Title: Finding priority ribosomes for future structural and pathogen research based upon global RNA and protein sequence analysis

Helena B. Cooper[1,2], Kurt L. Krause[1] & Paul P. Gardner[1].

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago.
[2]Department of Infectious Diseases, Central Clinical School, Monash University.

Author Note

Keywords: Bacterial Ribosome, Antibiotic Resistance, Phylogeny Analysis, Genomics.

We have no known conflicts of interest to disclose.

Correspondence concerning this article should be addressed to:

Paul P. Gardner, Department of Biochemistry, University of Otago, P. O. Box 56, Dunedin, 9054.

Email: paul.gardner@otago.ac.nz

Abstract

Ribosome-targeting antibiotics comprise over half of antibiotics used in medicine, but our fundamental knowledge of their binding sites is derived primarily from ribosome structures from non-pathogenic species. These include *Thermus thermophilus*, *Deinococcus radiodurans* and *Haloarcula marismortui,* as well as commensal and pathogenic *Escherichia coli*. Advancements in electron cryomicroscopy have allowed for the determination of more ribosome structures from pathogenic bacteria, with each study highlighting species-specific differences that had not been observed in the non-pathogenic structures. These observed differences suggest that more novel ribosome structures, particularly from pathogens, are required for a more accurate understanding of the level of diversity of the entire bacterial ribosome, potentially leading to potential advancements in antibiotic research. In this study, high accuracy covariance and hidden Markov models were used to annotate ribosomal RNA and protein sequences respectively from genomic sequence, allowing us to determine the underlying ribosomal sequence diversity using phylogenetic methods. This analysis provided evidence that the current non-pathogenic ribosome structures are not sufficient representatives of some pathogenic bacteria, such as *Campylobacter pylori,* or of whole phyla such as Bacteroidetes.

Significance Statement

The growing number of antibiotic resistance pathogenic bacteria are of critical concern to the health profession. Many of the current classes of antibiotics target the bacterial ribosome, the protein making factory for these species. However, much of our knowledge of the bacterial ribosome is based upon non-pathogenic bacteria that are highly divergent from the major pathogens of concern. We have analysed the genetic variation of the RNA and protein components of all available bacterial ribosomes. This has led us to identify the highest priority groups of bacteria that would benefit most from further analysis of their ribosome structures, from both a medical and evolutionary perspective.

Introduction

A global rise of antibiotic resistant bacteria has become an increasingly urgent problem in recent years, with the bacterial ribosome, particularly the ribosomal RNA (rRNA) component, being a common antibiotic target (1–3). Due to the limitations (recently renamed "Mycolicibacterium smegmatis") and requirements of X-ray crystallography, antibiotic binding studies initially used extremophiles *Thermus thermophilus (4)*, *Deinococcus radiodurans* [a 50S structure] (5, 6) and the archaeon *Haloarcula marismortui (7)*, with the pathogenic *Escherichia coli (8–10)* ribosome structure introduced in later years (2, 3, 11). However, improvements in electron cryomicroscopy have allowed more diverse ribosome structures to be analysed more recently, such as from non-pathogenic bacteria *Bacillus subtilis (12, 13), Flavobacterium johnsoniae (14)* and *Mycobacterium smegmatis* (15) (recently renamed "*Mycolicibacterium smegmatis*"), or from pathogens including *Acinetobacter baumannii (16, 17)*, *Enterococcus faecalis (18, 19), Listeria monocytogenes (19), Mycobacterium tuberculosis* [a 50S structure] *(20)*, *Pseudomonas aeruginosa* [a 50S structure] (21) and *Staphylococcus aureus* (19, 22, 23). Each pathogen ribosome study highlights species-specific differences in comparison to both non-pathogenic and pathogenic structures, implying that the non-pathogenic ribosome structures are not always optimal for inferring antibiotic binding across all bacteria (16, 20–22). As these structural differences could hinder antibiotic research it is important that solved ribosome structures, which form the basis of our understanding of the bacterial ribosome and ribosomal antibiotic binding, are suitably representative of pathogenic species (2, 3).

We further note that the specific target sites of antibiotics correspond to highly conserved, functional ribosome components, and that a survey focusing on these sites will not yield significant diversity. However, we also note that ribosome function has been found to be influenced by less well conserved components of structure (2, 22), and even by elements of structure quite distant from conserved antibiotic binding sites (24).

Therefore, the aim of this study is to broadly survey ribosome sequence space to determine how representative current ribosome structures are, ranging from the extremophiles (eg: *D. radiodurans, T. thermophilus*) for human pathogenic species (eg: *S. aureus, L. monocytogenes*), when compared to all bacterial species. Using this analysis, we propose new representative bacteria, including some surprising and diverse species, with the potential of yielding valuable new ribosome structural information that could be prioritised for future ribosomal structural studies.

Materials and Methods

Our analysis includes the full-length conserved components of the bacterial ribosome. We have elected not to specifically analyse antibiotic targets (e.g. Spectinomycin targets G1064 and C1992 of 16S rRNA (25)) as these are highly conserved functional components of the ribosome, which therefore carry a limited number of phylogenetically informative characters (26). Full-length sequences also allow a global overview of ribosome variation, some of which may change the behaviour of binding sites due to "action at a distance" (2, 22, 24).

A total of 3,758 bacterial genomes and the *H. marismortui* genome (Accession: AY596297.1) were obtained from the European Nucleotide Archive (27, 28). One representative sequence was retained per species for each rRNA and protein sequence, which were filtered to only include genomes with annotations covering 80% of the expected sequence length based upon consensus sequences. If paralogues were present, the sequence with the highest corresponding covariance model and hidden Markov model bit score for the species was used (29). Both 16S and 23S rRNA were annotated using barrnap v0.9 (https://github.com/tseemann/barrnap), and INFERNAL 1.1.2 was used to create sequence alignments using Rfam v14.3 covariance models (29, 30). Covariance models are the gold-standard for structural RNA alignment (29, 31), and the Rfam models are derived from the CRW database (32), which is the product of decades of manual alignment curation, resulting in some of the highest quality ribosomal RNA structural alignments available to date (32, 33). Ribosomal protein sequences from 32 universally conserved proteins were identified from six-frame translations of whole genomes and open-reading frame predictions, to account for potentially inconsistent or absent genome annotations. The selected sequences were aligned using HMMER v3.1 (hmmer.org), with protein hidden Markov models from Pfam v33.1 (34). Profile HMMs are the gold-standard for protein homology search and alignment (35, 36), Pfam is a widely used and highly curated database of protein domain alignments. The advantages of this approach are that our results are up-to-date and consistent with the current genome databases, contain both RNA and protein components of the bacterial ribosome and utilise the most accurate methods for alignment production. Whilst we acknowledge that there are existing ribosomal datasets (e.g. SILVA, GreenGenes, RDB, etc), these do not meet all the above requirements. Although there are alternative ribosomal RNA and protein alignments we could have used (e.g. (26, 37, 38)), which will presumably producing similar results, the Rfam and Pfam alignments have been continuously curated for more than 20 years and are shipped with statistical representations for effective homology searches (covariance and hidden Markov models). Phylogenetic trees for each alignment were generated using the maximum likelihood method from PHYLIP v3.697, with distance matrices of the pairwise distances between species computed in R v4.2.1 using ape v5.6 (39, 40). The distance matrices for each ribosomal gene were summed to create a single unified distance matrix, and this was used for multi-dimensional scaling (MDS) and visualisation (Fig. 1).

Results

Ribosomal sequence clustering suggests that several bacterial phyla are poorly represented by the current structures. To identify phyla that are poorly represented by solved ribosome structures, multidimensional scaling (MDS) was used to reduce two rRNA and 26 ribosomal protein-derived phylogeny trees, which captured the most dissimilarities between species. As assessing individual ribosomal sequences is unlikely to reflect structural variation, we combined 28 phylogenetic trees for each conserved rRNA and protein, creating one distance matrix prior to MDS (38). Only ribosomal genes that are conserved across the majority of the species in our dataset were used. Therefore, ribosomal proteins uS2, uS4, uL15, uL23, uL24 and uL30 were removed as there were no homologous sequences within Epsilonproteobacteria.

The resulting MDS plot shows four main clusters; two illustrate the non-monophyletic nature of Proteobacteria (38), one dominated by Actinobacteria and the fourth consisting of the remaining bacterial phyla (Fig. 1). The Proteobacteria cluster follows observations made in previous studies, such as Beta and Gammaproteobacteria being closely related, and Deltaproteobacteria and Oligoflexia being more related to non-Proteobacteria (38, 41). Due to these non-monophyletic properties, Proteobacteria classes will be treated as individual phyla for the rest of this study (38, 41). Bacteroidetes, Acidithiobacilla and Epsilonproteobacteria formed slightly isolated groups away from the larger clusters, suggesting that these could become more defined clusters if their sample sizes were larger (Fig. 1). Alphaproteobacteria is the only cluster without a solved structure, implying the presence of phyla specific variation that has not been captured by current structures (Fig. 1). Therefore, it is unlikely that all the bacterial phyla are well represented by the available ribosome structures, given that the solved structures in the multiple phyla cluster group together, instead of being evenly distributed throughout the clusters (Fig. 1).

Evaluation of current structures indicates that *Bacillus subtilis* is the most representative. To evaluate whether current ribosome structures from non-pathogens are sufficiently representative of the available sequenced bacterial ribosomes, phylogenetic distances were calculated from the summed distance matrix between 13 published solved structures (Supplementary Table 2) and 1,385 bacterial ribosomal sequences available in this study (11–16, 18, 20–22). The solved structure with the lowest recorded distance for each species is considered to be the most representative and assumes that species with similar primary sequences will form similar tertiary structures. As the minimum distance to the nearest solved structure increases, it becomes more likely that current structures are not suitably representative. These underrepresented species should be prioritised for future ribosome structural studies. Overall, *B. subtilis* was considered to be the most representative structure for 382 species, followed by *A. baumannii* for 264 species and 167 for *M. smegmatis* (Fig. 1). *D. radiodurans*, *T. thermophilus* and unsurprisingly the archaeon *H. marismortui* were the three least representative structures analysed and were not representative of any pathogenic species (Fig. 1), implying that the structures from these three species are becoming less representative with the introduction of ribosome structures from less divergent bacteria (11). *M. tuberculosis* was also representative for a small number of species, which is likely due to the presence of *M. smegmatis*, given that these two bacteria are closely related – effectively splitting the Actinobacteria (Fig. 1).

This observation does not imply that one structure is necessarily sufficient to represent a complete phyla or class, with *E. coli* and *P. aeruginosa* representing 149 and 110 species respectively, accounting for most Gammaproteobacteria (Fig. 1). Whilst these structures are very similar, they do capture some important differences (21). Each solved structure tends to solely be representative of the phyla it originated from, with *M. smegmatis* as the most representative of Actinobacteria, and both *E. coli* and *P. aeruginosa* representing Gammaproteobacteria exclusively. However, *B. subtilis* and *A. baumannii* were not representative of only their respective phyla, as the Gammaproteobacterium *A. baumannii* is representative of both Alpha- and Betaproteobacteria and the Bacillota/Firmicutes *B. subtilis* was the most representative for the remaining phyla without a ribosome structure (Table 1).

Therefore, we hypothesise that having at least one representative structure per phylum, or preferably one per class, would allow the majority of bacteria to be sufficiently represented.

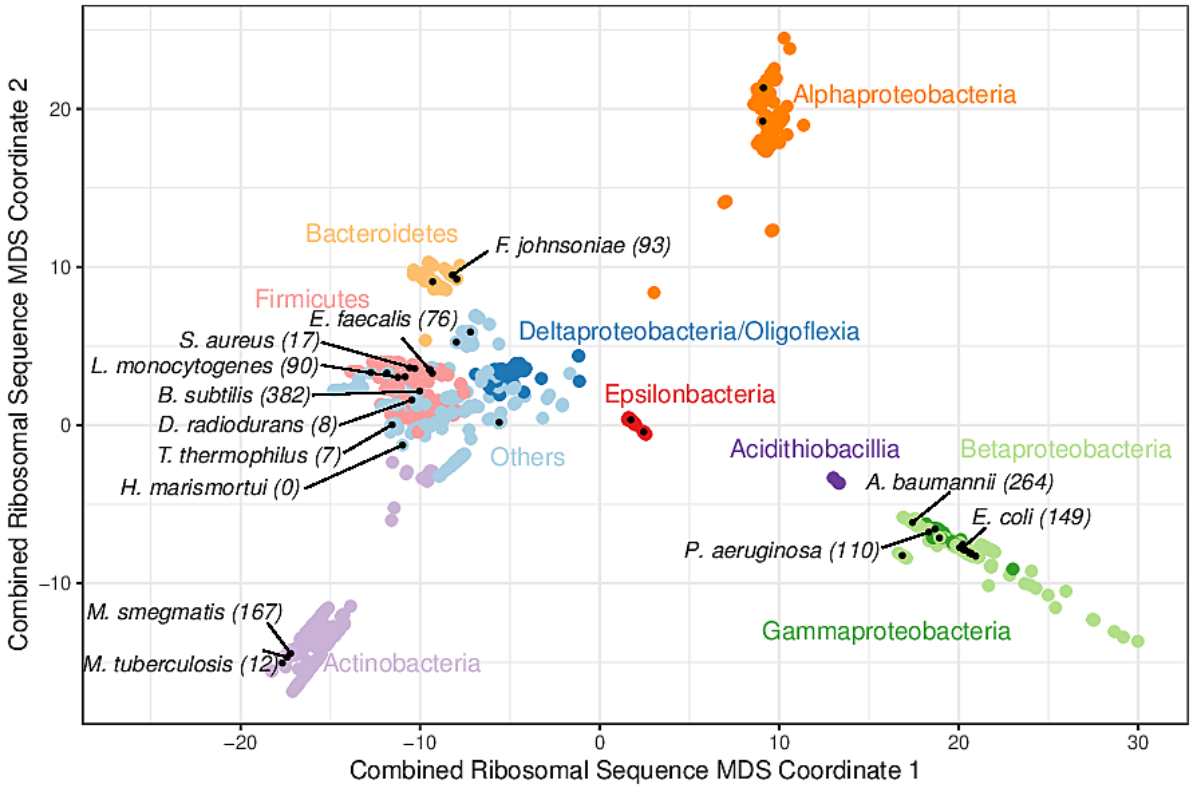


Fig. 1. A multiple-dimensional scaling (MDS) plot for combined phylogenetic distances from 16S, 23S rRNA and 26 universally conserved ribosomal proteins (N=1,396). All Proteobacteria are coloured by class, and other included phyla are coloured individually to highlight taxonomic relationships with clusters. Species with a solved ribosome structure (N=13), or known pathogenic bacteria without a structure (N=38) have been highlighted with a black dot. The number of bacteria that consider each species with solved structure to be representative, based on the minimum phylogenetic distance, has been labelled along with the species' names. The full list of species, MDS coordinates, the solved structures that were considered to be most representative and the corresponding minimum distance are available on Github.

Introducing new ribosome structures shows that an Epsilonproteobacteria representative, such as a *Campylobacter jejuni*, should be prioritised for structural work. To simulate the effect of having at least one representative ribosome structure per phyla, we selected one proposed structure per phyla which had the smallest average phylogenetic distance to all other members in the respective phyla. Only species with a minimum distance to a solved structure greater than 12.86 were considered, which is above the lower quartile of all the minimum distances recorded, as these species are most likely to be poorly represented by current structures. The available pathogens with a minimum distance above this threshold were prioritised due to their relevance in health research. The introduction of new proposed structures resulted in a decrease in the average minimum distance to a solved structure across each phylum, implying that having at least one

structure per phylum is a suitable sampling strategy for improving representation (Table 1). Of the ten highest priority structures, only Epsilonproteobacteria, Chlamydiae and Chlorobi had an average minimum distance per phyla below the lower quartile threshold, after the proposed structures were introduced (Table 1). This suggests that the other phyla listed are more diverse and may require additional structures to capture the remaining variation, or that a representative non-pathogenic ribosome would be a more beneficial representative. *Campylobacter jejuni* was identified as the highest priority structure (Table 1), as it was the most representative Epsilonproteobacteria pathogen, had the largest average minimum distance observed across all phyla and is a WHO priority pathogen (1). However, the selected ribosome for solving does not specifically need to be from a pathogenic strain, with either an attenuated lab strain (20, 22), or another species from the same genus (15, 18) being appropriate alternatives (Fig. 1).

Table 1. The ten highest priority ribosome structures to solve, with prioritisation given to pathogenic species, based on the average minimum distances to a solved structure prior to the introduction of each proposed structure (See Supplementary Table 1 for an extended list).

Proposed target species	Phylum	Min distance to closest structure	Avg min distance for phylum*
Campylobacter jejuni	Epsilonproteobacteria	41.07 (*B. subtilis*)	42.24 (10.11)
Chlamydophila trachomatis	Chlamydiae	40.25 (*B. subtilis*)	39.58 (8.27)
Opitutaceae bacterium	Verrucomicrobia	38.24 (*B. subtilis*)	37.89 (16.00)
Borrelia recurrentis	Spirochaetes	38.00 (*B. subtilis*)	39.41 (20.50)
Singulisphaera acidiphila	Planctomycetes	37.81 (*B. subtilis*)	39.54 (21.48)
Brucella melitensis	Alphaproteobacteria	37.73 (*A. baumannii*)	38.78 (19.26)
Chlorobium limicola	Chlorobi	37.52 (*F. johnsoniae*)	37.34 (5.64)
Ureaplasma urealyticum	Tenericutes	35.05 (*L. monocytogenes*)	32.16 (26.65)
Leptospirillum ferriphilum	Nitrospirae	33.96 (*B. subtilis*)	32.63 (16.34)
Persephonella marina	Aquificae	33.29 (*B. subtilis*)	34.06 (14.81)

* The first value is the average minimum distance prior to the proposed structures being selected, and the value in brackets is the recalculated average distance after the proposed structures have been incorporated.

Discussion

Our results reveal significant phylogenetic gaps in the existing cohort of solved ribosome structures. Filling these gaps will help identify clade-specific features of isolated bacterial clades that lack representative ribosome structures. This is important for the analysis of ribosome evolution, pathogen research, and potentially for antibiotic development.

There are two limitations with this approach to ribosome structure prioritisation: there is a bias towards species with available genomes, which are enriched in pathogenic species rather than being the most evolutionarily representative (42). In addition, our approach prioritises the most novel ribosomes within each major bacterial clade, with the risk of

over-emphasising the most diverse ribosomes. This is somewhat mitigated by selecting the most closely related pathogenic species as representative targets.

While having at least one representative ribosome structure per phyla is expected to capture the bulk of structural variation (Table 1), there is no guarantee that primary sequence differences result in significant changes at the tertiary level. A prominent example of this is *H. marismortui* which, although it is an archaean ribosome structure, has been used as an alternative to bacterial counterparts (11), despite having no close relatives at the primary level (Fig. 1). Another consideration is that the structural variation observed may be species-specific, rather than shared across a phylum, with species-specific differences having been observed between *P. aeruginosa* and *A. baumannii*, yet no phyla-specific differences when compared to *E. coli* and other available structures (16, 21). However, these three structures were observed to be relatively divergent from each other based upon the MDS analysis (Fig. 1), which reinforces the *in vivo* observation that *P. aeruginosa* and *A. baumannii* ribosomes are more similar to each other than to *E. coli,* suggesting that phylogenetic analyses have the potential to reflect structural variation (16).

Our list of high-priority bacterial ribosomes reveals that a number of bacterial clades are poorly represented by the existing solved structures (Table 1, Supplementary Table 1). Working down this list gives an opportunity to identify the most variant ribosomes based upon sequence analysis, which in many instances will reflect diverse ribosome structures. These will allow the development of more specific classes of antibiotics, as well as provide further information regarding the specifics of ribosome structure evolution.

Data Availability

All data described was last accessed in September 2022. Custom scripts, details of parameters, and dependencies used, and the accessions for all downloaded data and the resulting curated alignments and trees are available on GitHub: https://github.com/helena-bethany/ribosomal-variation

Funding

This work was supported by the Department of Biochemistry (University of Otago) as an Honours year research project and by an Australian Government Research Training Program (RTP) Scholarship .

Acknowledgements

The authors would like to thank Steven Gregory (University of Rhode Island) and Gerwald Jogl (Brown University) for their discussions and advice regarding this project.

References

1. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Pulcini C,

Kahlmeter G, Kluytmans J, Carmeli Y, Ouellette M, Outterson K, Patel J, Cavaleri M, Cox EM, Houchens CR, Grayson ML, Hansen P, Singh N, Theuretzbacher U, Magrini N, WHO Pathogens Priority List Working Group. 2018. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis 18:318–327.

2. Lin J, Zhou D, Steitz TA, Polikanov YS, Gagnon MG. 2018. Ribosome-Targeting Antibiotics: Modes of Action, Mechanisms of Resistance, and Implications for Drug Design. Annual Review of Biochemistry https://doi.org/10.1146/annurev-biochem-062917-011942.

3. Wilson DN. 2014. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. Nat Rev Microbiol 12:35–48.

4. Polikanov YS, Melnikov SV, Söll D, Steitz TA. 2015. Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. Nat Struct Mol Biol 22:342–344.

5. Harms JM, Wilson DN, Schluenzen F, Connell SR, Stachelhaus T, Zaborowska Z, Spahn CMT, Fucini P. 2008. Translational regulation via L11: molecular switches on the ribosome turned on and off by thiostrepton and micrococcin. Mol Cell 30:26–38.

6. Kaminishi T, Schedlbauer A, Fabbretti A, Brandi L, Ochoa-Lizarralde B, He C-G, Milón P, Connell SR, Gualerzi CO, Fucini P. 2015. Crystallographic characterization of the ribosomal binding site and molecular mechanism of action of Hygromycin A. Nucleic Acids Res 43:10015–10025.

7. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289:905–920.

8. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, Stark H. 2015. Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. Nature 520:567–570.

9. Noeske J, Wasserman MR, Terry DS, Altman RB, Blanchard SC, Cate JHD. 2015. High-resolution structure of the Escherichia coli ribosome. Nat Struct Mol Biol 22:336–341.

10. Watson ZL, Ward FR, Méheust R, Ad O, Schepartz A, Banfield JF, Cate JH. 2020. Structure of the bacterial ribosome at 2 Å resolution. Elife 9.

11. Fox GE. 2010. Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol 2:a003483.

12. Sohmen D, Chiba S, Shimokawa-Chiba N, Innis CA, Berninghausen O, Beckmann R, Ito K, Wilson DN. 2015. Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. Nat Commun 6:6941.

13. Crowe-McAuliffe C, Takada H, Murina V, Polte C, Kasvandik S, Tenson T, Ignatova Z, Atkinson GC, Wilson DN, Hauryliuk V. 2021. Structural Basis for Bacterial Ribosome-Associated Quality Control by RqcH and RqcP. Mol Cell 81:115–126.e7.

14. Jha V, Roy B, Jahagirdar D, McNutt ZA, Shatoff EA, Boleratz BL, Watkins DE, Bundschuh R, Basu K, Ortega J, Fredrick K. 2021. Structural basis of sequestration of the anti-Shine-Dalgarno sequence in the Bacteroidetes ribosome. Nucleic Acids Res 49:547–567.

15. Hentschel J, Burnside C, Mignot I, Leibundgut M, Boehringer D, Ban N. 2017. The Complete Structure of the Mycobacterium smegmatis 70S Ribosome. Cell Rep 20:149–160.

16. Morgan CE, Huang W, Rudin SD, Taylor DJ, Kirby JE, Bonomo RA, Yu EW. 2020. Cryo-electron Microscopy Structure of the Acinetobacter baumannii 70S Ribosome and Implications for New Antibiotic Development. MBio 11.

17. Zhang Z, Morgan CE, Bonomo RA, Yu EW. 2021. Cryo-EM Determination of Eravacycline-Bound Structures of the Ribosome and the Multidrug Efflux Pump AdeJ of Acinetobacter baumannii. MBio 12:e0103121.

18. Murphy EL, Singh KV, Avila B, Kleffmann T, Gregory ST, Murray BE, Krause KL, Khayat R, Jogl G. 2020. Cryo-electron microscopy structure of the 70S ribosome from Enterococcus faecalis. Sci Rep 10:16301.

19. Crowe-McAuliffe C, Murina V, Turnbull KJ, Kasari M, Mohamad M, Polte C, Takada H, Vaitkevicius K, Johansson J, Ignatova Z, Atkinson GC, O'Neill AJ, Hauryliuk V, Wilson DN. 2021. Structural basis of ABCF-mediated resistance to pleuromutilin, lincosamide, and streptogramin A antibiotics in Gram-positive pathogens. Nat Commun 12:3577.

20. Yang K, Chang J-Y, Cui Z, Li X, Meng R, Duan L, Thongchol J, Jakana J, Huwe CM, Sacchettini JC, Zhang J. 2017. Structural insights into species-specific features of the ribosome from the human pathogen Mycobacterium tuberculosis. Nucleic Acids Res 45:10884–10894.

21. Halfon Y, Jimenez-Fernandez A, La Rosa R, Espinosa Portero R, Krogh Johansen H, Matzov D, Eyal Z, Bashan A, Zimmerman E, Belousoff M, Molin S, Yonath A. 2019. Structure of Pseudomonas aeruginosa ribosomes from an aminoglycoside-resistant clinical isolate. Proc Natl Acad Sci U S A 116:22275–22281.

22. Eyal Z, Matzov D, Krupkin M, Wekselman I, Paukner S, Zimmerman E, Rozenberg H, Bashan A, Yonath A. 2015. Structural insights into species-specific features of the ribosome from the pathogen Staphylococcus aureus. Proc Natl Acad Sci U S A

112:E5805–14.

23. Halfon Y, Matzov D, Eyal Z, Bashan A, Zimmerman E, Kjeldgaard J, Ingmer H, Yonath A. 2019. Exit tunnel modulation as resistance mechanism of S. aureus erythromycin resistant mutant. Sci Rep 9:11460.

24. Belousoff MJ, Eyal Z, Radjainia M, Ahmed T, Bamert RS, Matzov D, Bashan A, Zimmerman E, Mishra S, Cameron D, Elmlund H, Peleg AY, Bhushan S, Lithgow T, Yonath A. 2017. Structural Basis for Linezolid Binding Site Rearrangement in the Staphylococcus aureus Ribosome. MBio 8.

25. Brink MF, Brink G, Verbeet M, de Boer HA. 1994. Spectinomycin interacts specifically with the residues G 1064 and C 1192 in 16S rRNA, thereby potentially freezing this molecule into an inactive conformation. Nucleic Acids Res 22:325–331.

26. Bernier CR, Petrov AS, Kovacs NA, Penev PI, Williams LD. 2018. Translation: The Universal Structural Core of Life. Mol Biol Evol 35:2065–2076.

27. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martinez-Villacorta J, Milano A, Pakseresht A, Rahman N, Rajan J, Reddy K, Richards E, Smirnov D, Sokolov A, Vijayaraja S, Cochrane G. 2020. The European Nucleotide Archive in 2019. Nucleic Acids Res 48:D70–D76.

28. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, Shannon P, Chiu Y, Weng RS, Gan RR, Hung P, Date SV, Marcotte E, Hood L, Ng WV. 2004. Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. Genome Res 14:2221–2234.

29. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches.

Bioinformatics 29:2933–2935.

30. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46:D335–D342.

31. Freyhult EK, Bollback JP, Gardner PP. 2007. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17:117–125.

32. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2.

33. Gutell RR, Lee JC, Cannone JJ. 2002. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol 12:301–310.

34. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432.

35. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

36. Söding J, Remmert M. 2011. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol 21:404–411.

37. Doris SM, Smith DR, Beamesderfer JN, Raphael BJ, Nathanson JA, Gerbi SA. 2015. Universal and domain-specific sequences in 23S-28S ribosomal RNA identified by

computational phylogenetics. RNA 21:1719–1730.

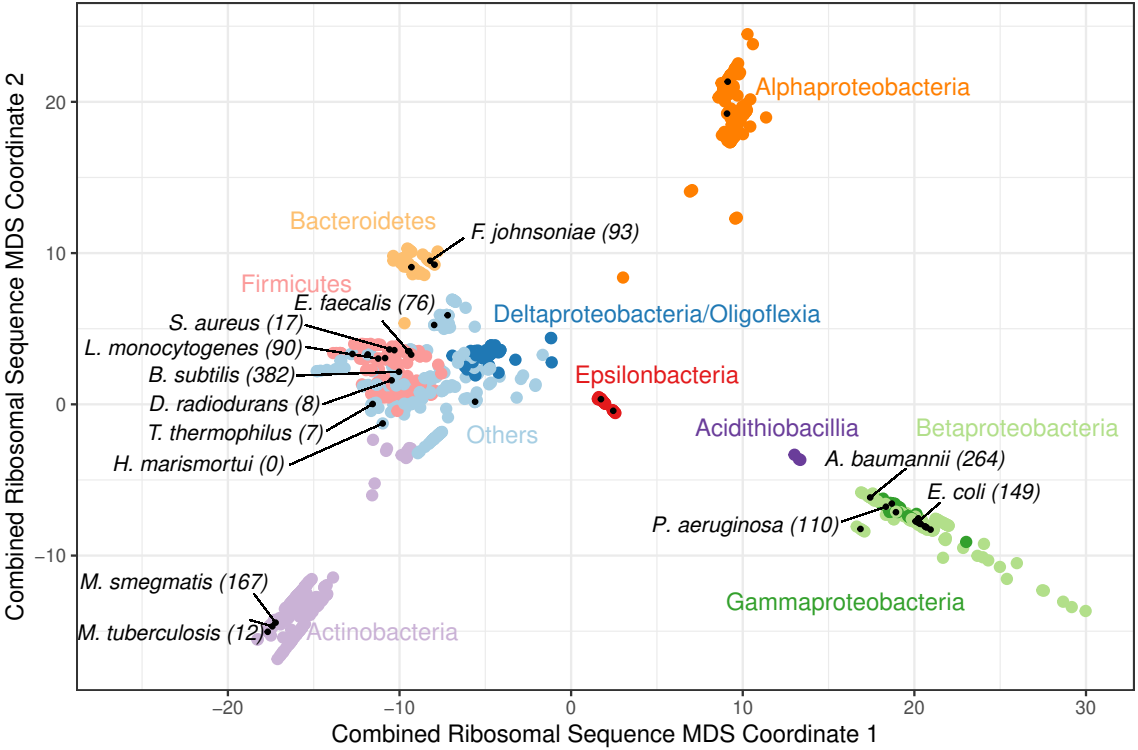
38. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.

39. Felsenstein J. 2009. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. http://www.evolution.gs.washington.edu/phylip.html. Department of Genome Sciences, University of Washington.

40. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

41. Waite DW, Chuvochina M, Pelikan C, Parks DH, Yilmaz P, Wagner M, Loy A, Naganuma T, Nakai R, Whitman WB, Hahn MW, Kuever J, Hugenholtz P. 2020. Proposal to reclassify the proteobacterial classes Deltaproteobacteria and Oligoflexia, and the phylum Thermodesulfobacteria into four phyla reflecting major functional capabilities. Int J Syst Evol Microbiol 70:5972–6016.

42. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060.


Alphaproteobacteria
Bacteroidetes
F. johnsoniae (93)
Firmicutes
E. faecalis (76)
S. aureus (17)
L. monocytogenes (90)
B. subtilis (382)
D. radiodurans (8)
T. thermophilus (7)
H. marismortui (0)
Deltaproteobacteria/Oligoflexia
Epsilonbacteria
Others
Acidithiobacillia
Betaproteobacteria
A. baumannii (264)
E. coli (149)
P. aeruginosa (110)
Gammaproteobacteria
M. smegmatis (167)
M. tuberculosis (12)
Actinobacteria
20
10
0
−10
−20
−10
0
10
20
30
Combined Ribosomal Sequence MDS Coordinate 2
Combined Ribosomal Sequence MDS Coordinate 1