oligo(U) specificity. Proc Natl Acad Sci USA 2001; 98:3685-9; PMID:11259661; http://dx.doi.org/10.1073/pnas.071033998.

80. Törö I, Basquin J, Teo-Dreher H, Suck D. Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile *Archaeoglobus fulgidus*. J Mol Biol 2002; 320:129-42; PMID:12079339; http://dx.doi.org/10.1016/S0022-2836(02)00406-0.

81. Mura C, Kozhukhovsky A, Gingery M, Phillips M, Eisenberg D. The oligomerization and ligand-binding properties of Sm-like archaeal proteins (SmAPs). Protein Sci 2003; 12:832-47; PMID:12649441; http://dx.doi.org/10.1110/ps.0224703.

82. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc Natl Acad Sci USA 2003; 100:4539-44; PMID:12668760; http://dx.doi.org/10.1073/pnas.0538042100.

83. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999; 12:85-94; PMID:10195279; http://dx.doi.org/10.1093/protein/12.2.85.

84. Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, et al. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. Cell 1999; 96:375-87; PMID:10025403; http://dx.doi.org/10.1016/S0092-8674(00)80550-4.

85. Mund M, Neu A, Ullmann J, Neu U, Sprangers R. Structure of the LSm657 complex: an assembly intermediate of the LSm1-7 and LSm2-8 rings. J Mol Biol 2011; 414:165-76; PMID:22001694; http://dx.doi.org/10.1016/j.jmb.2011.09.051.

86. Sprangers R, Groves MR, Sinning I, Sattler M. High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. J Mol Biol 2003; 327:507-20; PMID:12628254; http://dx.doi.org/10.1016/S0022-2836(03)00148-7.

87. Link TM, Valentin-Hansen P, Brennan RG. Structure of *Escherichia coli* Hfq bound to polyriboadenylate RNA. Proc Natl Acad Sci USA 2009; 106:19292-7; PMID:19889981; http://dx.doi.org/10.1073/pnas.0908744106.

88. Nielsen JS, Bøggild A, Andersen CBF, Nielsen G, Boysen A, Brodersen DE, et al. An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from *Methanococcus jannaschii*. RNA 2007; 13:2213-23; PMID:17959927; http://dx.doi.org/10.1261/rna.689007.

89. Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, et al. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. EMBO J 1995; 14:2076-88; PMID:7744013.

90. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res 2012; 40(Database issue):D290-301; PMID:22127870; http://dx.doi.org/10.1093/nar/gkr1065.

91. Eddy SR. Profile hidden Markov models. Bioinformatics 1998; 14:755-63; PMID:9918945; http://dx.doi.org/10.1093/bioinformatics/14.9.755.

92. Kilic T, Thore S, Suck D. Crystal structure of an archaeal Sm protein from *Sulfolobus solfataricus*. Proteins 2005; 61:689-93; PMID:16184597; http://dx.doi.org/10.1002/prot.20637.

93. Thore S, Mayer C, Sauter C, Weeks S, Suck D. Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya. J Biol Chem 2003; 278:1239-47; PMID:12409299; http://dx.doi.org/10.1074/jbc.M207685200.

94. Dayhoff MO. The origin and evolution of protein superfamilies. Fed Proc 1976; 35:2132-8; PMID:181273.

95. Dessailly B, Orengo C. (2009). Function Diversity Within Folds and Superfamilies. In *From Protein Structure to Function with Bioinformatics* (Rigden DJ, ed.), pp. 143-166. Springer Netherlands.

96. Fischer S, Benz J, Späth B, Maier LK, Straub J, Granzow M, et al. The archaeal Lsm protein binds to small RNAs. J Biol Chem 2010; 285:34429-38; PMID:20826804; http://dx.doi.org/10.1074/jbc.M110.118950.

97. Albrecht M, Lengauer T. Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. FEBS Lett 2004; 569:18-26; PMID:15225602; http://dx.doi.org/10.1016/j.febslet.2004.03.126.

98. Anantharaman V, Aravind L. Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability. BMC Genomics 2004; 5:45; PMID:15257761; http://dx.doi.org/10.1186/1471-2164-5-45.

99. Tritschler F, Eulalio A, Truffault V, Hartmann MD, Helms S, Schmidt S, et al. A divergent Sm fold in EDC3 proteins mediates DCP1 binding and P-body targeting. Mol Cell Biol 2007; 27:8600-11; PMID:17923697; http://dx.doi.org/10.1128/MCB.01506-07.

100. Krupovic M, Gonnet M, Hania WB, Forterre P, Erauso G. Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. PLoS One 2013; 8:e49044; PMID:23326305; http://dx.doi.org/10.1371/journal.pone.0049044.

101. Bass RB, Strop P, Barclay M, Rees DC. Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. Science 2002; 298:1582-7; PMID:12446901; http://dx.doi.org/10.1126/science.1077945.

102. Achsel T, Brahms H, Kastner B, Bachi A, Wilm M, Lührmann R. A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro. EMBO J 1999; 18:5789-802; PMID:10523320; http://dx.doi.org/10.1093/emboj/18.20.5789.

103. Walke S, Bragado-Nilsson E, Séraphin B, Nagai K. Stoichiometry of the Sm proteins in yeast spliceosomal snRNPs supports the heptamer ring model of the core domain. J Mol Biol 2001; 308:49-58; PMID:11302706; http://dx.doi.org/10.1006/jmbi.2001.4549.

104. Stark H, Dube P, Lührmann R, Kastner B. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. Nature 2001; 409:539-42; PMID:11206553; http://dx.doi.org/10.1038/35054102.

105. Plessel G, Lührmann R, Kastner B. Electron microscopy of assembly intermediates of the snRNP core: morphological similarities between the RNA-free (E.F.G) protein heteromer and the intact snRNP core. J Mol Biol 1997; 265:87-94; PMID:9020971; http://dx.doi.org/10.1006/jmbi.1996.0713.

106. Wu D, Jiang S, Bowler MW, Song H. Crystal structures of Lsm3, Lsm4 and Lsm5/6/7 from *Schizosaccharomyces pombe*. PLoS One 2012; 7:e36768; PMID:22615807; http://dx.doi.org/10.1371/journal.pone.0036768.

107. Naidoo N, Harrop SJ, Sobti M, Haynes PA, Szymczyna BR, Williamson JR, et al. Crystal structure of Lsm3 octamer from *Saccharomyces cerevisiae*: implications for Lsm ring organisation and recruitment. J Mol Biol 2008; 377:1357-71; PMID:18329667; http://dx.doi.org/10.1016/j.jmb.2008.01.007.

108. Kilic T, Sanglier S, Van Dorsselaer A, Suck D. Oligomerization behavior of the archaeal Sm2-type protein from *Archaeoglobus fulgidus*. Protein Sci 2006; 15:2310-7; PMID:16963646; http://dx.doi.org/10.1110/ps.062191506.

109. Panja S, Woodson SA. Hexamer to monomer equilibrium of E. coli Hfq in solution and its impact on RNA annealing. J Mol Biol 2012; 417:406-12; PMID:22326348; http://dx.doi.org/10.1016/j.jmb.2012.02.009.

110. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. RNA 2011; 17:489-500; PMID:21205841; http://dx.doi.org/10.1261/rna.2452111.

111. Arluison V, Mura C, Guzmán MR, Liquier J, Pellegrini O, Gingery M, et al. Three-dimensional structures of fibrillar Sm proteins: Hfq and other Sm-like proteins. J Mol Biol 2006; 356:86-96; PMID:16337963; http://dx.doi.org/10.1016/j.jmb.2005.11.010.

112. Murzin AG. Structural principles for the propeller assembly of beta-sheets: the preference for seven-fold symmetry. Proteins 1992; 14:191-201; PMID:1409568; http://dx.doi.org/10.1002/prot.340140206.

113. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol 2011; 3: In press; PMID:20980440; http://dx.doi.org/10.1101/cshperspect.a003798.

114. Lykke-Andersen J, Aagaard C, Semionenkov M, Garrett RA. Archaeal introns: splicing, intercellular mobility and evolution. Trends Biochem Sci 1997; 22:326-31; PMID:9301331; http://dx.doi.org/10.1016/S0968-0004(97)01113-4.

115. Sugahara J, Kikuta K, Fujishima K, Yachie N, Tomita M, Kanai A. Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. Mol Biol Evol 2008; 25:2709-16; PMID:18832079; http://dx.doi.org/10.1093/molbev/msn216.

116. Tang TH, Rozhdestvensky TS, d'Orval BC, Bortolin ML, Huber H, Charpentier B, et al. RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. Nucleic Acids Res 2002; 30:921-30; PMID:11842103; http://dx.doi.org/10.1093/nar/30.4.921.

117. Mayer C, Suck D, Poch O. The archaeal homolog of the Imp4 protein, a eukaryotic U3 snoRNP component. Trends Biochem Sci 2001; 26:143-4; PMID:11246005; http://dx.doi.org/10.1016/S0968-0004(00)01779-5.

118. Hartmann RK, Gössringer M, Späth B, Fischer S, Marchfelder A. The making of tRNAs and more - RNase P and tRNase Z. Prog Mol Biol Transl Sci 2009; 85:319-68; PMID:19215776; http://dx.doi.org/10.1016/S0079-6603(08)00808-8.

119. Hartung S, Hopfner KP. Lessons from structural and biochemical studies on the archaeal exosome. Biochem Soc Trans 2009; 37:83-7; PMID:19143607; http://dx.doi.org/10.1042/BST0370083.

120. Watanabe Y, Yokobori S, Inaba T, Yamagishi A, Oshima T, Kawarabayasi Y, et al. Introns in protein-coding genes in Archaea. FEBS Lett 2002; 510:27-30; PMID:11755525; http://dx.doi.org/10.1016/S0014-5793(01)03219-7.

121. Fischer S, Benz J, Späth B, Jellen-Ritter A, Heyer R, Dörr M, et al. Regulatory RNAs in *Haloferax volcanii*. Biochem Soc Trans 2011; 39:159-62; PMID:21265765; http://dx.doi.org/10.1042/BST0390159.

122. Darnell R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. Cold Spring Harb Protoc 2012; 2012:1146-60; PMID:23118367; http://dx.doi.org/10.1101/pdb.prot072132.

123. Marchfelder A, Fischer S, Brendel J, Stoll B, Maier LK, Jäger D, et al. Small RNAs for defence and regulation in archaea. Extremophiles 2012; 16:685-96; PMID:22763819; http://dx.doi.org/10.1007/s00792-012-0469-5.

124. Wilusz CJ, Wilusz J. Eukaryotic Lsm proteins: lessons from bacteria. Nat Struct Mol Biol 2005; 12:1031-6; PMID:16327775; http://dx.doi.org/10.1038/nsmb1037.

125. Sittka A, Sharma CM, Rolle K, Vogel J. Deep sequencing of Salmonella RNA associated with heterologous Hfq proteins *in vivo* reveals small RNAs as a major target class and identifies RNA processing phenotypes. RNA Biol 2009; 6:266-75; PMID:19333007; http://dx.doi.org/10.4161/rna.6.3.8332.

126. Hirata A, Klein BJ, Murakami KS. The X-ray crystal structure of RNA polymerase from Archaea. Nature 2008; 451:851-4; PMID:18235446; http://dx.doi.org/10.1038/nature06530.

127. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. Homologs of small nucleolar RNAs in Archaea. Science 2000; 288:517-22; PMID:10775111; http://dx.doi.org/10.1126/science.288.5465.517.

128. Terns MP, Terns RM. Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin. Gene Expr 2002; 10:17-39; PMID:11868985.

129. Kufel J, Allmang C, Petfalski E, Beggs J, Tollervey D. Lsm Proteins are required for normal processing and stability of ribosomal RNAs. J Biol Chem 2003; 278:2147-56; PMID:12438310; http://dx.doi.org/10.1074/jbc.M208856200.

130. Fernandez CF, Pannone BK, Chen X, Fuchs G, Wolin SL. An Lsm2-Lsm7 complex in Saccharomyces cerevisiae associates with the small nucleolar RNA snR5. Mol Biol Cell 2004; 15:2842-52; PMID:15075370; http://dx.doi.org/10.1091/mbc.E04-02-0116.

131. Lee T, Feig AL. The RNA binding protein Hfq interacts specifically with tRNAs. RNA 2008; 14:514-23; PMID:18230766; http://dx.doi.org/10.1261/rna.531408.

132. Pellizzoni L, Yong J, Dreyfuss G. Essential role for the SMN complex in the specificity of snRNP assembly. Science 2002; 298:1775-9; PMID:12459587; http://dx.doi.org/10.1126/science.1074962.

133. Maurer-Stroh S, Dickens NJ, Hughes-Davies L, Kouzarides T, Eisenhaber F, Ponting CP. The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. Trends Biochem Sci 2003; 28:69-74; PMID:12575993; http://dx.doi.org/10.1016/S0968-0004(03)00004-5.

134. Pek JW, Anand A, Kai T. Tudor domain proteins in development. Development 2012; 139:2255-66; PMID:22669818; http://dx.doi.org/10.1242/dev.073304.

135. Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci USA 2009; 106:21878-82; PMID:19996181; http://dx.doi.org/10.1073/pnas.0909051106.

136. Bernick DL, Dennis PP, Lui LM, Lowe TM. Diversity of Antisense and Other Non-Coding RNAs in Archaea Revealed by Comparative Small RNA Sequencing in Four Pyrobaculum Species. Front Microbiol 2012; 3:231; PMID:22783241; http://dx.doi.org/10.3389/fmicb.2012.00231.

137. Jäger D, Pernitzsch SR, Richter AS, Backofen R, Sharma CM, Schmitz RA. An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains. Nucleic Acids Res 2012; 40:10964-79; PMID:22965121; http://dx.doi.org/10.1093/nar/gks847.

138. Gebetsberger J, Zywicki M, Künzi A, Polacek N. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in Haloferax volcanii. Archaea 2012; 2012:260909; PMID:23326205; http://dx.doi.org/10.1155/2012/260909.

139. Lee SR, Collins K. Starvation-induced cleavage of the tRNA anticodon loop in Tetrahymena thermophila. J Biol Chem 2005; 280:42744-9; PMID:16272149; http://dx.doi.org/10.1074/jbc.M510356200.

140. Randau L, Schröder I, Söll D. Life without RNase P. Nature 2008; 453:120-3; PMID:18451863; http://dx.doi.org/10.1038/nature06833.

141. Boto L. Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci 2010; 277:819-27; PMID:19864285; http://dx.doi.org/10.1098/rspb.2009.1679.

142. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 2011; 7:e1001284; PMID:21298028; http://dx.doi.org/10.1371/journal.pgen.1001284.

143. Blombach F, Makarova KS, Marrero J, Siebers B, Koonin EV, van der Oost J. Identification of an ortholog of the eukaryotic RNA polymerase III subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA polymerases for coding and non-coding RNAs in Archaea. Biol Direct 2009; 4:39; PMID:19828044; http://dx.doi.org/10.1186/1745-6150-4-39.

144. Schuster-Böckler B, Schultz J, Rahmann S. HMM Logos for visualization of protein families. BMC Bioinformatics 2004; 5:7; PMID:14736340; http://dx.doi.org/10.1186/1471-2105-5-7.

145. Durbin R, Eddy SR, Krogh A, Mitchison G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.