AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Bacteriology®

RESEARCH ARTICLE

Nonessential Ribosomal Proteins in Bacteria and Archaea Identified Using Clusters of Orthologous Genes

Michael Y. Galperin,[a] Yuri I. Wolf,[a] Sofya K. Garushyants,[a] Roberto Vera Alvarez,[a] Eugene V. Koonin[a]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

ABSTRACT Ribosomal proteins (RPs) are highly conserved across the bacterial and archaeal domains. Although many RPs are essential for survival, genome analysis demonstrates the absence of some RP genes in many bacterial and archaeal genomes. Furthermore, global transposon mutagenesis and/or targeted deletion studies showed that elimination of some RP genes had only a moderate effect on the bacterial growth rate. Here, we systematically analyzed the evolutionary conservation of RPs in prokaryotes by compiling a list of the ribosomal genes that are missing from one or more genomes in the recently updated version of the Clusters of Orthologous Genes (COG) database. Some of these absences occurred because the respective genes carried frameshifts, presumably resulting from sequencing errors, while others were overlooked and not translated during genome annotation. Apart from these annotation errors, we identified multiple genuine losses of RP genes in a variety of bacteria and archaea. Some of these losses are clade specific, whereas others occur in symbionts and parasites with dramatically reduced genomes. The lists of computationally and experimentally defined nonessential ribosomal genes show a substantial overlap, revealing a common trend in prokaryote ribosome evolution that could be linked to the architecture and assembly of the ribosomes. Thus, RPs that are located at the surface of the ribosome and/or are incorporated at a late stage of ribosome assembly are more likely to be nonessential and to be lost during microbial evolution, particularly in the course of genome compaction.

IMPORTANCE In many prokaryote genomes, one or more ribosomal protein (RP) genes are missing. Analysis of 1,309 prokaryote genomes included in the Clusters of Orthologous Genes (COG) database shows that only about half of the RPs are universally conserved in bacteria and archaea. In contrast, up to 16 other RPs are missing in some genomes, primarily tiny (<1 Mb) genomes of host-associated bacteria and archaea. Six bacterial and nine archaeally specific ribosomal proteins show clear patterns of lineage-specific gene loss. Most of the RPs that are frequently lost from bacterial genomes are located on the ribosome periphery and are nonessential in *Escherichia coli* and *Bacillus subtilis*. These results reveal general trends and common constraints in the architecture and evolution of ribosomes in prokaryotes.

KEYWORDS essential genes, gene loss, genome analysis, ribosomal proteins, ribosome synthesis

Ribosomes are macromolecular cell factories that consist of rRNAs and ribosomal proteins (RPs) and are responsible for the translation of all mRNAs. Bacterial ribosomes that have been thoroughly characterized in model organisms, such as *Escherichia coli* and *Bacillus subtilis*, typically contain 54 core RPs, including 33 in the large subunit and 21 in the small subunit (1–5). Archaeal ribosomes include up to 66 proteins, of which 33 are universal, i.e., shared with bacteria and eukaryotes (18 in the large ribosomal subunit and 15 in the small subunit), and 33 proteins are shared only

Citation Galperin MY, Wolf YI, Garushyants SK, Vera Alvarez R, Koonin EV. 2021. Nonessential ribosomal proteins in bacteria and archaea identified using Clusters of Orthologous Genes. J Bacteriol 203:e00058-21. https://doi.org/10.1128/JB.00058-21.

Editor Tina M. Henkin, Ohio State University

Address correspondence to Michael Y. Galperin, galperin@ncbi.nlm.nih.gov, or Eugene V. Koonin, koonin@ncbi.nlm.nih.gov.

Dedicated to the memory of Alexander S. Spirin (1931 to 2020), the founder and the long-time chair of the Department of Molecular Biology of the Moscow State University and a member of the Russian Academy of Sciences and U.S. National Academy of Sciences, who made pivotal contributions to the study of the topography of ribosomal proteins and protein folding in the ribosome.

Received 31 January 2021
Accepted 11 March 2021
Accepted manuscript posted online 22 March 2021
Published 7 May 2021

with eukaryotes. The list of core RPs in several model organisms is provided in Table S1 in the supplemental material.

Several lines of evidence indicate that some RPs can be nonessential, at least, in some organisms and under certain conditions. First, experiments on genome-wide mutagenesis have resulted in the generation of mutants with a deletion or transposon insertion in a variety of RP genes. Such mutants were viable but grew slower than the wild type (6, 7). Such experiments have been performed in a wide variety of bacteria but so far not in archaea. The global mutagenesis approach has some potential caveats, such as conditional lethality (mutations in each of the two genes are tolerated individually but not together) and functional compensation by paralogs. For example, many bacteria carry paralogs of zinc-containing RPs L31 and L36 that do not bind zinc and, under zinc limitation, replace these RPs (8–10). Similarly, *B. subtilis* carries two paralogs of L31, L33, and S14 that could each partly compensate for the loss of the respective RP function (11, 12). In addition, the absence of certain RP genes can be compensated by changes in the intracellular milieu, such as, for example, a high level of Mg^{2+} ions (13, 14). Gene essentiality data derived from genome-wide mutagenesis studies are well represented in the literature and are also available in online databases, such as the Database of Essential Genes (DEG) (15) and the Online Gene Essentiality database (OGEE) (16). In addition to the global mutagenesis studies, data on RP gene essentiality have been obtained by monitoring the effects of suppressing gene expression, e.g., with antisense RNA (17–19).

Another general approach for the prediction of (non)essential RPs is by using comparative genomics (1–5). The absence of a particular gene in a complete microbial genome (or, better yet, in several related genomes) strongly suggests that this gene is nonessential, at least for growth on a rich medium. This approach also has several caveats, such as the problems with genome completeness and sequencing quality, as well as the presence of paralogs or other forms of functional compensation. However, it is inexpensive, high throughput, and readily applies to hard-to-grow and even noncultured bacteria and archaea. Genome comparisons have proven particularly fruitful for the analysis of the highly reduced genomes of intracellular parasites, insect cell symbionts, and the near-minimal genomes of axenically growing mollicutes (20–27). Collectively, these studies suggest that the number of truly essential RP genes could be as small as 33 (23).

The universal presence of most RPs in organisms from all three domains of life makes them a key component of the small set of highly conserved genes that can be used for the construction of deep-rooted phylogenetic trees and the global Tree of Life (4, 28, 29). Therefore, understanding the evolution of RPs and differentiating universal, essential RPs from dispensable ones that are occasionally lost during evolution are important tasks in evolutionary biology.

Here, we report patterns of the presence and absence of RP genes in the current release of the Clusters of Orthologous Genes (COGs) database (30). The COG database is a particularly convenient tool for the analysis of gene gain and loss because it includes a limited number of high-quality complete microbial genomes and features of COG-specific patterns of the presence and absence of evolutionarily conserved genes in the respective organisms (31–33). In other words, COG profiles show which protein families (COGs) are absent in the given genome(s). In addition, the COG construction algorithm (34, 35) provides for the detection of even highly diverged orthologous proteins that are not necessarily recognized as orthologs by other tools (31–33, 36, 37). Phyletic patterns of COGs have been previously used to reconstruct the ancestral states and evolution of various functional systems, including the minimal and ancestral sets of RPs (2, 38). Owing to these features, the COG database allows for straightforward identification of the genomes that do not encode the given RP.

The current version of the COG database (30) features a selection of COGs grouped by metabolic pathways and functional complexes, including the RPs of the 50S and 30S ribosome subunits as well as a group of archaeally specific RPs. Examination of the

phyletic patterns of the COGs for all three groups allowed us to (i) compile the list of about 500 RP genes missing in some bacterial and/or archaeal genomes (some actually lost and some missing because of sequencing problems), (ii) identify more than 50 RP genes that have been overlooked in the course of genome annotation, (iii) establish the patterns of RP gene loss during bacterial and archaeal evolution, and (iv) correlate the experimentally derived and computationally generated sets of the likely nonessential RP genes.

RESULTS

Delineation of the ribosomal protein set. The conserved ribosomal protein (RP) set, extracted from the current release of the COG database (30), consisted of 54 core bacterial RPs, including 33 from the 50S subunit (L1 to L7/L12, L9 to L11, L13 to L25, and L27 to L36) and 21 (S1 to S21) from the 30S subunit (1–5). Several additional proteins, such as S22 (RpsV, Sra) and S31e (Thx), which are associated with ribosomes in some bacteria (39, 40), are not covered in the COG database and have not been included in the analyzed set. The archaeal RP gene set included 66 genes, of which 33 are shared with bacteria and eukaryotes, and 33 RPs that are shared only with eukaryotes. The list of core RPs from model organisms, such as *Escherichia coli* K-12, *Bacillus subtilis* strain 168, *Mycoplasma pneumoniae* M129, *Aeropyrum pernix* K1, and *Haloarcula marismortui* ATCC 43049, that were analyzed here is presented in Table S1 in the supplemental material. This table shows that the archaeally specific RP set is quite variable; *A. pernix* encodes seven RPs that are missing in *H. marismortui*.

Frameshifted and unannotated ribosomal protein genes. Before analyzing the patterns of RP loss across the diversity of bacteria and archaea, it was necessary to identify and eliminate artifacts that could result from sequencing or annotation errors. To ensure the quality of the genome collection, members of the International Nucleotide Sequence Database Collaboration, the DNA Database of Japan, EBI European Nucleotide Archive, and NCBI GenBank routinely check new genome submissions for the presence of certain RPs (41). Nevertheless, due to the sheer number of sequenced genomes, errors occasionally crop up, which becomes evident when the same organisms repeatedly show up as missing certain RPs despite having relatively large genomes and in the absence of similar problems in related organisms. Another tell-tale sign of sequencing problems is the presence of frameshifted genes that are present in a full-length form in other members of the same lineage (see Table S2 in the supplemental material). There are good reasons to suspect that many if not most of these frameshifts represent sequencing errors, rather than genuine mutations or cases of programmed translational frameshifting that is not known to be a common mechanism of RP translation (42). For example, the 6.09-Mb genome of the betaproteobacterium *Mitsuaria* sp. strain 7 misses the genes for L13, L21, L25, L27, and S9 proteins, which is unique among the genomes of this size. Likewise, the 3.97-Mb genome of the alphaproteobacterium "*Candidatus* Filomicrobium marinum Y" lacks the L1, L7/L12, L10, L11, S7, and S12 genes (Table S2) and is the only genome where the genes for the S7 and S12 proteins are missing.

Another widespread cause of missing RPs is the automated genome annotation, which sometimes fails to recognize genuine protein-coding genes, particularly short ones. As a result, these overlooked open reading frames (ORFs) are not included in the respective protein sets. Sequencing and annotation problems often show up in the same genomes, making their quality suspect and putting into question the apparent absence of certain RPs. As an example, in the 4.28-Mb GenBank entry for the halophilic gammaproteobacterium *Salinicola tamaricis* F01, *rplF* (encoding the L6 protein), *rplI* (L9), *rplL* (L7/L12), *rplY* (L25), and *rpsD* (S4) genes are frameshifted; the *rpsB* (S2) gene is absent; and two full-length genes, namely, *rplD* (L4) and *rpsR* (S18), are present but have been overlooked in the course of genome annotation. Similarly, the current GenBank entry for *Sulfobacillus acidophilus* strain TPY, a member of *Clostridia*, lacks the genes for RPs L27, L28, L32, L33, L36, and S14, which are encoded in the genome but have been overlooked in course of annotation (with the exception of L33, all these

genes are present in the GenBank entry for the type strain of *S. acidophilus*). These genes have been easily found by TBLASTn search (43) using the respective RPs from closely related clostridial genomes as queries (see Table S3 in the supplemental material for details). Two more organisms, namely, *Pelotomaculum thermopropionicum* and "*Candidatus* Methylomirabilis oxyfera" had five overlooked ribosomal genes each (Table S3). Two or more unannotated RP genes were found in seven more bacterial genomes.

The tiny (606 kb) genome of the nanoarchaeon "*Candidatus* Nanopusillus acidilobi" presented a different problem. The current protein set of "*Ca.* Nanopusillus acidilobi" in GenBank misses 14 RPs that are found in almost all other archaeal genomes. However, a detailed examination of this genome showed that only four RP genes were truly missing (Table 1), the gene encoding S14 protein was frameshifted (Table S2), and the gene for L37e had been overlooked and could be found by TBLASTn (Table S3). Full-length ORFs coding for eight other RPs, namely, L6p/L9e (genomic locus tag Nps_02895), L15e (Nps_01385), L16/L10ae (Nps_03305), L22 (Nps_03365), L24 (Nps_02910), L35ae (Nps_03205), S6e (Nps_01880), and S15p/S13e (Nps_01520), were correctly identified at the annotation stage and described in the respective publication (44). However, for some unknown reason, these genes were assumed to be disrupted and were erroneously marked as pseudogenes in the GenBank entry for "*Ca.* Nanopusillus acidilobi" (see Table S3 for details). As a result, the RPs encoded by these genes, which are all longer than 110 amino acids, never made it into the protein database. The same problem on a lesser scale was observed for the other nanoarchaeon in the current COG collection, "Nanohaloarchaea archaeon SG9," where the genes for L24e, L40e, and S28e proteins were overlooked, whereas genes encoding L18 and S2 were marked as pseudogenes and left untranslated (Table S3). Correcting such annotation problems is important for assessing the essentiality of RPs in biologically interesting but poorly studied groups of microorganisms.

Loss of ribosomal protein genes in tiny genomes. Several previous studies investigated the gene contents in organisms with small genome sizes and reported a widespread absence of certain RP genes (2, 23, 27, 45). The most extensive loss of RP genes was observed in the tiny genomes of obligate insect symbionts that include members of *Alphaproteobacteria*, *Betaproteobacteria*, and *Bacteroidetes*. These genomes have undergone dramatic compaction, resulting in genome sizes of less than 1.0 Mb and widespread loss of one or more RP genes (23, 27, 46). Indeed, in some of these tiny genomes, the loss of RP genes was extensive, such that up to 16 RP genes could be missing and several more genes had highly diverged sequences (Table 1). A massive loss of RP genes was also observed in the 593-kb genome of the bryozoan symbiont "bacterium AB1," which is currently unclassified and apparently belongs to a novel major bacterial lineage (47).

As an example, a comparison of the organization of the widely conserved *spc* operon *rplNXE-rpsNH-rplFR-rpsE-rpmD-rplO-secY-rpmJ* (48, 49) in the seven smallest proteobacterial genomes showed that six of them missed *rplX*, the second gene of the operon that encodes L24 (Table 1). In four of these six genomes, *rplN* and *rplE* genes were adjacent with no gap between them (see Fig. S1 in the supplemental material), whereas "*Candidatus* Vidania fulgoroidea OLIH" and "*Candidatus* Hodgkinia cicadicola Dsem" had 139-bp and 160-bp gaps, respectively, but the translated ORFs (GenBank accession numbers AXN02546.1 and ACT34268.1) showed no discernible sequence similarity to L24. In contrast, "*Candidatus* Zinderia insecticola CARI" had a typical *rplX* gene. A similar picture was observed at the distal end of the *spc* operon; five of these seven small genomes lacked the *rpmD* gene with no gap between *rpsE* and *rplO*, whereas "*Ca.* Zinderia insecticola CARI" and "*Candidatus* Tremblaya phenacola PAVE" had the *rpmD* gene, encoding a diverged variant of L30 in the former and a typical one in the latter. Essentially the same pattern was found for the gradual loss of *rplW* (L23) and widespread loss of *rpmC* (L29) genes in the S10 operon (Table 1). These findings suggest that the RP gene loss typically involves sequence divergence and the loss of

TABLE 1 Ribosomal genes missing in organisms with tiny genomes[a]

Organism name[b]	Genome size (kb)	Taxonomy	Missing and highly diverged protein(s) (*n*)[c]	Protein(s) found by TBLASTn
Bacteria				
"*Ca.* Nasuia deltocephalinicola NAS-ALF"	112.1	*Betaproteobacteria*	L1, **L9**, L10, *L13*, *L18*, L19, *L21*, *L22*, **L24**, *L28*, **L29**, **L30**, L31, L32, *L33*, *L35*, S16, *S18*, S20, *S21* (11)	
"*Ca.* Vidania fulgoroideae OLIH"	136.1	*Betaproteobacteria*	**L9**, L10, *L17*, L19, **L21**, *L22*, *L23*, **L24**, *L28*, **L29**, **L30**, *L31*, L32, L35, *S2*, *S6*, *S15*, *S16*, *S17*, S20, **S21** (11)	
"*Ca.* Hodgkinia cicadicola Dsem"	143.8	*Alphaproteobacteria*	L1, **L9**, *L19*, **L23**, **L24**, **L29**, **L30**, L31, *L32*, L34, **S15**, S20, **S21** (11)	S6
"*Ca.* Tremblaya phenacola PAVE"	171.5	*Betaproteobacteria*	**L9**, **L21**, **L23**, **L24**, **L29**, *L32*, L34 (6)	
"*Ca.* Carsonella ruddii DC"	174.0	*Gammaproteobacteria*	**L9**, L10, L17, L18, L19, ***L21***, **L23**, **L24**, L25, **L29**, **L30**, L32, L34, L35, S6, ***S15***, *S18*, S20, **S21** (16)	
"*Ca.* Sulcia muelleri PUNC"	190.7	*Bacteroidetes*	**L23**, **L24**, **L29**, **L30** (4)	
"*Ca.* Zinderia insecticola CARI"	208.6	*Betaproteobacteria*	**L9**, **L23**, *L28*, **L29**, ***L30***, L35, *S6*, *S18*, *S20* (4)	
"*Ca.* Uzinura diaspidicola ASNER"	263.4	*Bacteroidetes*	**L29**	
"*Ca.* Walczuchella monophlebidarum"	309.3	*Bacteroidetes*	**L29**	
"*Ca.* Mikella endobia"	352.8	*Gammaproteobacteria*	—	
"*Ca.* Portiera aleyrodidarum"	357.5	*Gammaproteobacteria*	**L30**	
"*Ca.* Evansia_muelleri"	357.5	*Gammaproteobacteria*	**L9**, **L30**	
"*Ca.* Profftella armatura DC"	464.9	*Betaproteobacteria*	—	
"*Ca.* Purcelliella pentastirinorum"	479.9	*Gammaproteobacteria*	—	
"*Ca.* Moranella endobia"	538.2	*Gammaproteobacteria*	—	
Mycoplasma genitalium G37[b]	580.1	*Mollicutes*	L25, **L30**, S1	
"*Ca.* Riesia pediculicola"	582.1	*Gammaproteobacteria*	**L30**	
Bacterium AB1	593.4	N/A	***L9***, L10, L19, **L21**, **L23**, L25, **L29**, **L30**, L31, L32, L33, L35, S6, **S15**, *S18*, S20, **S21** (15)	L34
Cand. division Kazan bacterium GW2011_GWA1_50_15	602.6	Other bacteria	**L30**, **S21**	L34
Blattabacterium sp. (*Blattella germanica*) strain Bge	641.0	*Bacteroidetes*	**L30**	
Buchnera aphidicola APS (*Acyrthosiphon pisum*)	655.7	*Gammaproteobacteria*	—	
"*Ca.* Hepatoplasma crinochetorum Av"[b]	657.1	*Mollicutes*	**L9**, L25, **L30**, S1, **S21**	
"*Ca.* Nanosynbacter lyticus TM7x"	705.1	Other bacteria	**L9**, L25, **L30**, L32	
"*Ca.* Campbellbacteria bacterium GW2011 OD1 34 28"	752.6	Other bacteria	L1, **L29**, **L30**	L36
"*Ca.* Blochmannia pennsylvanicus BPEN"	791.7	*Gammaproteobacteria*	**L30**	
"*Ca.* Woesebacteria bacterium GW2011 GWF1_31_35"	819.5	Other bacteria	**L9**, **L29**, **L30**	
"*Ca.* Fokinia solitaria"	837.3	*Alphaproteobacteria*	**L30**	
Cand. division TM6 bacterium GW2011 GWF2_28_16	853.1	Other bacteria	**L9**, **L30**, L32, **S21**	L36
Neorickettsia sennetsu Miyayama	859.0	*Alphaproteobacteria*	**L30**	
Cand. division WWE3 bacterium RAAC2_WWE3_1	878.1	Other bacteria	**L9**, **L30**, L32	L34, L36, S14
Berkelbacteria bacterium GW2011 GWE1_39_12	915.1	Other bacteria	**L30**	L36
"*Ca.* Xiphinematobacter Idaho Grape"	915.9	*Verrucomicrobia*	—	
Tropheryma whipplei Twist	927.3	*Actinobacteria*	S21	
"*Ca.* Wolfebacteria bacterium GW2011_GWB1_47_1"	984.4	Other bacteria	L1, **L30**, L33, **S21**	L32, L34
Archaea[d]				
Nanoarchaeum equitans Kin4-M	490.9	Other archaea	**L13e**, L40e, S25e, **S30**	L24e, L37e
"*Ca.* Nanopusillus acidilobi"	605.9	Other archaea	**L13e**, L29, **L39e**, S27e, **S30**	L6/L9e, L16/L10ae, L15e, L22, L24, L35ae, L37e, S6e, S15/S13e

(Continued on next page)

TABLE 1 (Continued)

Organism name[b]	Genome size (kb)	Taxonomy	Missing and highly diverged protein(s) (*n*)[c]	Protein(s) found by TBLASTn
"*Ca.* Mancarchaeum acidiphilum Mia14"	952.3	Other archaea	**L13e**, L20a/L18a, L35ae, L37e, S17e, S25e, S27e, **S30**	
Nanohaloarchaea archaeon SG9	1,118.6	*Euryarchaeota*	**L13e**, L14e, L20a/L18a, L30e, L31e, L34e, L35ae, **L39e**, **S30**	L18, L24e, L40e, S2, S28e
Archaeon GW2011_AR15	1.157.8	Other archaea	**L13e**, L20a/L18a, L40e, S25e, S26e, **S30**	

[a]Organism names, genome sizes, and taxonomic assignments are taken from the NCBI Taxonomy database (81) and are listed as in the COG database (30). The organisms are listed in the order of their genome sizes. Cand., candidate; *Ca.*, *Candidatus*; N/A, not available.

[b]For genome sizes over 600 kb, only selected organisms are shown. Only two representatives of *Tenericutes* (*Mollicutes*) are included. See text for discussion.

[c]Ribosomal proteins that are missing in several distinct lineages are shown in bold; highly diverged proteins and fragments not recognized by the standard CD-search (82) are in italics. A dash indicates the presence of the full set of RPs.

[d]No complete archaeal genomes sequenced so far encode L9, L7/L12, L17, L19, L20, L21, L25, L27, L28, L31 to L36, S1, S6, S16, S18, S20, and S21 (see Table S1). The proteins listed here are those present in other, larger archaeal genomes.

RP function, followed by the complete elimination of the respective ORF, often without the loss of the operon structure.

However, not all bacteria with tiny genomes display a massive loss of RP genes, and indeed, some of them retain nearly all RPs. The 263-kb genome of the flavobacterium "*Candidatus* Uzinura diaspidicola," an endosymbiont of armored scale insects, misses only a single RP gene, *rpmC*, that encodes L29 (Table 1). Similarly, the absence of *rpmC*, but no other RP gene, was observed in another flavobacterium, "*Candidatus* Walczuchella monophlebidarum," which has a slightly larger 309-kb genome. The 641-kb genome of yet another member of *Bacteroidetes*, *Blattabacterium* sp., also misses a single RP gene, namely, in this case, the L30-encoding *rpmD*. The *rpmD* gene is also the only one missing in the genomes of the alphaproteobacterium *Neorickettsia sennetsu* (859 kb) and in some gammaproteobacteria, such as "*Candidatus* Portiera aleyrodidarum" (357 kb), "*Candidatus* Riesia pediculicola" (582 kb), and "*Candidatus* Blochmannia pennsylvanicus" (792 kb). The 837-kb genome of "*Candidatus* Fokinia solitaria," an obligate intracellular endosymbiont of the ciliate *Paramecium* sp., lacks the genes for both L29 and L30 (Table 1).

Some tiny genomes actually encode the full set of core RPs (Fig. 1A). In the investigated genome set, the smallest such genome (353 kb) was from the gammaproteobacterial symbiont of mealybugs "*Candidatus* Mikella endobia." This bacterium inhabits the cytoplasm of the betaproteobacterium "*Candidatus* Tremblaya princeps," which has an even smaller (171 kb) genome (46) and lacks the genes for eight RPs (Table 1). Other insect endosymbionts with tiny genomes that encode the full set of RPs include the alphaproteobacterial psyllid symbiont "*Candidatus* Profftella armatura" (465 kb) and the gammaproteobacterium "*Candidatus* Purcelliella pentastirinorum" and "*Candidatus* Moranella endobia" (genome sizes, 480 kb and 539 kb, respectively) (Table 1). "*Ca.* Moranella endobia" is also an intracellular symbiont of "*Ca.* Tremblaya princeps" (46).

All 122 archaeal genomes included in the COG database lack 21 bacterially-specific RPs, namely, L9, L7/L12, L17, L19 to L21, L25, L27, L28, L31 to L36, S1, S6, S16, S18, S20, and S21 (1, 4, 5) (see Table S1). Only 5 of these 122 archaeal genomes are smaller than 1.2 Mb (Fig. 1B); 3 of these small genomes come from the DPANN superphylum, one comes from *Euryarchaeota*, and one remains unclassified. These genomes show conservation of all the universal RPs and most archaeally specific RP genes. Each of these five genomes lacks the genes for L13e and S30, and in some of them, L20a/L18a and L39e genes are missing as well (Table 1). As mentioned above, a substantial number of RPs, namely, nine in "*Ca.* Nanopusillus acidilobi" and five in "Nanohaloarchaea archaeon SG9," are encoded in the respective genomes and are only missing in GenBank owing to the errors in genome submission (Table 1; Table S3).

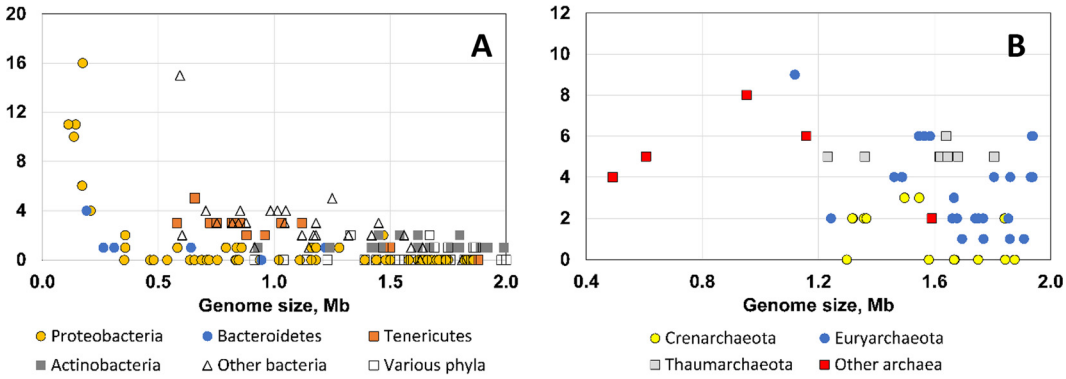

FIG 1 Loss of ribosomal genes in bacteria and archaea with small genome sizes. The numbers of ribosomal protein genes missing in various bacteria (A) and archaea (B) are shown as a function of the genome size. Each symbol indicates a representative organism from those included in the COG database (a single genome per genus). In panel A, yellow circles indicate the genomes of members of *Proteobacteria*; blue circles, *Bacteroidetes*; orange squares, *Tenericutes*; gray squares, *Actinobacteria*; empty squares, representatives of various phyla with 5 to 10 members in COGs (*Aquificae*, *Chlamydiae*, *Chloroflexi*, *Cyanobacteria*, *Fusobacteria*, *Spirochaetes*, *Synergistetes*, *Thermotogae*, and *Verrucomicrobia*); and triangles, representatives of poorly sampled phyla (the "Other bacteria" group in COGs). In panel B, yellow circles indicate genomes of members of *Crenarchaeota*; blue circles, *Euryarchaeota*; gray squares, *Thaumarchaeota*; and red squares, representatives of poorly sampled phyla (the "Other archaea" group in COGs). See Table 1 for the names of representative organisms.

Lineage-specific loss of ribosomal protein genes. Figure 1A shows that, at genome sizes over 1.5 Mb, bacterial genomes rarely lack more than three RPs. At slightly larger genome sizes, most organisms contain the full RP sets. The exact position of the boundary between RP-missing and RP-complete protein sets varies between bacterial lineages but is typically around 2.0 Mb. The lowest such boundary at 0.8 Mb was detected in *Gammaproteobacteria*; the only gammaproteobacterial genome in the COGs that is larger than 0.8 Mb but is missing any RP genes is the abovementioned genome of *Salinicola tamaricis*, where the absence of the *rpsB* gene is likely due to a sequencing error. In the analyzed genome set, the boundary for *Betaproteobacteria* and *Chloroflexi* lies at 1.70 Mb, for *Bacteroidetes* at 1.88 Mb, and for *Alphaproteobacteria* at 2.01 Mb, whereas for *Cyanobacteria* it is 3.34 Mb.

Irrespective of the genome size, no RP gene loss was observed in any representatives of the phyla *Aquificae* (9 genomes, 1.50 to 1.98 Mb), *Chlamydiae* (6 genomes, 1.04 to 3.07 Mb), *Chlorobi* (5 genomes, 2.15 to 3.29 Mb), *Spirochaetes* (11 genomes, 1.14 to 4.70 Mb), and *Synergistetes* (5 genomes, 1.85 to 3.59 Mb) and the proteobacterial class *Epsilonproteobacteria* (12 genomes, 1.64 to 3.19 Mb) that are covered in the current version of COGs. Among poorly represented phyla (the "Other bacteria" group in COGs), the full set of RP genes was found in both members of *Armatimonadetes*, *Gemmatimonadetes*, and *Ignavibacteriae* and all three members of *Thermodesulfobacteria*. *Acidobacteria*, *Deltaproteobacteria*, and *Verrucomicrobia* had a single RP gene missing in a single organism, which could be due to the sequencing problems.

In certain lineages, however, loss of ribosomal genes was consistently detected in regular-size genomes of free-living bacteria and archaea. As shown in Table 2, this type of RP gene loss is often lineage specific. A striking example is the previously reported absence of the *rpsU* (S21) gene in every member of the phylum *Actinobacteria* (4). This trend still held true for the 155 actinobacterial genomes from 149 genera included in the current version of the COG database. An additional check in the NCBI protein database showed that the S21 protein is not encoded by any genome from the phylum *Actinobacteria* sequenced to date. This protein is also missing in all representatives of the phyla *Deinococcus-Thermus*, *Fusobacteria*, and *Thermotogae* (Table 2; see also Table S3 in reference 4). All six representatives of the phylum *Fusobacteria* also lack the *rplY* (L25) gene, which is absent in certain lineages of *Actinobacteria*, *Firmicutes*, and *Tenericutes* as well (Table 2).

Similar lineage-specific patterns of gene loss were detected also in lower-level taxa.

TABLE 2 Lineage-specific loss of ribosomal proteins

Protein(s) (*n*)	Protein characteristic (*n* missing/*n* organisms)[a]
50S subunit	
L2–L6, L14–L16, L20 (9)	Always present
L7/L12, L11, L27, L36 (4)	Missing only in poorly sequenced genomes
L9, L17–L19, L22, L23 (6)	Missing only in tiny genomes (Table 1)
L1, L10, L13, L21, L24, L28 (6)	Missing in some tiny genomes (Table 1) and one additional (poorly sequenced?) genome
L25	Missing in some tiny genomes (Table 1) and in *Coriobacteriia* (11/11), *Bacillales* (6/50), *Lactobacillales* (14/23), *Mollicutes* (12/14), *Negativicutes* (10/10)
L29, L31, L32, L33	Missing in tiny genomes (Table 1) and several other genomes
L30	Missing in some tiny genomes (Table 1), *Halanaerobiales* (6/6), *Pelagibacterales* (2/2), *Rickettsiales* (5/9)
L34	Missing in some tiny genomes (Table 1), *Halanaerobiales* (6/6), *Planctomycetes* (12/14)
L35	Missing in some tiny genomes (Table 1), *Mollicutes* (2/14)
30S subunit	
S1	Missing in *Mollicutes* (11/14), *Dehalococcoidia* (2/2), *Erysipelothrix* (1/1)
S3-S5, S8, S10, S11, S13, S14, S17–S19 (11)	Always present
S6, S15, S16, S20 (4)	Missing only in tiny genomes (Table 1)
S2, S7, S9, S12 (4)	Missing in a single genome, possible sequencing error
S21	Missing in *Actinobacteria* (155/155), *Deinococcus-Thermus* (6/6), *Fusobacteria* (6/6), *Halanaerobiales* (5/6), *Thermotogae* (9/9)
Archaeal ribosomes	
L7ae, L12e, L15e, L18e, L19e, L24e, L32e, L37ae/L43a, L44e, S3ae, S4e, S6e, S8e, S19e, S24e, S28e (16)	Always present
L21e, L31e, L37e, S17e (4)	Missing in 1 genome out of 122
L40e, S27e	Missing in 2 genomes out of 122
L13e	Missing in *Crenarchaeota* (12/25), *Euryarchaeota* (79/79), *Thaumarchaeota* (11/12)
L14e	Missing in *Archaeoglobi* (3/3), *Halobacteria* (31/31), *Methanomicrobia* (18/18), *Thermoplasmata* (10/10), *Thaumarchaeota* (11/12)
L20a/L18a	Missing in *Halobacteriales* (4/11), *Natrialbales* (5/11), *Thaumarchaeota* (12/12)
L30e	Missing in *Halobacteria* (31/31), *Thermoplasmata* (5/10)
L34e	Missing in *Archaeoglobi* (3/3), *Halobacteria* (31/31), *Methanomicrobia* (18/18), *Thermoplasmata* (10/10), *Thaumarchaeota* (12/12)
L35ae	Missing in *Archaeoglobi* (3/3), *Halobacteria* (31/31), *Methanomicrobia* (18/18), *Thermoplasmata* (10/10), *Thaumarchaeota* (12/12)
S25e, S26e, S30	Missing in *Euryarchaeota* (79/79), tiny genomes
S27ae	Missing in *Haloferacales* (7/10)

[a]Data from the 1,309 genomes in the COG database (30); divergent genes (Table 1), frameshifts, and point mutations in RP genes (Table S3), as well as newly translated RPs (Table S4), are not counted as missing.

Thus, in the clostridial order *Halanaerobiales*, five of the six members, namely, *Acetohalobium arabaticum*, *Halanaerobium hydrogeniformans*, *Halobacteroides halobius*, *Halocella* sp. strain SP3-1, and *Halothermothrix orenii*, with genomes in the 2.5- to 4.0-Mb range, lack *rpmD* (L30), *rpmH* (L34), and *rpsU* (S21) genes, whereas the remaining member *Anoxybacter fermentans* only lacks *rpmD* (L30) and *rpmH* (L34). No other clostridial member in the COG system misses the *rpmH* or *rpsU* genes, pinpointing the loss of these genes to the base of the *Halanaerobiales* lineage. Likewise, in the order *Lactobacillales* (class *Bacilli*), the *rplY* (L25) gene is lost in members of three families, namely, *Lactobacillaceae*, *Leuconostocaceae*, and *Streptococcaceae*, but present in the members of *Aerococcaceae*, *Carnobacteriaceae*, and *Enterococcaceae*.

Among the *Archaea*, the L30e protein is missing in all representatives of the euryarchaeal class *Halobacteria* and in all but one representative of the order *Thermoplasmatales* (Table 2).

Widespread ribosomal protein gene loss in *Mollicutes*. The phylum *Tenericutes* presents a remarkable case of RP loss. Most early studies of gene essentiality focused on *Mycoplasma genitalium*, which has a 580-kb genome, the smallest among the known bacteria that are capable of axenic growth and can be obtained in pure culture (the recently sequenced genomes of several strains of *M. genitalium* are all at least

579.5 kb long) (20, 50). The genomes of *M. genitalium* and its close relative *Mycoplasma pneumoniae* were found to lack *rplY* (encoding L25 protein), *rpmD* (L30), and *rpsA* (S1) genes. Genes for all other core RPs were present and, with the possible exception of *rpmB* (L28), *rpsT* (S20), and *rpmGB* (encoding a paralog of L33), none could be disrupted by transposon mutagenesis (20, 24, 50).

Essentially the same pattern of the absence of the genes for L25, L30, and S1 has been detected in other mollicutes as well (23, 45). In the current version of the COGs, the coverage of this group was expanded to include representatives of 12 genera of *Mollicutes* and 2 recently sequenced unclassified members of the phylum *Tenericutes* (30). Among these 14 genomes, the genes for L25, L30, and S1 were missing, besides *Mycoplasma* spp., in representatives of 4 other genera, namely, *Entomoplasma luminosum*, *Mesoplasma florum*, *Spiroplasma chrysopicola*, and *Ureaplasma parvum*. The genome of "*Candidatus* Hepatoplasma crinochetorum," in addition to missing genes L25, L30, and S1, also lacked the genes for L9 and S21; whereas in four other mollicutes, L25 and S1 were missing; and two of these genomes additionally lacked L35. Finally, the slightly larger (1.5 Mb) genome of *Acholeplasma laidlawii* only lacked L25, whereas the two unclassified members of *Tenericutes*, namely, "*Candidatus* Izimaplasma strain HR1" and "Tenericutes bacterium MO-XQ" with their even larger genomes (1.88 and 2.16 Mb, respectively), were found to encode the full set of core RP genes.

Experimentally identified nonessential ribosomal proteins. Over the past 15 to 20 years, numerous studies have been published aiming at the identification of the essential genes in a variety of bacteria (see Table S4 in the supplemental material). In the course of these projects, many genes, including certain RP-encoding genes, were identified as nonessential because their inactivation through transposon insertion or in-frame deletion proved to be nonlethal. We reviewed the relevant literature and compiled the lists of RP genes that have been successfully inactivated and therefore deemed nonessential (Table S4). Table S4 shows that the lists of nonessential RPs can vary dramatically between closely related organisms and, in some cases, even in experiments performed by different groups on the same bacterial strains. It should be noted that these lists include only the genes that have been explicitly reported to be disrupted. As an example, a detailed study of *Bacteroides thetaiotaomicron* strain VPI-5482 (51) identified only 24 essential RP genes, suggesting that others could have been successfully inactivated. However, only mutants lacking L9 and L19 were used in subsequent experiments, positively marking these proteins as nonessential for *B. thetaiotaomicron*. Accordingly, numerous studies that centered on the essential genes and did not report the details of the disruption of nonessential genes (e.g., reference 52) have been ignored. Nevertheless, a comparison of the data obtained on a variety of distinct organisms clearly shows that certain RPs are far more likely to be nonessential than the rest of the set. Furthermore, thorough analyses performed in *E. coli* and *B. subtilis* (6, 7, 53, 54) have resulted in closely similar lists of nonessential RPs (Table S1 and S4).

Loss propensity versus nonessentiality of ribosomal proteins. Table 1 and 2 show that certain RP genes are repeatedly identified as being prone to be lost in a variety of bacteria and archaea. Notably, some of the same genes could be successfully deleted in different organisms (Table S1 and S4). Indeed, a comparison of the data in Table 1, 2, and S4 reveals a consistent pattern: the genes that are often missing in tiny genomes are also nonessential in *E. coli* and/or *B. subtilis* (Table 3). Conversely, the genes that are always found even in tiny bacterial genomes could not be deleted from *E. coli* and, with the sole exception of L15, from *B. subtilis* (Table 3). Table S1 also shows that 12 genes that are dispensable in *E. coli* and/or *B. subtilis* are bacterially specific, that is, missing in archaea and yeast. Indeed, the list of 26 dispensable RP genes (Table 3) includes 16 (of 21) bacterially-specific RPs and 10 (of 33) universal RPs. Thus, bacterially-specific RPs appear more likely to be nonessential than universal ones.

We are unaware of systematic efforts on disruption or deletion of archaeal RPs. However, Table 2 shows that of the 33 archaeally specific RPs, 22 are conserved in nearly all analyzed genomes, whereas the rest exhibit lineage-specific gene losses.

TABLE 3 Comparison of nonessentiality and gene loss of ribosomal proteins

Protein name[a]	Gene name	Deletion[b]		Loss or disruption in tiny genomes[c]								Lineage-specific gene loss[d]
		E. coli	*B. subtilis*	Nas	Vid	Hod	Tre	Car	Sul	Zin	AB1	
uL1	*rplA*	Y	Y	Y	–	Y	–	–	–	–	–	
bL9	*rplI*	Y	Y	Y	Y	Y	Y	Y	–	Y	D	
uL10	*rplJ*	–	–	Y	Y	–	–	Y	–	–	Y	
uL11	*rplK*	Y	Y	–	–	–	–	–	–	–	–	
bL17	*rplQ*	–	–	–	D	–	–	Y	–	–	–	
uL18	*rplR*	–	–	D	–	–	–	Y	–	–	–	
bL19	*rplS*	–	–	Y	Y	D	–	Y	–	–	Y	
bL21	*rplU*	–	–	Y	Y	Y	Y	D	–	–	Y	
uL22	*rplV*	–	Y	D	D	–	–	–	–	–	–	
uL23	*rplW*	–	Y	–	D	Y	Y	Y	Y	Y	Y	
uL24	*rplX*	–	–	Y	Y	Y	Y	Y	Y	–	–	
bL25	*rplY*	Y	Y	–	–	–	–	Y	–	–	–	*Actinobacteria*, *Firmicutes*
bL28	*rpmB*	–	Y	D	D	–	–	–	–	D	–	
uL29	*rpmC*	–	Y	Y	Y	Y	Y	Y	Y	Y	Y	
uL30	*rpmD*	–	–	Y	Y	Y	–	Y	Y	Y	Y	*Clostridia*, *Proteobacteria*
bL31	*rpmE*	Y	Y	Y	Y	Y	–	–	–	–	Y	*Tenericutes*
bL32	*rpmF*	Y	Y	Y	Y	Y	D	Y	–	–	Y	
bL33	*rpmG*	Y	Y	–	–	–	–	–	–	–	Y	
bL34	*rpmH*	–	Y	–	–	Y	Y	Y	–	–	–	*Firmicutes*, *Planctomycetes*
bL35	*rpmI*	Y	Y	D	Y	–	–	Y	–	Y	Y	*Tenericutes*
bL36	*rpmJ*	Y	Y	–	–	–	–	–	–	–	–	
bS6	*rpsF*	Y	Y	–	Y	–	–	Y	–	D	Y	
uS15	*rpsO*	Y	–	–	D	Y	–	D	–	–	Y	
bS18	*rpsR*	–	–	D	–	–	–	D	–	D	D	
bS20	*rpsT*	Y	Y	Y	Y	Y	–	Y	–	D	Y	
bS21	*rpsU*	Y	Y	Y	Y	Y	–	–	–	–	Y	*Actinobacteria*, *Deinococcus-Thermus*, *Fusobacteria*, *Thermotogae*

[a]Ribosomal protein names according to the universal nomenclature (56). Prefix "u" indicates universal conservation of the protein, prefix "b" indicates proteins that are specific for bacteria. A detailed list of core ribosomal proteins in several model organisms with the respective UniProt entries is provided in the Table S1 and, in expanded form in the Excel format, as Table S6. An expanded version of this table is provided as Table S7.

[b]Successfully generated deletion mutants in *E. coli* (6) and *B. subtilis* (7) are indicated as Y, absence of such mutants is indicated by a dash. See Table S4 for details.

[c]Organisms with tiny genomes are listed in the order of their genome sizes (same as in Table 1), as follows: Nas, "*Ca.* Nasuia deltocephalinicola strain NAS-ALF"; Vid, "*Ca.* Vidania fulgoroideae OLIH"; Hod, "*Ca.* Hodgkinia cicadicola Dsem"; Tre, "*Ca.* Tremblaya phenacola PAVE"; Car, "*Ca.* Carsonella ruddii DC"; Sul, "*Ca.* Sulcia muelleri PUNC"; Zin, "*Ca.* Zinderia insecticola CARI"; AB1, "bacterium AB1." Y indicates the loss of the respective gene, divergence and/or disruption of the gene is indicated by D, and presence of the gene is shown by a dash.

[d]Bacterial phyla containing the lineages that exhibit loss of the respective genes (from Table 2).

Loss of ribosomal protein genes in evolution versus ribosome structure and assembly. It is instructive to compare the pattern of RP loss during prokaryote evolution with the location of the respective RPs in the ribosome structure (55–60) and a related characteristic, the order of RP joining during the ribosome assembly (3, 61–65). Fig. 2 and Table S5 show that neither of these features provides a clear-cut prediction of the RP loss propensity and/or (non)essentiality. Indeed, frequently lost bacterial RPs, such as L9, L25, L29, L30, and S21, are located on the surface of the ribosome (Fig. 2A and B). However, L32, which is lost in many tiny genomes, has a significant buried area, whereas L34, which is lost in two bacterial lineages, is mostly buried in the ribosome structure. Conversely, several other surface RPs with relatively small buried areas (L16, L27, S16, S17, and S18) are rarely lost and could not be deleted in either *E. coli* or *B. subtilis* (Table 2 and 3). Likewise, most of the frequently lost RPs (L7/L12, L9, L25, L30, L32, and S21) (see Table S5) are incorporated into the ribosome at the late stages of its assembly (65). However, some of the RPs that join the ribosome early and interact with either 16S (S6 and S20) or 23S (L21, L24, L29, and L34) rRNA can also be lost or deleted (Table 3), whereas the late-addition RPs L6, L16, L27, S2, S3, S10, S13, S14, and S19 are seldom lost (see Table S5 in the supplemental material) and could not be deleted in *E. coli* or *B. subtilis*. Thus, there seems to be, at best, only a weak trend in the expected direction, namely, that RPs that are located on the surface of the ribosome and are attached late during the ribosome assembly are frequently lost in evolution and are often nonessential. A detailed accounting

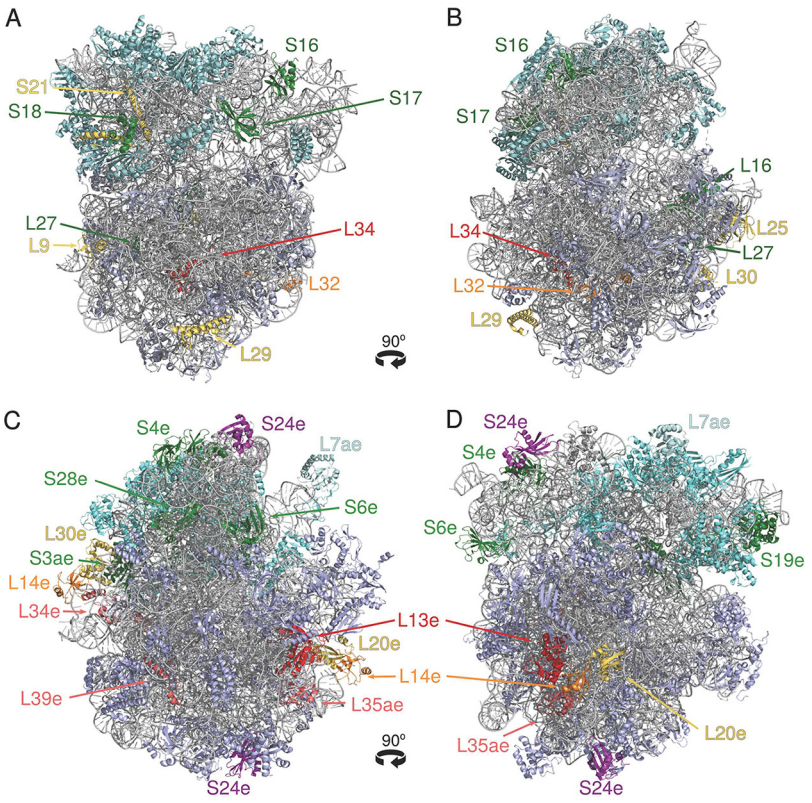

FIG 2 Localization of certain frequently and rarely lost surface proteins in the ribosomes of *Escherichia coli* and *Pyrococcus furiosus*. (A, B) Crystal structure of the *E. coli* ribosome (PDB 7K00), solved at 2-Å resolution by Watson et al. (60). (C, D) Cryo-electron microscopy (cryo-EM) structure of the *P. furiosus* ribosome (PDB 4V6U), solved at 6.6-Å resolution by Armache et al. (59). The rRNAs are shown in gray. Unless indicated otherwise, 50S subunit proteins are in lavender and 30S subunit proteins are in cyan. Proteins mentioned in the text are indicated by bright colors, as follows: frequently lost surface proteins are in yellow; rarely lost ones are in green; and other proteins described in the text are in red, orange, and magenta. The structures were visualized and colored using PyMOL v. 1.0 (Schrödinger, LLC).

of specific protein-rRNA and protein-protein contacts (27) could eventually provide a better predictor of the RP loss propensity (nonessentiality).

Similar trends are detectable among the archaeally specific RPs. The L13e protein, which is lost in all tiny archaeal genomes (Table 1) and is also missing in euryarchaea and nearly all thaumarchaea (Table 2), is only partially surface exposed; its N-terminal loop and the first α-helix project deep into the core of the 50S subunit (Fig. 2C and D). Of the two "promiscuous" surface proteins L14e and S24e that are present in two copies in the archaeal ribosome (59) (Fig. 2C and D), L14e is often lost but S24e is never missing in archaea (Table 2). Among other frequently lost archaeally specific RPs (Table 2), L20a/L18a (also known as LX) and L30e are surface proteins, but L34e and L35ae are mostly buried and L39e is only partially surface exposed (58, 59). Conversely, surface-exposed proteins S3ae, S4e, S6e, S19e, and S28e (Fig. 2C and D) were never found to be missing in any of the analyzed archaeal genomes (Table 2).

DISCUSSION

The overall conservation of the translation machinery among the bacteria, archaea, and eukaryotes (see Table S1 in the supplemental material) is the strongest evidence of the common origin of all organisms, which allows their inclusion in a single, universal Tree of Life (28, 29, 66). Indeed, 33 RPs are universal, that is, in all likelihood they have been conserved throughout the more than 3.5 billion years of the evolution of life (67, 68). Therefore, it is remarkable that genes for some of these universal RPs can

be lost in many bacterial and archaeal organisms with tiny genomes (Table 1), as well as in certain bacterial and archaeal lineages with larger genomes (Table 2), and can be deleted from the genomes of model bacteria without substantial loss of viability (Table 3 and S4). Furthermore, most of these genes are also missing (see Table S6 and S7 in the supplemental material) in the largely overlapping mitochondrial and chloroplast RP gene sets (69). By analogy with the gene transfer from plastids and mitochondria to the nucleus, some of the RP genes that are missing in the tiny genomes of intracellular symbionts might have been transferred from the symbiont to the host genome (23, 46). In one case, a 21-kDa product of an aphid host gene has been reported to be specifically produced in the bacteriocyte (70). While this could explain the massive RP loss in some of the tiny genomes, the possibility of interspecies RP transfer was not investigated in this work.

Here, we sought to trace the loss of RP genes in a relatively small, well-defined set of bacterial and archaeal genomes covered by the recent release of the COG database (30). This work was prompted by the observation that relatively few of the RP COGs had "perfect" phyletic patterns, that is, included representatives of all 1,309 organisms (or, in the case of domain-specific RPs, all representatives of either 1,187 bacteria or 122 archaea, respectively). Based on the previous studies (1–5, 23, 27, 45), the missing RPs were expected to come primarily from the highly degraded genomes with an additional contribution of lineage-specific gene loss. These expectations proved to be largely correct, with organisms with tiny genomes (Table 1) and lineage-specific gene loss (Table 2) accounting for a large fraction of imperfect phyletic patterns among the RPs.

In addition, we identified multiple instances of frameshifted ORFs (Table S2) that were likely generated by sequencing errors. In certain cases, these frameshifts occurred in long stretches of identical nucleotides, which raises the possibility that some of them could represent authentic programmed frameshifts (42). This possibility, however, seems unlikely in cases where the genome of a closely related bacterium encodes an intact full-length ORF. Imperfect phyletic patterns can also be caused by problems in genome annotation whereby certain ORFs, particularly short ones, are overlooked by the annotation software (Table S3). Given the widespread loss of RP genes (Table 1 and 2), it would not be realistic to require every newly sequenced genome to contain the full set of RP genes. Nevertheless, when a bacterial or archaeal genome of more than 1 Mb long lacks any of the 43 widely conserved RPs (Table 2), it should raise a red flag. Furthermore, the example of "*Ca.* Nanopusillus acidilobi" (Table 1) shows that short and/or divergent RPs from poorly studied bacteria and archaea should not be deemed pseudogenes without clear evidence that this is indeed the case. In particular, as shown in Table 2, almost all archaeal genomes encode the 33 universal and 20 archaeally specific RPs, so that the absence of any of these genes in an archaeal genome is highly unlikely.

So, what conclusions can be drawn from the patterns of RP loss—and conservation—shown in Table 1, 2, and 3? First, these observations validate the previously noted trend of an independent loss of orthologous RP genes in several phylogenetically distant lineages (4). Examples include the loss of L25 in certain members of *Actinobacteria*, *Firmicutes*, and *Mollicutes*; loss of L30 in some members of *Clostridia* and *Alphaproteobacteria*; loss of L34 in certain *Clostridia* and *Planctomycetes* members; and the loss of S21 in several distinct bacterial phyla (Table 2). Among archaeally specific RPs, it is worth noting the simultaneous absence of L13e, L14e, L20a, L34e, and L35ae proteins in many members of *Euryarchaeota* and *Thaumarchaeota* (Table 2).

The second prominent trend is the gradual loss of RPs within a single lineage. Thus, previous analyses of the mollicute genomes reported the absence of the genes for L25, L30, and S1 (23, 45). This pattern was confirmed here for several mollicute genomes, albeit not for the two recently sequenced unclassified members of *Tenericutes*. These comparisons allowed us to reconstruct a possible scenario of RP gene loss in the phylum *Tenericutes* (Fig. 3). It appears that progressive genome reduction during the

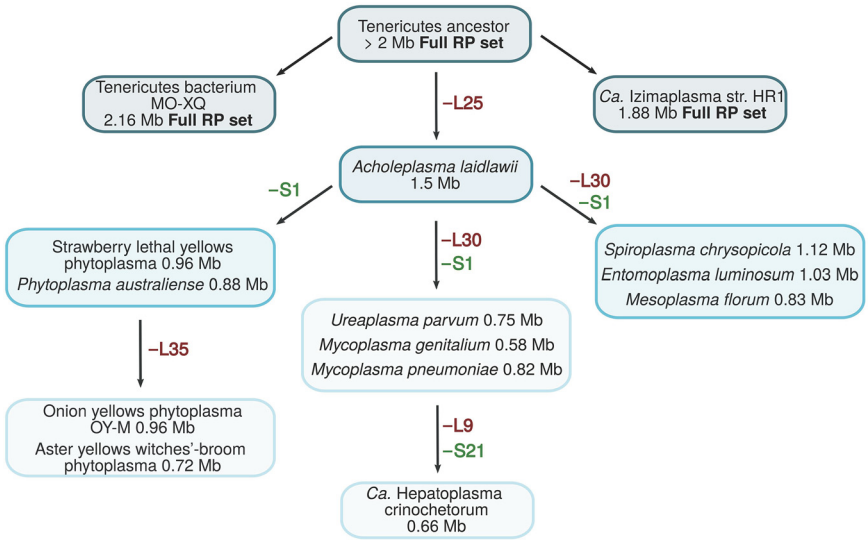

FIG 3 A possible scenario of ribosomal protein loss in *Mollicutes* members (*Tenericutes*). The organisms are those listed in the COG database. Based on the phylogeny of this group, which divides *Mollicutes* into at least four distinct lineages, namely, *Acholeplasma*/*Phytoplasma*, *Spiroplasma*/*Mesoplasma*, *Mycoplasma pneumoniae*/*Ureaplasma*, and *Mycoplasma hominis* (84), the loss of L30 and S1 may have occurred independently on two or more occasions. The loss of L25 may have occurred at an early stage in the evolution of this group or has occurred several times.

evolution of the *Mollicutes* first led to the loss of L25 and then S1, followed by either L30 or L35, and culminated in the loss of two more genes in "*Ca.* Hepatoplasma crinochetorum" (Fig. 3). Remarkably, the same set of genes coding for L25, L30, and S1 that is missing in *Mycoplasma* spp. is also missing in the genome of *Erysipelothrix rhusiopathiae*, a member of the *Firmicutes* branch that is closest to *Mollicutes* (71).

The loss of RPs in phylum- or class-level lineages generally correlates with two other hallmarks of nonessentiality, namely, the availability of deletion mutants in model organisms and frequency of loss in tiny genomes (Table 3). The genes that encode apparently nonessential RPs, but for which few or no losses were observed in the genomes included in the COGs, are likely to be lost in other bacterial or archaeal genomes and especially in small ones. For example, the loss of L1 and/or L9, which was detected in several tiny genomes from "Other bacteria" (Table 1), has been reported to be widespread among the "Candidate Phyla Radiation" (*Patescibacteria*), a vast and diversified group of poorly characterized bacteria that are thought to be symbionts or parasites of other bacteria (72).

An interesting aspect of the nonessential RP gene set is its potential use in synthetic biology. In previous attempts to construct a "minimal" bacterial cell, either fully synthetic (25, 73–75) or highly streamlined (76), the researchers aimed at obtaining rapidly growing microorganisms and chose not to modify their RP gene content. Accordingly, the synthetic *Mycoplasma genitalium* JCVI-1.0 (GenBank accession number CP000925) and both synthetic versions of *Mycoplasma mycoides*, namely, JCVI-syn1.0 (CP002027) and JCVI-syn3.0 (CP016816), included all 50 RP genes that are normally found in these organisms (which do not encode L25, L30, and S1). The synthetic genome of *Caulobacter ethensis* 2.0 included all core RPs of *Caulobacter crescentus* except for L1 and L6 (75). The MiniBacillus project ended up including all 54 core RP genes (Table S1), as well as YlxQ (L7ae) and paralogs of L6, L33, and S14 (76, 77). Future attempts at constructing streamlined bacterial genomes might involve attempts to substantially reduce the sets of RP genes. In contrast, in archaea, the overall conservation of the RPs leaves few choices for such gene deletion.

It should be noted that the absence of RP genes was discussed here—and elsewhere—in terms of genome compaction and lineage-specific gene loss, based on the

presence of the respective RP genes in the genomes of closely related organisms. However, a recent study (78) has shown that certain RPs are encoded by phages, which indicated the distinct possibility of the acquisition of the RP genes through lateral transfer. The L7/L12 and S21 genes appear to be most widespread in phages, and some phages also encode L9 and S30. Furthermore, analysis of viral metagenome sequences has demonstrated the occasional presence of genes for L11, L19, L31, L33, S6, S9, S15, and S20 and, less frequently, for L2 and L10 (78). The presence of such genes in phage genomes could be explained by the pressure on the phage to provide the cell with its own RPs to accelerate translation, particularly when the respective genes are missing in the host genome. Indeed, the RPs listed above are often lost, both in organisms with tiny genomes and in specific bacterial lineages (Table 1 and 2).

Overall, the observations presented here show that the evolution of RPs is more malleable and dynamic than previously thought. It remains to be seen whether additional massive sequencing of diverse bacterial and archaeal genomes leads to further erosion of the set of universal RPs and/or of those that are conserved within the archaeal or bacterial domains of life.

MATERIALS AND METHODS

Genome coverage and protein selection. The list of bacterial and archaeal genomes used in this work was taken from the recent release of the COG database (30). This set includes 1,309 complete genomes of 1,187 bacteria and 122 archaea, most of them with a single representative of the respective genus (1,234 named genera; see https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/cog-20.org.csv for the full list).

The list of RPs analyzed in this work was also taken from the COG database (the "Ribosome 30S subunit," "Ribosome 50S subunit," and "Archaeal ribosomal proteins" groups in the COG pathways list; https://www.ncbi.nlm.nih.gov/research/cog/pathways). This list included 54 bacterial (or universal) proteins and 33 archaeally specific proteins, as listed in the Table S1 in the supplemental material. Two auxiliary RPs, namely, S22 (RpsV, Sra) and S31e (Thx), were not included in this survey because there were no respective COGs in the database. The S22 protein is mostly expressed during the stationary phase and appears to be nonessential for the viability of *E. coli* (40). S31e (Thx) is part of the 30S subunit in *Thermus thermophilus* (57) and is mostly found in *Bacteroidetes*, *Proteobacteria*, and several other phyla. The RNA-binding protein L7ae (YlxQ) is associated with archaeal ribosomes but apparently not with bacterial ribosomes (79). The S1 protein was deemed present when the respective ORF included three or more S1-like domains. The archaeal protein set did not include L38e and L41e proteins (arCOG04057 and arCOG06624 in reference 80, respectively), which are not represented in the current set of COGs.

Identification of missing ribosomal genes. The list of RPs missing from each genome was taken from the phyletic profiles of the respective COGs. The nucleotide sequences of the respective genomes were searched with representative RP sequences (taken either from Table S1 or from closely related taxa) using the recent version of the TBLASTn program (43) that allows the selection of specific organisms based on the NCBI taxonomy (81) assignments. The resulting BLAST hits (cutoff E value, 0.1) were verified using CD-search (82) and compared against the protein sets in GenBank and RefSeq databases. The identity of the RPs that were listed in GenBank and/or UniProt but were not recognized by the standard CD-search was checked using CD-search with relaxed parameters (E value cutoff of 100) and HHpred (83); such RPs are listed as "highly diverged" in Table 1. The confirmed genuine RP ORFs that were missing in GenBank were classified as either frameshifted (or interrupted by a stop codon or missing a recognizable start codon) or overlooked; for the overlooked ORFs, the full-size ORFs were translated from the genomic sequences using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/). Representative frameshifted and overlooked ORFs are listed, respectively, in Tables S2 and S3.

The RPs that produced no statistically significant hits in TBLASTn searches were classified as missing in the respective genomes. These genomes were classified into tiny (less than 1 Mb long for bacteria or 1.2 Mb for archaea) and regular size; they were further sorted by phyla according to their COG assignments (which rely on the NCBI taxonomy database).

Identification of nonessential ribosomal genes. The lists of nonessential ribosomal genes (Table S4) were collected from the literature and two online databases, namely, Database of Essential Genes (DEG) (15) and the Online Gene Essentiality database (OGEE) (16). Since these databases, and most of the original literature, focused on essential genes, the supplemental material files for each paper were individually checked to select those genes that had been positively identified as nonessential and ignore those genes that were not listed as essential but whose status had not been specified.

To assess the ribosomal localization of selected RPs, the structures of ribosomes of *E. coli* (PDB 7K00) (60) and *Pyrococcus furiosus* (PDB 4V6U) (59) were downloaded from the Protein Data Bank and visualized using PyMOL v. 1.0 (Schrödinger, LLC). Individual surface proteins were colored based on their loss propensity (Table 2).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.5 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.04 MB.

ACKNOWLEDGMENT

This study was supported by the Intramural Research Program of the U.S. National Library of Medicine at the National Institutes of Health.

REFERENCES

1. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 30:5382–5390. https://doi.org/10.1093/nar/gkf693.
2. Mushegian A. 2005. Protein content of minimal and ancestral ribosome. RNA 11:1400–1406. https://doi.org/10.1261/rna.2180205.
3. Kaczanowska M, Rydén-Aulin M. 2007. Ribosome biogenesis and the translation process in *Escherichia coli*. Microbiol Mol Biol Rev 71:477–494. https://doi.org/10.1128/MMBR.00013-07.
4. Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS One 7:e36972. https://doi.org/10.1371/journal.pone.0036972.
5. Ban N, Beckmann R, Cate JH, Dinman JD, Dragon F, Ellis SR, Lafontaine DL, Lindahl L, Liljas A, Lipton JM, McAlear MA, Moore PB, Noller HF, Ortega J, Panse VG, Ramakrishnan V, Spahn CM, Steitz TA, Tchorzewski M, Tollervey D, Warren AJ, Williamson JR, Wilson D, Yonath A, Yusupov M. 2014. A new system for naming ribosomal proteins. Curr Opin Struct Biol 24:165–169. https://doi.org/10.1016/j.sbi.2014.01.002.
6. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006.0008. https://doi.org/10.1038/msb4100050.
7. Akanuma G, Nanamiya H, Natori Y, Yano K, Suzuki S, Omata S, Ishizuka M, Sekine Y, Kawamura F. 2012. Inactivation of ribosomal protein genes in *Bacillus subtilis* reveals importance of each ribosomal protein for cell proliferation and cell differentiation. J Bacteriol 194:6282–6291. https://doi.org/10.1128/JB.01544-12.
8. Makarova KS, Ponomarev VA, Koonin EV. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biol 2:e0033. https://doi.org/10.1186/gb-2001-2-9-research0033.
9. Panina EM, Mironov AA, Gelfand MS. 2003. Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proc Natl Acad Sci U S A 100:9912–9917. https://doi.org/10.1073/pnas.1733691100.
10. Ueta M, Wada C, Wada A. 2020. YkgM and YkgO maintain translation by replacing their paralogs, zinc-binding ribosomal proteins L31 and L36, with identical activities. Genes Cells 25:562–581. https://doi.org/10.1111/gtc.12796.
11. Nanamiya H, Akanuma G, Natori Y, Murayama R, Kosono S, Kudo T, Kobayashi K, Ogasawara N, Park SM, Ochi K, Kawamura F. 2004. Zinc is a key factor in controlling alternation of two types of L31 protein in the *Bacillus subtilis* ribosome. Mol Microbiol 52:273–283. https://doi.org/10.1111/j.1365-2958.2003.03972.x.
12. Natori Y, Nanamiya H, Akanuma G, Kosono S, Kudo T, Ochi K, Kawamura F. 2007. A fail-safe system for the ribosome under zinc-limiting conditions in *Bacillus subtilis*. Mol Microbiol 63:294–307. https://doi.org/10.1111/j.1365-2958.2006.05513.x.
13. Akanuma G, Kobayashi A, Suzuki S, Kawamura F, Shiwa Y, Watanabe S, Yoshikawa H, Hanai R, Ishizuka M. 2014. Defect in the formation of 70S ribosomes caused by lack of ribosomal protein L34 can be suppressed by magnesium. J Bacteriol 196:3820–3830. https://doi.org/10.1128/JB.01896-14.
14. Akanuma G, Yamazaki K, Yagishi Y, Iizuka Y, Ishizuka M, Kawamura F, Kato-Yamada Y. 2018. Magnesium suppresses defects in the formation of 70S ribosomes as well as in sporulation caused by lack of several individual ribosomal proteins. J Bacteriol 200:e00212-18. https://doi.org/10.1128/JB.00212-18.
15. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, Zhang R. 2021. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res 49:D677–D686. https://doi.org/10.1093/nar/gkaa917.
16. Gurumayum S, Jiang P, Hao X, Campos TL, Young ND, Korhonen PK, Gasser RB, Bork P, Zhao XM, He LJ, Chen WH. 2021. OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. Nucleic Acids Res 49:D998–D1003. https://doi.org/10.1093/nar/gkaa884.
17. Ji Y, Zhang B, Van SF, Horn Warren P, Woodnutt G, Burnham MK, Rosenberg M. 2001. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science 293:2266–2269. https://doi.org/10.1126/science.1063566.
18. Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, Wall D, Wang L, Brown-Driver V, Froelich JM, C KG, King P, McCarthy M, Malone C, Misiner B, Robbins D, Tan Z, Zhu Zy ZY, Carr G, Mosca DA, Zamudio C, Foulkes JG, Zyskind JW. 2002. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. Mol Microbiol 43:1387–1400. https://doi.org/10.1046/j.1365-2958.2002.02832.x.
19. Yin D, Ji Y. 2002. Genomic analysis using conditional phenotypes generated by antisense RNA. Curr Opin Microbiol 5:330–333. https://doi.org/10.1016/s1369-5274(02)00315-6.
20. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, III, Smith HO, Venter JC. 2006. Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A 103:425–430. https://doi.org/10.1073/pnas.0510013103.
21. French CT, Lao P, Loraine AE, Matthews BT, Yu H, Dybvig K. 2008. Large-scale transposon mutagenesis of *Mycoplasma pulmonis*. Mol Microbiol 69:67–76. https://doi.org/10.1111/j.1365-2958.2008.06262.x.
22. Dybvig K, Lao P, Jordan DS, Simmons WL. 2010. Fewer essential genes in mycoplasmas than previous studies suggest. FEMS Microbiol Lett 311:51–55. https://doi.org/10.1111/j.1574-6968.2010.02078.x.
23. McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26. https://doi.org/10.1038/nrmicro2670.
24. Lluch-Senar M, Delgado J, Chen WH, Llorens-Rico V, O'Reilly FJ, Wodke JA, Unal EB, Yus E, Martinez S, Nichols RJ, Ferrar T, Vivancos A, Schmeisky A, Stülke J, van Noort V, Gavin AC, Bork P, Serrano L. 2015. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol Syst Biol 11:780. https://doi.org/10.15252/msb.20145558.
25. Glass JI, Merryman C, Wise KS, Hutchison CA, III, Smith HO. 2017. Minimal cells—real and imagined. Cold Spring Harb Perspect Biol 9:a023861. https://doi.org/10.1101/cshperspect.a023861.
26. Breuer M, Earnest TM, Merryman C, Wise KS, Sun L, Lynott MR, Hutchison CA, Smith HO, Lapek JD, Gonzalez DJ, de Crecy-Lagard V, Haas D, Hanson AD, Labhsetwar P, Glass JI, Luthey-Schulten Z. 2019. Essential metabolism for a minimal cell. Elife 8:e36842. https://doi.org/10.7554/eLife.36842.
27. Nikolaeva DD, Gelfand MS, Garushyants SK. 2021. Simplification of ribosomes in bacteria with tiny genomes. Mol Biol Evol 38:58–66. https://doi.org/10.1093/molbev/msaa184.
28. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287. https://doi.org/10.1126/science.1123061.
29. Fournier GP, Gogarten JP. 2010. Rooting the ribosomal tree of life. Mol Biol Evol 27:1792–1801. https://doi.org/10.1093/molbev/msq057.
30. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018.

31. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637. https://doi.org/10.1126/science.278.5338.631.
32. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. https://doi.org/10.1093/nar/28.1.33.
33. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–D269. https://doi.org/10.1093/nar/gku1223.
34. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28. https://doi.org/10.1093/nar/29.1.22.
35. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics 26:1481–1487. https://doi.org/10.1093/bioinformatics/btq229.
36. Natale DA, Galperin MY, Tatusov RL, Koonin EV. 2000. Using the COG database to improve gene recognition in complete genomes. Genetica 108:9–17. https://doi.org/10.1023/a:1004031323748.
37. Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. 2019. Microbial genome analysis: the COG approach. Brief Bioinform 20:1063–1070. https://doi.org/10.1093/bib/bbx117.
38. Koonin EV, Galperin MY. 2003. Sequence-evolution-function: computational approaches in comparative genomics. Kluwer Academic, Boston, MA.
39. Choli T, Franceschi F, Yonath A, Wittmann-Liebold B. 1993. Isolation and characterization of a new ribosomal protein from the thermophilic eubacteria, *Thermus thermophilus*, *T. aquaticus* and *T. flavus*. Biol Chem Hoppe Seyler 374:377–383. https://doi.org/10.1515/bchm3.1993.374.1-6.377.
40. Izutsu K, Wada C, Komine Y, Sako T, Ueguchi C, Nakura S, Wada A. 2001. *Escherichia coli* ribosome-associated protein SRA, whose copy number increases during stationary phase. J Bacteriol 183:2765–2773. https://doi.org/10.1128/JB.183.9.2765-2773.2001.
41. Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T. 2011. Solving the problem: genome annotation standards before the data deluge. Stand Genomic Sci 5:168–193. https://doi.org/10.4056/sigs.2084864.
42. Farabaugh PJ. 1996. Programmed translational frameshifting. Annu Rev Genet 30:507–528. https://doi.org/10.1146/annurev.genet.30.1.507.
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
44. Wurch L, Giannone RJ, Belisle BS, Swift C, Utturkar S, Hettich RL, Reysenbach AL, Podar M. 2016. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. Nat Commun 7:12115. https://doi.org/10.1038/ncomms12115.
45. Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citti C, Barre A, Yoshizawa S, Fourmy D, de Crecy-Lagard V, Blanchard A. 2014. Predicting the minimal translation apparatus: lessons from the reductive evolution of mollicutes. PLoS Genet 10:e1004363. https://doi.org/10.1371/journal.pgen.1004363.
46. McCutcheon JP, von Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol 21:1366–1372. https://doi.org/10.1016/j.cub.2011.06.051.
47. Miller IJ, Weyna TR, Fong SS, Lim-Fong GE, Kwan JC. 2016. Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome. Sci Rep 6:34362. https://doi.org/10.1038/srep34362.
48. Cerretti DP, Dean D, Davis GR, Bedwell DM, Nomura M. 1983. The *spc* ribosomal protein operon of *Escherichia coli*: sequence and cotranscription of the ribosomal protein genes and a protein export gene. Nucleic Acids Res 11:2599–2616. https://doi.org/10.1093/nar/11.9.2599.
49. Koonin EV, Galperin MY. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr Opin Genet Dev 7:757–763. https://doi.org/10.1016/s0959-437x(97)80037-8.
50. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 286:2165–2169. https://doi.org/10.1126/science.286.5447.2165.
51. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, Knight R, Gordon JI. 2009. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 6:279–289. https://doi.org/10.1016/j.chom.2009.08.003.
52. Veeranagouda Y, Husain F, Tenorio EL, Wexler HM. 2014. Identification of genes required for the survival of *B. fragilis* using massive parallel sequencing of a saturated transposon mutant library. BMC Genomics 15:429. https://doi.org/10.1186/1471-2164-15-429.
53. Commichau FM, Pietack N, Stülke J. 2013. Essential genes in *Bacillus subtilis*: a re-evaluation after ten years. Mol Biosyst 9:1068–1075. https://doi.org/10.1039/c3mb25595f.
54. Shoji S, Dambacher CM, Shajani Z, Williamson JR, Schultz PG. 2011. Systematic chromosomal deletion of bacterial ribosomal protein genes. J Mol Biol 413:751–761. https://doi.org/10.1016/j.jmb.2011.09.004.
55. Cate JH, Yusupov MM, Yusupova GZ, Earnest TN, Noller HF. 1999. X-ray crystal structures of 70S ribosome functional complexes. Science 285:2095–2104. https://doi.org/10.1126/science.285.5436.2095.
56. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905–920. https://doi.org/10.1126/science.289.5481.905.
57. Wimberly BT, Brodersen DE, Clemons WM, Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit. Nature 407:327–339. https://doi.org/10.1038/35030006.
58. Greber BJ, Boehringer D, Godinic-Mikulcic V, Crnkovic A, Ibba M, Weygand-Durasevic I, Ban N. 2012. Cryo-EM structure of the archaeal 50S ribosomal subunit in complex with initiation factor 6 and implications for ribosome evolution. J Mol Biol 418:145–160. https://doi.org/10.1016/j.jmb.2012.01.018.
59. Armache JP, Anger AM, Marquez V, Franckenberg S, Frohlich T, Villa E, Berninghausen O, Thomm M, Arnold GJ, Beckmann R, Wilson DN. 2013. Promiscuous behaviour of archaeal ribosomal proteins: implications for eukaryotic ribosome evolution. Nucleic Acids Res 41:1284–1293. https://doi.org/10.1093/nar/gks1259.
60. Watson ZL, Ward FR, Meheust R, Ad O, Schepartz A, Banfield JF, Cate JH. 2020. Structure of the bacterial ribosome at 2 Å resolution. Elife 9:e60482. https://doi.org/10.7554/eLife.60482.
61. Held WA, Ballou B, Mizushima S, Nomura M. 1974. Assembly mapping of 30 S ribosomal proteins from *Escherichia coli*. Further studies. J Biol Chem 249:3103–3111. https://doi.org/10.1016/S0021-9258(19)42644-6.
62. Herold M, Nierhaus KH. 1987. Incorporation of six additional proteins to complete the assembly map of the 50 S subunit from *Escherichia coli* ribosomes. J Biol Chem 262:8826–8833. https://doi.org/10.1016/S0021-9258(18)47489-3.
63. Nierhaus KH. 1991. The assembly of prokaryotic ribosomes. Biochimie 73:739–755. https://doi.org/10.1016/0300-9084(91)90054-5.
64. Shajani Z, Sykes MT, Williamson JR. 2011. Assembly of bacterial ribosomes. Annu Rev Biochem 80:501–526. https://doi.org/10.1146/annurev-biochem-062608-160432.
65. Chen SS, Williamson JR. 2013. Characterization of the ribosome biogenesis landscape in *E. coli* using quantitative mass spectrometry. J Mol Biol 425:767–779. https://doi.org/10.1016/j.jmb.2012.11.040.
66. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229.
67. Hsiao C, Mohan S, Kalahar BK, Williams LD. 2009. Peeling the onion: ribosomes are ancient molecular fossils. Mol Biol Evol 26:2415–2425. https://doi.org/10.1093/molbev/msp163.
68. Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud NV, Williams LD. 2015. History of the ribosome and the origin of translation. Proc Natl Acad Sci U S A 112:15396–15401. https://doi.org/10.1073/pnas.1509761112.
69. Maier UG, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF, Martin WF. 2013. Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. Genome Biol Evol 5:2318–2329. https://doi.org/10.1093/gbe/evt181.
70. Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima SY. 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. Curr Biol 24:R640–R641. https://doi.org/10.1016/j.cub.2014.06.038.
71. Yutin N, Galperin MY. 2013. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. Environ Microbiol 15:2631–2641. https://doi.org/10.1111/1462-2920.12173.

72. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. https://doi.org/10.1038/nature14486.
73. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, III, Smith HO, Venter JC. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52–56. https://doi.org/10.1126/science.1190719.
74. Hutchison CA, III, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi ZQ, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. 2016. Design and synthesis of a minimal bacterial genome. Science 351:aad6253. https://doi.org/10.1126/science.aad6253.
75. Venetz JE, Del Medico L, Wolfle A, Schächle P, Bucher Y, Appert D, Tschan F, Flores-Tinoco CE, van Kooten M, Guennoun R, Deutsch S, Christen M, Christen B. 2019. Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. Proc Natl Acad Sci U S A 116:8070–8079. https://doi.org/10.1073/pnas.1818259116.
76. Reuß DR, Commichau FM, Gundlach J, Zhu B, Stülke J. 2016. The blueprint of a minimal cell: MiniBacillus. Microbiol Mol Biol Rev 80:955–987. https://doi.org/10.1128/MMBR.00029-16.
77. Reuß DR, Altenbuchner J, Mäder U, Rath H, Ischebeck T, Sappa PK, Thürmer A, Guérin C, Nicolas P, Steil L, Zhu B, Feussner I, Klumpp S, Daniel R, Commichau FM, Völker U, Stülke J. 2017. Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res 27:289–299. https://doi.org/10.1101/gr.215293.116.
78. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, Gillet R, Forterre P, Krupovic M. 2019. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun 10:752. https://doi.org/10.1038/s41467-019-08672-6.
79. Baird NJ, Zhang J, Hamma T, Ferre-D'Amare AR. 2012. YbxF and YlxQ are bacterial homologs of L7Ae and bind K-turns but not K-loops. RNA 18:759–770. https://doi.org/10.1261/rna.031518.111.
80. Makarova KS, Wolf YI, Koonin EV. 2015. Archaeal Clusters of Orthologous Genes (arCOGs): an update and application for analysis of shared features between *Thermococcales*, *Methanococcales*, and *Methanobacteriales*. Life (Basel) 5:818–840. https://doi.org/10.3390/life5010818.
81. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020: baaa062. https://doi.org/10.1093/database/baaa062.
82. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48: D265–D268. https://doi.org/10.1093/nar/gkz991.
83. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.
84. Gupta RS, Sawnani S, Adeolu M, Alnajar S, Oren A. 2018. Phylogenetic framework for the phylum Tenericutes based on genome sequence data: proposal for the creation of a new order *Mycoplasmoidales* ord. nov., containing two new families *Mycoplasmoidaceae* fam. nov. and *Metamycoplasmataceae* fam. nov. harbouring *Eperythrozoon*, *Ureaplasma* and five novel genera. Antonie Van Leeuwenhoek 111:1583–1630. https://doi.org/10.1007/s10482-018-1047-3.