Simplification of ribosomes in bacteria with tiny genomes

Daria D. Nikolaeva[1,3], Mikhail S. Gelfand[2,3,4], Sofya K. Garushyants[2*]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye Gory 1-73, Moscow, 119234, Russia;

[2] Skolkovo Institute of Science and Technology, Nobelya Ulitsa 3, Moscow, 121205, Russia;

[3] Institute for Information Transmission Problems (Kharkevich Institute), Bolshoy Karetny 19-1, Moscow, 127051, Russia;

[4] Faculty of Computer Science, National Research University Higher School of Economics, Kochnovsky pr. 3, Moscow, 125319, Russia.

* Correspondence: Sofya K. Garushyants garushyants@iitp.ru

Keywords: ribosome; bacteria; evolution; genome reduction; rRNA; ribosomal protein

Abstract

The ribosome is an essential cellular machine performing protein biosynthesis. Its structure and composition are highly conserved in all species. However, some bacteria have been reported to have an incomplete set of ribosomal proteins. We have analyzed ribosomal protein composition in 214 small bacterial genomes (< 1 Mb) and found that although the ribosome composition is fairly stable, some ribosomal proteins may be absent, especially in bacteria with dramatically reduced genomes. The protein composition of the large subunit is less conserved than that of the small subunit. We have identified the set of frequently lost ribosomal proteins and demonstrated that they tend to be situated on the ribosome surface and have fewer contacts to other ribosome components. Moreover, loss of some proteins is correlated, at least in one case leading to the loss of a complete functional block. Additionally, the reduction of rRNA is also common in bacteria with short genomes, and deletions mostly occur in free loops. The loss of ribosomal protein uL24 occurs simultaneously with the loss of the corresponding binding site in rRNA. Finally, the loss of the anti-Shine-Dalgarno sequence is not correlated with the genome size or the loss of particular ribosomal proteins.

Introduction

The ribosome is a universal biosynthesis machinery present in all eukaryotes and prokaryotes. A bacterial ribosome is comprised of the small (30S) and large (50S) subunits which together form the 70S particle (Kurland 1972; Ramakrishnan 2002). The bacterial ribosome consists of multiple proteins (RP, r-proteins) and three rRNA molecules – 16S in the small subunit, 23S and 5S in the large subunit (Kurland 1972). The main catalytic functions of the ribosome, such as the peptide bond formation, mRNA decoding, and translocation of mRNA and tRNA after the peptide bond formation, are performed by ribosomal RNA (rRNA) (Green & Noller 1997; Nissen et al. 2000; Schmeing et al. 2003). Moreover, rRNA molecules also determine the ribosomal spatial organization providing sites for binding of the ribosomal proteins (r-proteins) (Khaitovich et al. 1999). The ribosome of *Escherichia coli* contains 21 proteins in the 30S subunit (bS1–bS21) and 33 proteins in the 50S subunit (uL1–bL36) (r-protein nomenclature is according to Ban et al. 2014) (Schuwirth et al. 2005; Kaczanowska & Rydén-Aulin 2007). The role of r-proteins is to stabilize the

molecular machine and to regulate the ribosomal activity (Aseev and Boni 2011). While the key role in the protein biosynthesis is played by rRNA, the r-protein composition also tends to be conserved in most bacteria (Roberts et al. 2008). Moreover, 33 r-proteins are conserved among different domains of life (Roberts et al. 2008; Smith et al. 2008; Lecompte et al. 2002).

The protein composition of bacterial ribosomes has been studied intensively. The analysis of single-gene knockout mutants of *E. coli* has showed that these bacteria are able to grow in the absence of proteins bS6, uS15, bS20, bS21, uL1, bL9, uL11, and bL25, but in most cases the growth rate of these knockout mutants was reduced (Baba et al. 2006). In another study of an *E. coli* knockout collection, nine r-proteins (uL15, bL21, uL24, bL27, uL29, uL30, bL34, and uS17) have been found to be nonessential for survival in experimental conditions (Shoji et al. 2011). A similar study in *Bacillus subtilis* has identified 20 r-proteins non-essential to growth in experimental conditions (Akanuma et al. 2012). Based on this result, the smaller r-proteins were proposed to have been incorporated into the ribosome relatively recently, and hence to be less essential. A phylogenetic analysis of 995 completely sequenced bacterial genomes shows that 44 r-proteins are strictly ubiquitous, proteins bS16, bL9, bL19, bL31, bL34, and bL36 are rarely missing, while bS21, S22, bThx, bL25, and uL30 are absent in a large fraction of bacteria (Yutin et al. 2012). A comparative analysis of the translation apparatus of *Mollicutes* has shown that five r-proteins (uS14, S22, bL7, bL25, and bL31) are missing in all studied genomes, uL30 is missing in almost all of them, and bS1 has been lost in seven independent events in different clades (Grosjean et al. 2014). Finally, an analysis of endosymbiotic bacteria with small genomes from a variety of phyla has demonstrated that they are lacking the largest fraction of r-proteins, as only 17 of 21 small-subunit and 16 of 32 large-subunit r-proteins are universally present in these bacteria (McCutcheon and Moran 2012). For example, *Candidatus* Tremblaya princeps, an endosymbiont of mealybugs (*Pseudococcidae* family), has only 139 genes suggesting little possibility to synthesize the complete translation apparatus (McCutcheon and Moran 2012). While the list of essential r-proteins is slightly different between these analyses, bL25, uL30, bL31, bS21, S22, and bThx have been reported to be the least essential.

Less is known about whether there are any universal features common to these non-essential proteins, and how the r-protein loss is linked to the genome reduction. Here, we analyzed the r-protein composition in bacteria with small genomes from a variety of phyla.

We have identified a set of frequently lost proteins and found two patterns of correlative r-protein loss. A majority of frequently lost proteins have been shown to be located on the ribosome surface and to form less contacts with other proteins, compared to preserved r-proteins. The analysis of deletions in 16S and 23S rRNA has yielded a deletion of the r-protein uL24 binding site coinciding with the loss of the protein itself. We also show that the loss of anti-Shine-Dalgarno sequence is not correlated with the loss of particular ribosomal proteins or with the bacterial genome size.

Results

Some ribosomal proteins are missing in bacteria with small genomes

We re-annotated all r-proteins in 214 bacterial strains with small genomes (less than 1 Mbp) from 38 genera of the following phyla: *Proteobacteria*, *Bacteroidetes*, *Spirochaetes*, *Tenericutes*, and *Actinobacteria* (Table S1, Materials and Methods). For that, we used 65 Pfam domains: 63 domains for the canonical set of ribosomal proteins, and two domains for the trigger-factor protein (TF), a ribosome-associated chaperone (Table S2).

Our results (Fig. S1) show that two proteins are absent in almost all considered strains, S22 (all strains) and bThx (present in *Candidatus* Walczuchella monophlebidarum only).

All except 11 r-proteins are lost in at least one strain from our dataset. At that, proteins bL9, uL24, bL25, uL29, uL30, bL32, bL34, bL36, bS21, and TF are lost frequently, as they are absent in at least 19 strains from a variety of phyla. R-proteins of the small subunit are more likely to be retained than the large-subunit r-proteins (Fig. S1). The only frequently lost small-subunit protein bS21 is absent in 54 strains from seven genera, while the most frequently lost protein from the large subunit, uL30, is absent in 138 genomes from 18 genera. All frequently lost proteins had been lost independently multiple times in bacteria (Fig. 1, Fig. S2). The largest number of r-protein losses was observed in *Candidatus* Tremblaya princeps, *Candidatus* Hodgkinia cicadocola, and *Candidatus* Carsonella rudii, bacteria with the shortest genomes in our dataset.

R-protein loss depends on the level of genome reduction

The comparison of the r-protein composition and the genome size revealed a

correlation between the genome size and the number of retained ribosomal proteins ($r^2 = 0.6$, $p < 2.2\times10^{-16}$), but the slope is dramatically different for the tiny genomes (Fig 2a). Indeed, for genomes shorter than 350 Kb there is an even stronger correlation between the genome size and the number of r-proteins ($r^2 = 0.7$, $p = 5.7\times10^{-6}$), while the remaining genomes in the dataset show no correlation at all ($p = 0.53$). This pattern holds both for the complete ribosome and the individual subunits (Fig. 2bc), the slope being steeper for the large subunit.

Patterns of ribosomal protein loss

Ten most frequently missing proteins had been lost independently in different phyla, showing parallel events (Fig. 1, Fig. S2). To analyze possible dependencies between frequent losses, we calculated the Jaccard similarity coefficient between phyletic profiles, that is, vectors of the protein presence/absence in all strains (Fig. 3). Two clusters were observed, uL24+uL29+bL9+TF and bL25+uL30+bS21. These patterns were also found when using the Pearson correlation coefficient between vectors of the protein presence/absence in all strains (Fig. S3). As all frequently lost proteins had been lost in bacteria from different phyla, the observed patterns could not be explained by the phylogenetic structure of the dataset.

The observed patterns are consistent with positions of the r-proteins in the ribosome. Indeed, bL25 and uL30 are positioned near each other (Fig. 4b), consistent with their tendency to be lost cooperatively. The correlated loss of proteins uL24, uL29, and TF can be explained by the TF position in the ribosomal structure, as the latter is located on the ribosome surface where it interacts with r-proteins uL23, uL24, and uL29 (Fig. 4c). The absence of TF leads to the loss of the r-proteins it is connected to.

A more general tendency is that frequently lost proteins as a rule are situated at the ribosome surface (Fig. 4abd). However, surprisingly, proteins bL34 and bL36 are deeply immersed within the structure (Fig. 4bd).

Frequently lost r-proteins have less contacts than preserved proteins

Further, we studied what features differ between frequently lost and preserved r-proteins. For that we fitted a logistic regression model to our data, regressing the protein status (either preserved or frequently lost) on the evolutionary rate, protein length, and the number of contacts with all adjacent proteins and rRNA. Only the number of contacts is a significant determinant of the protein status ($p = 0.022$).

Additionally, we analyzed the differences in the evolutionary rate (Fig. 5a) and the

number of contacts (Fig. 5b) between preserved and frequently lost proteins. Indeed, frequently lost proteins have significantly less contacts than preserved proteins (means 231 and 478, respectively, $p = 1.4\times10^{-4}$), as the former are mainly situated on the ribosome surface.

Surprisingly, there was no significant difference in evolutionary rate between preserved and frequently lost proteins. Moreover, the evolutionary rate of a protein is slightly correlated with the protein length ($r^2 = 0.3877335$, $p = 0.05548$), and short proteins evolve slower (Fig. 5c).

Deletions in rRNAs and the loss of r-proteins

To study possible correlations between the loss of r-proteins and their rRNA binding sites, we analyzed 16S and 23S rRNA multiple alignments and identified nine deletions (four in 23S and five in 16S) that had occurred in more than two genera and were longer than ten nucleotides in at least one species. Of these, one deletion coincided with an r-protein binding site (uL24) and occurred only in strains which had lost the uL24. All other deletions affect rRNA free loops not involved in the r-protein binding.

Anti-SD loss is frequent among strains from different taxa

Analysis of the anti-Shine-Dalgarno sequence in 16S rRNA demonstrated frequent loss that, however, showed little correlation with the number of lost r-proteins ($r^2 = 0.28$, $p = 9.4\times10^{-5}$) or genome reduction ($r^2 = 0.24$, $p = 8.3\times10^{-4}$) (Fig. 6a). There is also no correlation between the loss of the anti-SD sequence and any of frequently lost r-protein (Fig. 6b). The anti-SD presence/absence pattern is not taxonomy-specific, as each phylum has strains with and without the anti-SD in 16S rRNA (Fig. 6a).

Discussion

The bacterial genome size is evolutionarily labile, and radical genome reduction has occurred multiple times independently throughout the bacterial domain. Availability of multiple independent genome reduction events allows for a systematic study of the patterns of gene loss associated with them.

While individual r-proteins can be lost in bacteria with genomes of any size, for the tiniest of bacterial genomes, further reduction in the genome size is associated with the loss

of generally conserved ribosomal proteins. The fact that the ribosomal proteins are among the last to leave a shrinking genome illustrates that they tend to be less dispensable than most protein. Two r-proteins absent in the largest number of small-genome bacteria are S22 and bThx. Previous studies indicate that these proteins are largely non-essential. Indeed, the S22 gene is mostly expressed at the stationary growth phase, and its deletion does not affect the viability of *E. coli* mutants (Izutsu et al. 2001). bThx is found only in thermophilic bacteria and has been shown to stabilize the organization of RNA elements at the top of the 30S subunit head in *Thermus thermophilus* (Choli et al. 1993; Wimberly et al. 2000), which implies that this protein is only essential for survival at high temperatures. However, the low prevalence of these two proteins may indicate not the ease of their loss, but rather their late arrival. S22 is present only in some enterobacteria (Izutsu et al. 2001), and bThx homologs are only present in thermophilic bacteria from genus *Thermus* (Choli et al. 1993; Wimberly et al. 2000). Therefore, both S22 and bThx seem to have been acquired only by some bacterial groups after the last common ancestor of bacteria.

Our list of frequently lost proteins is largely consistent with previous observations (Baba et al. 2006; Shoji et al. 2011; Akanuma et al. 2012; Yutin et al. 2012; Grosjean et al. 2014). For example, among six non-ubiquitous r-proteins reported by Yutin et al. 2012, three (bL25, uL30, bS21) are among the r-proteins identified as frequently lost in our dataset, in addition to two proteins (S22 and bThx) that only occur in some bacterial lineages (see above). However, our focus on bacteria with tiny genomes allowed us to study less dispensable r-proteins. Proteins bL34 and bL36 identified previously as present in almost all genomes (Yutin et al. 2012) are frequently lost in our dataset. While these proteins are generally considered essential, they can be dispensable under certain circumstances. The absence of bL34 has been shown to affect cell growth, but the cell function can be restored by increasing the magnesium ion flow (Shoji et al. 2011; Akanuma et al. 2014). The absence of bL36 is only essential for cell growth at high temperatures (Ikegami et al. 2005). Moreover, bL36 has been shown to be missing in *Bacteroidetes* (Yutin et al. 2012). R-proteins uL24 and uL29 identified here as frequently lost had been lost in at least two independent events, and only in bacteria with drastically reduced genomes. The knockouts of both these proteins have shown to be viable (Shoji et al. 2011). Interestingly, r-proteins bL31 and uS14 have been identified as lost in all *Mollicutes* (Grosjean et al., 2014), but we have

observed that only few strains of *Mollicutes* lack these proteins. This difference is likely caused by a more sensitive annotation procedure used here.

What are the factors that affect the propensity of a protein to be lost? Previously, the proteins frequently lost in evolution have been shown to evolve rapidly, have fewer interactions with other proteins, and lower expression levels (Krylov et al. 2003). The expression levels of different r-proteins are naturally similar, so we have not considered them in our model. Here, we observe that only the number of contacts is a significant determinant of the r-protein loss, and not the substitution rate or protein length. The number of contacts is an indirect measure of the exposure of an r-protein on the ribosome surface. Indeed, r-proteins that are frequently lost tend to be located on the ribosome surface, the exceptions being bL34 and bL36 (Fig. 4).

The order in which r-proteins are assembled in the ribosome complex (Chen & Williamson 2013) provides indirect evidence of the order in which r-proteins have been incorporated in the ribosome in course of evolution. Thus, r-proteins that appear only in the late ribosome intermediates should be relatively young. Seven out of ten frequently lost proteins r-proteins – bS21, bL9, bL25, uL29, bL32, bL34 and bL36 – are included only in the late assembly intermediates. Thus, it can be speculated that frequently lost r-proteins had been incorporated in the ribosome late in evolution.

Another interesting question is whether r-proteins are lost at random, or there are patterns of ribosome simplification, where the loss of one protein leads to the loss of others. Our dataset contained many genomes with incomplete sets of r-proteins, that allowed us to study the patterns of common protein loss.

One such pattern is the loss of the entire block around the ribosome exit tunnel (TF, bL9, uL24, uL29). The trigger-factor (TF) is a chaperone, associated with the ribosome exit channel (Hoffman et al. 2010) and is the most frequently lost protein from the pattern. TF is surrounded by uL23, uL29, and uL24, directly interacting with uL23 and uL29 (Ferbitz et al. 2004). The role of bL9 in this pattern is less clear, because it does not interact directly with TF. This protein has been shown to reduce translation frameshifting (Dunkle et al. 2010), which means that its loss is also associated with the ribosome fidelity. The loss of this functional block is surprising, as even highly reduced genomes tend to keep chaperones (McCutcheon and Moran 2012).

One more pattern is the cooperative loss of three proteins (bS21, bL25 and uL30),

two of which, bL25 and uL30, are located close to each other on the ribosome surface, while bS21 is not close and even belongs to a different subunit. Multiple independent losses of bL25 and uL30, and even bS21, are supported by previous observations (Yutin et al. 2012; Lecompte et al. 2002). This pattern still remains unexplained, because uL30 is highly conserved in Archaea and Eukarya (Lecompte et al. 2002), where it is thought to be essential for the selenocysteine recognition (Chavatte et al. 2005). The role of this protein in bacteria is unclear. The loss of bL25 is tolerated in *E. coli*, but the mutant bacteria have a reduced growth rate (Baba et al. 2006). bL25 has been proposed to be essential for interaction with r-protein uL16, which is necessary for ribosome stability (Anikaev et al. 2016). bS21 is required for the recognition of native templates, and its function resembles the function of bS1 (Van Duin and Wijnands 1981). Thus, the loss of proteins in this cluster should reduce the ribosome effectiveness.

In conclusions, proteins are not lost at random, and there are at least two distinct patterns of loss of closely situated proteins, at least in one case constituting a complete functional block. Overall, frequently lost proteins have less contacts, are located on the ribosome surface, and have been incorporated in the ribosome late in evolution. That is consistent with loss not affecting the ribosome function, especially in symbiotic bacteria that live in stable host environments.

Materials and Methods

Dataset of small bacterial genomes

The list of all bacterial species with complete genomes not exceeding 1 Mbp was compiled from the IMG/M database (Chen et al. 2017). The genomic data for all strains of these species (214 genomes in total, Table S1) were downloaded from the NCBI FTP; files with protein sequences in February 2017 (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/), files with RNA sequences in September 2017 (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/).

The genomes of some strains in the selected species exceeded 1 Mb, but were still retained in the dataset.

Annotation of protein domains

Ribosomal proteins and trigger factor (TF) were re-annotated in the downloaded genomes (the complete list of proteins is given in Supplementary Table S2) using HMM matrices from the Pfam-A database (Finn et al. 2016). The HMM-profiles for the selected domains were scanned against each genome with the HMMER software (Mistry et al. 2013). A protein was considered to be present in a given genome if the corresponding domain had a hit in the HMMER search with E-value < 0.001; only the best such hit was retained for further analysis. If the HMMER domain bias had the same order of magnitude as the similarity score, additional filtering was performed as described below.

For each domain we created a multiple alignment of proteins with confident HMMer predictions. Protein alignments were built using the GUIDANCE2 Server (Penn et al. 2010) with the MAFFT program for building multiple sequence alignments. A candidate protein with a high domain bias value was added to the domain alignment and considered as present if it fitted the multiple sequence alignment keeping the GUIDANCE alignment confidence score above the recommended threshold (0.8). If the confidence score was below the threshold, the protein was removed from the dataset.

Proteins consisting of two domains (uL2, uL5, bL7/L12, bL9, uL11, uS4, uS5, TF) were considered to be present if both domains were found.

Identification of independent r-protein losses

To map the phyletic patterns of r-protein losses, we constructed the maximum likelihood tree of concatenated multiple alignment of all preserved r-proteins (see below) using PhyML v. 3.1 (Guindon et al. 2010) with the default parameters. Prior to the tree construction all columns containing gaps were removed from the concatenated alignment. This tree was used to calculate the number of independent losses of each protein.

Estimation of the evolutionary rate

The evolutionary rate of a protein was defined as the average branch length in the respective phylogenetic tree. The maximum likelihood phylogenetic tree was built for each ribosomal protein using PhyML with 100 bootstrap replicates (LG+G model was utilized with tree topology and branch lengths optimised ("-o tl" parameters) and with gamma parameter identified for each protein alignment using ProtTest (version 3.4.2) (Darriba et al.

2011)). The evolutionary rate for a protein was compared to the average rates for the set of preserved proteins (uL5, uL14, uL15, bL20, uS5, uS7, uS8, uS9, uS11, uS13, and uS17). For each studied protein, the corresponding rates for every preserved protein from the set were calculated in the subtree with the same set of species as in the tree of the studied protein. To compare the results for proteins with different phyletic profiles, we calculated the relative evolutionary score, i.e. the protein's evolutionary rate divided by the corresponding average rate for the set of preserved r-proteins. For those proteins that consist of two domains (see Table S3) the same gamma parameters and, subsequently, the same evolutionary rates were obtained, with an exception of ribosomal protein bL9 - for this protein we utilize an average evolutionary rate between the two evolutionary rates for domains.

Calculation of the number of contacts

The number of contacts was estimated by measuring pairwise distances between atoms in ribosome PDB structures (PDB ID: 5H5U and 2D3O) (Ma et al. 2016; Schlünzen et al. 2005). Atoms were defined as contacting if the distance between them was at most 5Å. All protein self-contacts were disregarded. When multiple atoms of a studied protein were connected with the same atom in the ribosome structure, only one such contact was retained. The pairwise distances were calculated with the PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC and script pairwise_dist.py. For each protein the total number of contacts was calculated.

rRNA analysis

16S and 23S rRNA alignments were built using SINA Alignment Service (Pruesse et al. 2012), and then all common gaps were removed.

A strain was defined as having the anti-Shine-Dalgarno sequence if there was an anti-SD consensus CCUCCU at the 3'-end of the strain's 16S rRNA.

Statistical analysis and data visualization

All statistical analyses were performed using R. We used the logistic regression (function glm, family=binomial()) to find determinants of r-proteins loss. Correlation coefficient was calculated in R with the cor() function. The Jaccard similarity index was calculated with the vegan package in R. The dependency between the total number of r-proteins and the genome length (Fig. 2) was built in R using the ggplot2 package. Curve

smoothing was performed using the “loess” method. Phylogenetic trees were visualised with the iTOL server (Letunic and Bork 2011).

Acknowledgements

The project was initiated by Sofya Sherstneva and Anna Ergemlidze at the Summer School of Molecular and Theoretical Biology for high school students, supported by the Zimin Foundation. We thank Arthur Zalevsky (FBB MSU) and Georgii Bazykin (IITP, Skoltech) for useful discussions and suggestions. This work was supported by the Russian Science Foundation (grant 18-14-00358).

References

Akanuma G et al. 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. doi: 10.1128/JB.01544-12.

Akanuma G et al. 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. doi: 10.1128/JB.01896-14.

Anikaev AY, Isaev AB, Korobeinikova AV, Garber MB, Gongadze GM. 2016. Role of protein L25 and its contact with protein L16 in maintaining the active state of *Escherichia coli* ribosomes *in vivo*. Biochemistry Moscow. 81:19–27. doi: 10.1134/S0006297916010028.

Aseev LV, Boni IV. 2011. Extraribosomal functions of bacterial ribosomal proteins. Mol Biol. 45:739. doi: 10.1134/S0026893311050025.

Baba T et al. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2:2006.0008. doi: 10.1038/msb4100050.

Ban, Nenad, et al. 2014. A new system for naming ribosomal proteins. Current opinion in structural biology 24: 165-169. doi: 10.1016/j.sbi.2014.01.002

Bokov K, Steinberg SV. 2009. A hierarchical model for evolution of 23S ribosomal RNA. Nature. 457:977–980. doi: 10.1038/nature07749.

Chavatte L, Brown BA, Driscoll DM. 2005. Ribosomal protein L30 is a component of the UGA-selenocysteine recoding machinery in eukaryotes. Nature Structural & Molecular

Biology. 12:408–416. doi: 10.1038/nsmb922.

Chen I-MA et al. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 45:D507–D516. doi: 10.1093/nar/gkw929.

Chen SS, Williamson JR. 2013. Characterization of the Ribosome Biogenesis Landscape in E. coli Using Quantitative Mass Spectrometry. Journal of Molecular Biology. 425:767–779. doi: 10.1016/j.jmb.2012.11.040.

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*. Biological Chemistry Hoppe-Seyler. 374:377–384. doi: 10.1515/bchm3.1993.374.1-6.377.

Darriba, Diego, et al. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27.8: 1164-1165. doi: 10.1093/bioinformatics/btr088

Dunkle JA, Xiong L, Mankin AS, Cate JHD. 2010. Structures of the Escherichia coli ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. PNAS. 107:17152–17157. doi: 10.1073/pnas.1007988107.

Ferbitz L et al. 2004. Trigger factor in complex with the ribosome forms a molecular cradle for nascent proteins. Nature. 431:590–596. doi: 10.1038/nature02899.

Finn RD et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285. doi: 10.1093/nar/gkv1344.

Hoffmann A, Bukau B, Kramer G. 2010. Structure and function of the molecular chaperone Trigger Factor. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 1803:650–661. doi: 10.1016/j.bbamcr.2010.01.017.

Green R, Noller HF. 1997. Ribosomes and Translation. Annu. Rev. Biochem. 66:679–716. doi: 10.1146/annurev.biochem.66.1.679.

Grosjean H et al. 2014. Predicting the Minimal Translation Apparatus: Lessons from the Reductive Evolution of Mollicutes Gojobori, T, editor. PLoS Genetics. 10:e1004363. doi: 10.1371/journal.pgen.1004363.

Guindon S et al. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 59:307–321. doi: 10.1093/sysbio/syq010.

Ikegami A, Nishiyama K, Matsuyama S, Tokuda H. 2005. Disruption of rpmJ Encoding Ribosomal Protein L36 Decreases the Expression of secY Upstream of the spc Operon

and Inhibits Protein Translocation in Escherichia coli. Bioscience, Biotechnology, and Biochemistry. 69:1595–1602. doi: 10.1271/bbb.69.1595.

Izutsu K et al. 2001. Escherichia coli ribosome-associated protein SRA, whose copy number increases during stationary phase. J. Bacteriol. 183:2765–2773. doi: 10.1128/JB.183.9.2765-2773.2001.

Kaczanowska M, Rydén-Aulin M. 2007. Ribosome Biogenesis and the Translation Process in Escherichia coli. Microbiol. Mol. Biol. Rev. 71:477–494. doi: 10.1128/MMBR.00013-07.

Khaitovich P, Mankin AS, Green R, Lancaster L, Noller HF. 1999. Characterization of functionally active subribosomal particles from Thermus aquaticus. PNAS. 96:85–90. doi: 10.1073/pnas.96.1.85.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. Genome Res. 13:2229–2235. doi: 10.1101/gr.1589103.

Kurland CG. 1972. Structure and Function of the Bacterial Ribosome. Annual Review of Biochemistry. 41:377–408. doi: 10.1146/annurev.bi.41.070172.002113.

Lecompte O, Ripp R, Thierry J-C, 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.

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucl. Acids Res. 39: W475–W478. doi: 10.1093/nar/gkr201.

Ma C et al. 2017. Mechanistic insights into the alternative translation termination by ArfA and RF2. Nature. 541:550–553. doi: 10.1038/nature20822.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology. 10:13–26. doi: 10.1038/nrmicro2670.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41:e121–e121. doi: 10.1093/nar/gkt263.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. Science. 289:920–930. doi: 10.1126/science.289.5481.920.

Penn O et al. 2010. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Research. 38:W23–W28. doi: 10.1093/nar/gkq443.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 28:1823–1829. doi: 10.1093/bioinformatics/bts252.

Ramakrishnan V. 2002. Ribosome Structure and the Mechanism of Translation. Cell. 108:557–572. doi: 10.1016/S0092-8674(02)00619-0.

Roberts E, Sethi A, Montoya J, Woese CR, Luthey-Schulten Z. 2008. Molecular signatures of ribosomal evolution. PNAS. 105:13953–13958. doi: 10.1073/pnas.0804861105.

Schlünzen F et al. 2005. The Binding Mode of the Trigger Factor on the Ribosome: Implications for Protein Folding and SRP Interaction. Structure. 13:1685–1694. doi: 10.1016/j.str.2005.08.007.

Schmeing TM, Moore PB, Steitz TA. 2003. Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit. RNA. 9:1345–1352. doi: 10.1261/rna.5120503.

Schuwirth BS et al. 2005. Structures of the Bacterial Ribosome at 3.5 Å Resolution. Science. 310:827–834. doi: 10.1126/science.1117230.

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. doi: 10.1016/j.jmb.2011.09.004.

Smith TF, Lee JC, Gutell RR, Hartman H. 2008. The origin and evolution of the ribosome. Biol Direct. 3:16. doi: 10.1186/1745-6150-3-16.

Van Duin J, Wijnands R. 1981. The Function of Ribosomal Protein S21 in Protein Synthesis. European Journal of Biochemistry. 118:615–619. doi: 10.1111/j.1432-1033.1981.tb05563.x.

Wimberly BT et al. 2000. Structure of the 30S ribosomal subunit. Nature. 407:327–339. doi: 10.1038/35030006.

Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of Prokaryotic Ribosomal Proteins. PLOS ONE. 7:e36972. doi: 10.1371/journal.pone.0036972.

Figures

Figure 1. Maximum likelihood phylogenetic tree of analyzed bacterial species, constructed by concatenated alignment of preserved ribosomal proteins. The tree was constructed by PhyML with 100 bootstrap replicates. In this representation branch lengths are ignored (the tree with branch lengths is provided as Fig. S2). Bootstrap values in the range 0.9–1 are shown by blue circles. The absence of one of ten frequently lost proteins (bL9, uL24, bL25, uL29, uL30, bL32, bL34, bL36, bS21, TF) in a strain is marked by a colored circle. Leaves are colored by the phyla: *Actinobacteria* – black, *Bacteroidetes* – yellow, *Proteobacteria* – blue, *Spirochaetes* – green, *Tenericutes* – red, unclassified bacteria – grey.

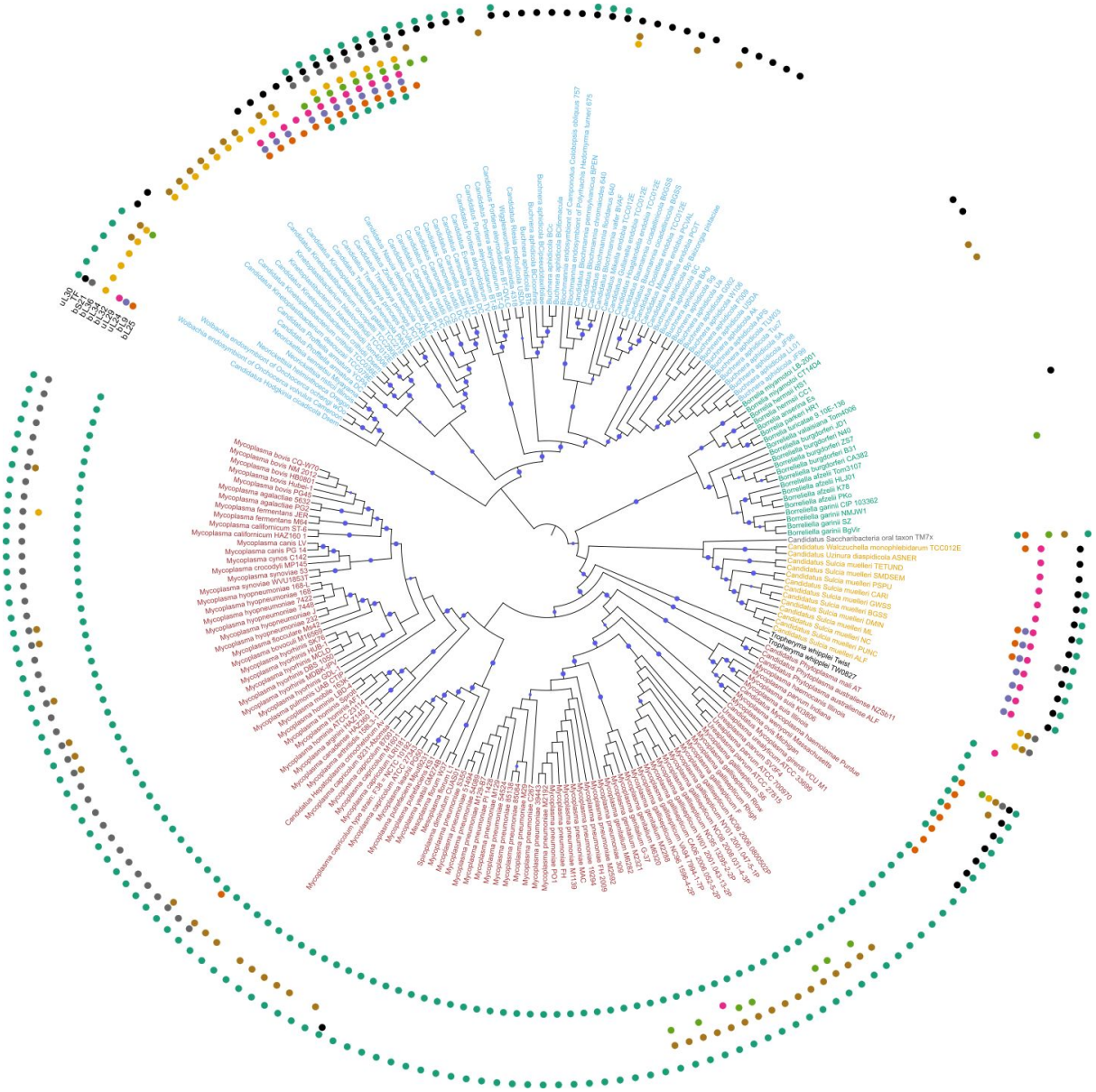

Figure 2. Dependencies of the number of ribosomal proteins (RP) on the genome length (in bp): (a) in the complete ribosome, (b) in the 50S subunit, and (c) in the 30S subunit. Each dot represents a bacterial genome, the dots are colored by phyla listed in the legend. The trend line is generated by the loess regression.

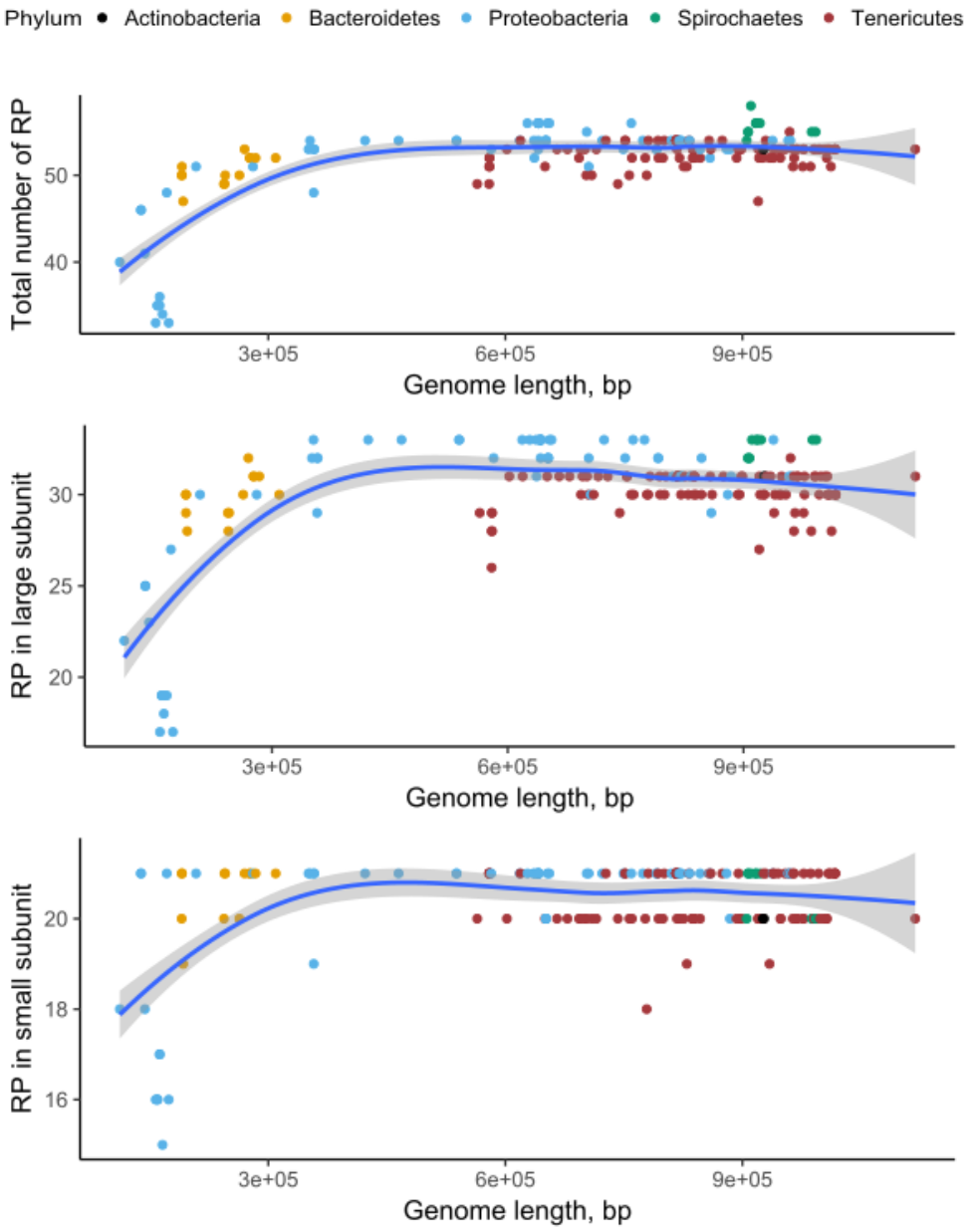

Figure 3. The Jaccard similarity coefficient between vectors of ribosomal protein presence/absence in bacterial strains. The color spectrum from red to blue corresponds to the similarity range from 0 to 1. Only ten frequently lost proteins are considered.

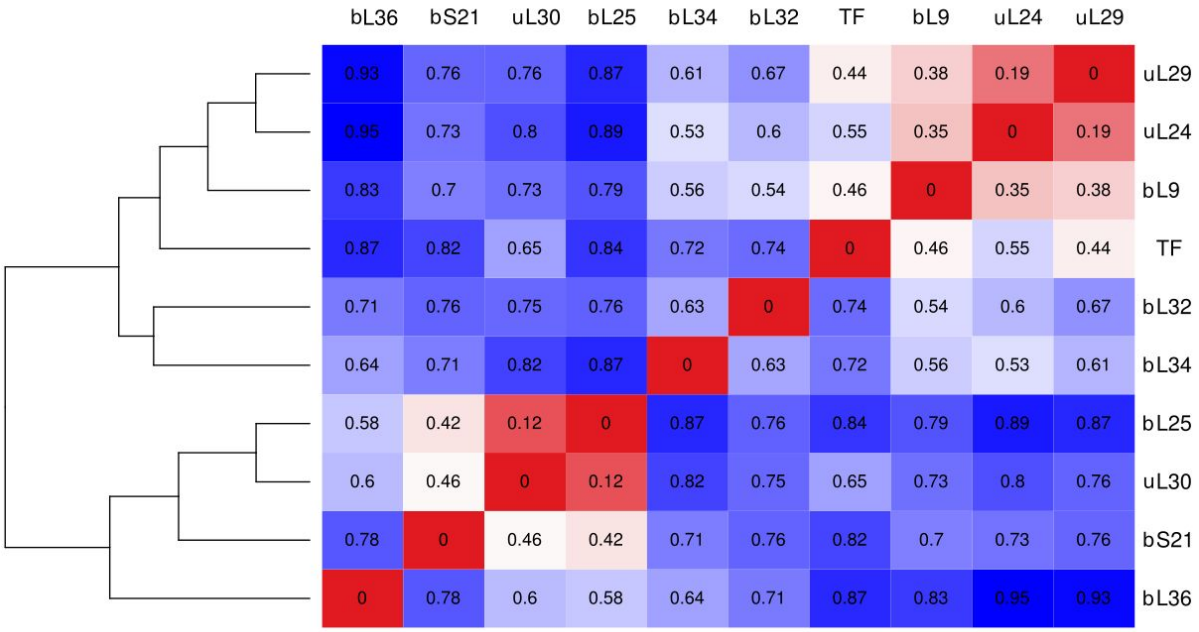

Figure 4. Ribosome. Proteins of interest are colored. Protein labels are of the same color as the respective proteins in the structure. (a), (b), (d) PDB ID: 5H5U (*Escherichia coli*). (c) PDB ID: 2D3O (*Deinococcus radiodurans*). (a) Crown view of the 70S ribosome to help understand the position of frequently lost ribosomal proteins labeled as "(lost)". The typical crown view position is marked using several ribosomal proteins (magenta), the position of central protuberance and 5S rRNA (light green) and the position of 16S rRNA (blue) and small subunit. In (b), (c), (d) only frequently lost ribosomal proteins are labeled. (b), (d) Various perspectives of the ribosome showing the position of frequently lost proteins. (c) 23S rRNA and frequently lost proteins which are connected with trigger factor.

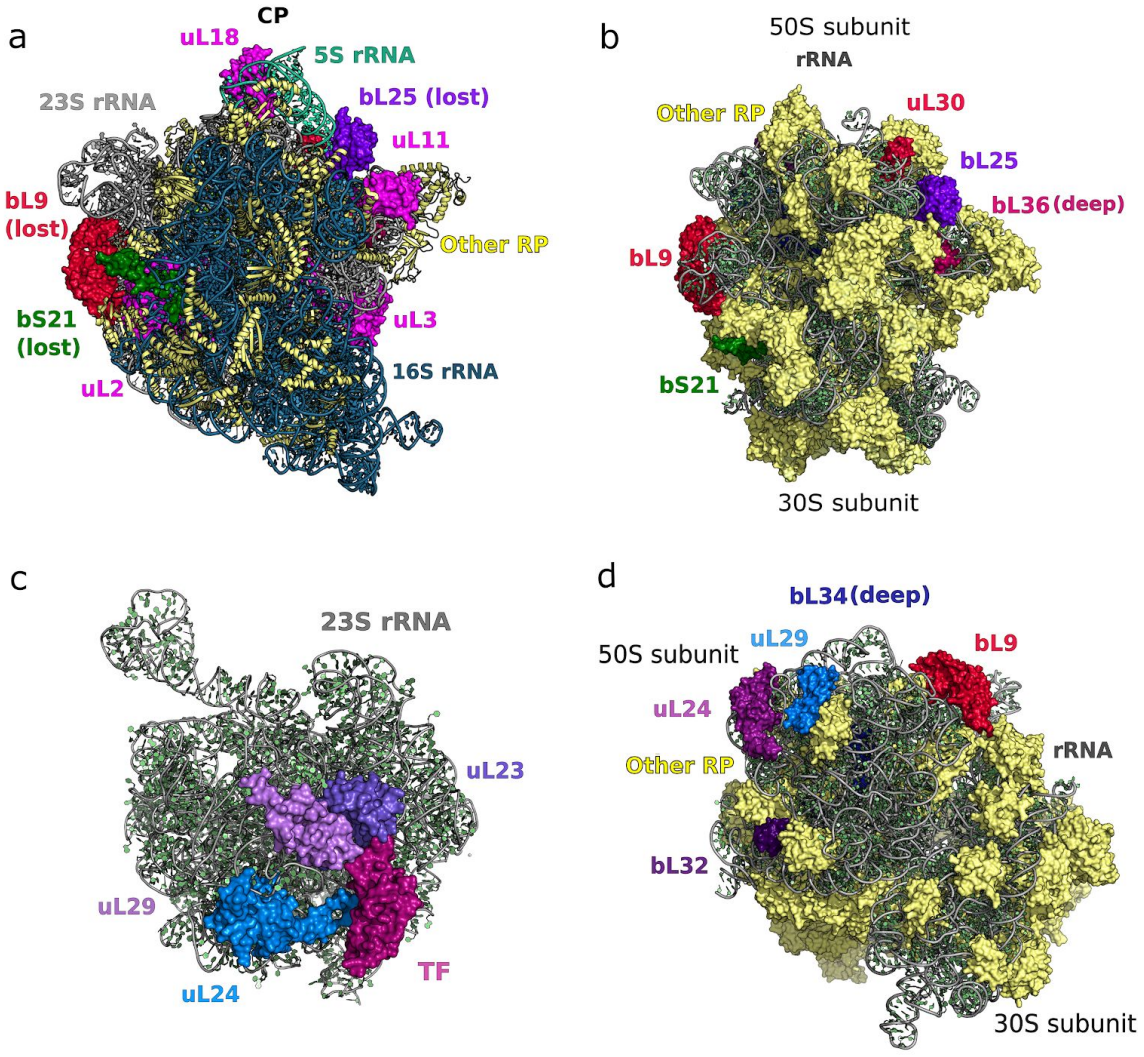

Figure 5. Determinants of the ribosomal protein loss. The differences in the evolutionary rate (a) and the number of contacts (b) between the preserved (blue) and frequently lost (red) ribosomal proteins. (c) The dependency between the evolutionary rate and the protein length (number of amino acid residues) for both preserved and frequently lost ribosomal proteins.

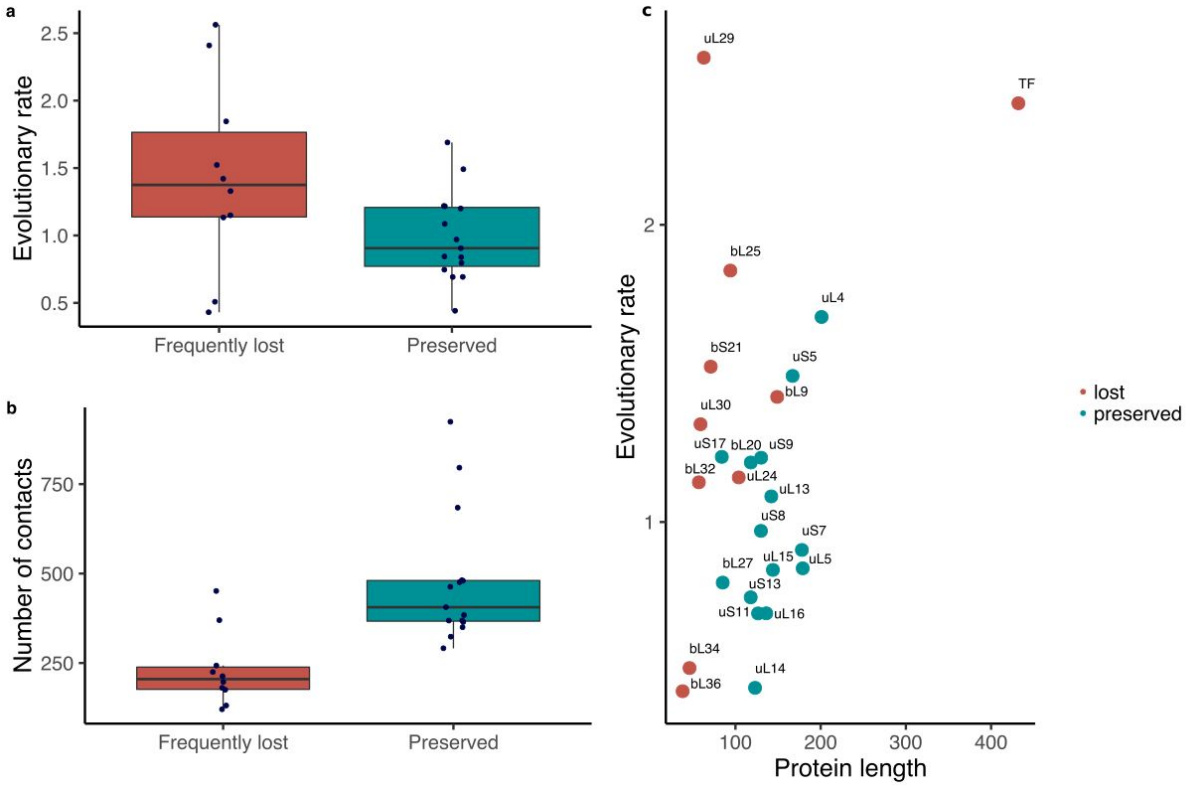
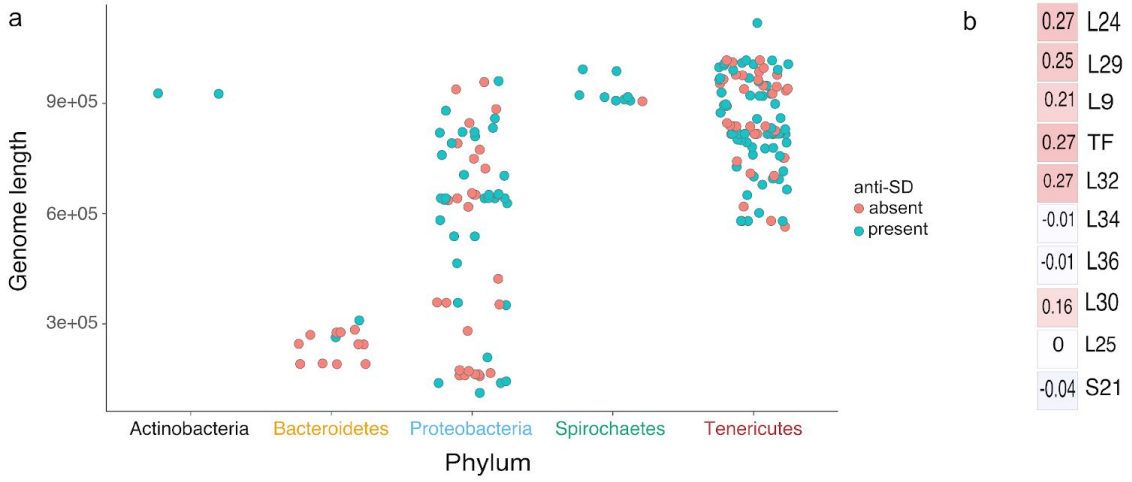
Figure 6. Loss of ribosomal proteins and the anti-Shine-Dalgarno box. (a) Dependency between the presence/absence of the anti-SD sequence and the genome length (bp). Each dot represents one of 214 strains. The anti-SD sequence is present in 16S rRNA of strains colored blue and absent in strains represented by red dots. The dots are grouped into bars by phyla. (b) The Pearson correlation coefficient between vectors of ribosomal protein and anti-SD presence/absence in bacterial strains. The color spectrum from blue to red corresponds to the correlation coefficient range from -1 to 1.