Investigating bacterial ribosomal sequence variation in regards to future structural and antibiotic research.

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

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago.

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 the commensal or 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 to get a more accurate understanding of the level of diversity in the bacterial ribosome, leading to potential advancements in antibiotic research. In this study, 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 be of the most benefit for 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 and requirements of X-ray crystallography, antibiotic binding studies initially used extremophiles *Thermus thermophilus*, *Deinococcus radiodurans* and *Haloarcula marismortui*, with the pathogenic *Escherichia coli* ribosome structure introduced in later years (2–4). However, improvements in electron cryomicroscopy have allowed more diverse ribosome structures to be analysed, such as from non-pathogenic bacteria *Mycobacterium smegmatis* and *Bacillus subtilis* or from pathogens including *Enterococcus faecalis, Staphylococcus aureus*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii* (5–11). Each pathogen ribosome study highlighted 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 (8–11). 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 suitable for representing pathogenic species (2, 3). Therefore, the aim of this study is to determine how representative *D. radiodurans*, *T. thermophilus*, *H. marismortui* and *E. coli* are of all bacterial species, particularly pathogens, and to prioritise representative bacterial ribosomes that would be the most beneficial to have solved structures for.

Materials and Methods

A total of 3,758 bacterial genomes and the *H. marismortui* genome (Accession: AY596297.1) were obtained from the European Nucleotide Archive (12, 13). One representative sequence was retained per species for each rRNA and protein sequence, which were filtered to only include those annotations with at least 80% of the expected sequence length based upon consensus sequences. If paralogues were present, the sequence with the highest corresponding bit score for the species was used (14). 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 (14, 15). Ribosomal protein sequences from 32 universally conserved proteins were annotated in 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 (16). 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 v4.0.3 using ape v5.4 (17, 18). 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 it is unlikely all bacterial phyla are well represented by current structures. To identify phyla that are underrepresented by solved ribosome structures, MDS was used to reduce two rRNA and 26 ribosomal protein-derived phylogeny trees to capture the most dissimilarities between species. As individually assessing each ribosomal sequence is unlikely to reflect structural variation, the 28 phylogeny trees were combined to create one distance matrix prior to MDS (19). Only ribosomal genes which are conserved across the majority of bacteria were used, resulting in ribosomal proteins uS2, uS4, uL15, uL23, uL24 and uL30 being removed, as there were no homologous sequences present in any of the Epsilonproteobacteria or in *H. marismortui*. The resulting MDS plot showed four main clusters; two driven by the non-monophyletic nature of Proteobacteria (19), one driven by Actinobacteria and the fourth consisting of the remaining bacterial phyla (Fig. 1). The Proteobacteria clustering follows observations made in previous studies, such as Beta and Gammaproteobacteria being the most closely related or Deltaproteobacteria and Oligoflexia being more related to non-Proteobacteria (19, 20). Due to these non-monophyletic properties, Proteobacteria classes will be treated as individual phyla for the rest of this study (19, 20). Bacteroidetes, Acidithiobacilla and Epsilonproteobacteria tended to form 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 appeared to be the most underrepresented cluster as they were 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 bacterial phyla are well represented by the set of currently available ribosome structures, given that the solved structures in the multiple phyla cluster group together, instead of being evenly distributed throughout the cluster (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 any bacterial ribosome, phylogenetic distances were calculated from the summed distance matrix between 11 published solved structures and 1,385 other bacteria available in this study (4–11). 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, meaning that these underrepresented species should be prioritised for future ribosome structural studies. Overall, *B. subtilis* was considered to be the most representative structure for 534 species, followed by *A. baumannii* for 268 species and 167 for *M. smegmatis* (Fig. 1 & Table 1). *D. radiodurans*, *T. thermophilus* and *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 relevant with the introduction of ribosome structures from less divergent bacteria (4). *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 (Fig. 1).

This observation does not imply that one structure is necessarily sufficient to represent a phyla or class, with *E. coli* and *P. aeruginosa* representing 149 and 111 species respectively, accounting for most Gammaproteobacteria (Fig. 1). Each solved structure tends to only be representative of the phyla it originated from, with *M. smegmatis* being 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 clades, as *A. baumannii* was the most representative for Alpha and Betaproteobacteria and *B. subtilis* was the most representative for all remaining phyla without a ribosome structure (Table 1). Therefore, we would hypothesise that having at least one representative structure per phylum, or preferably one per class, would likely allow the majority of bacteria to be sufficiently represented.

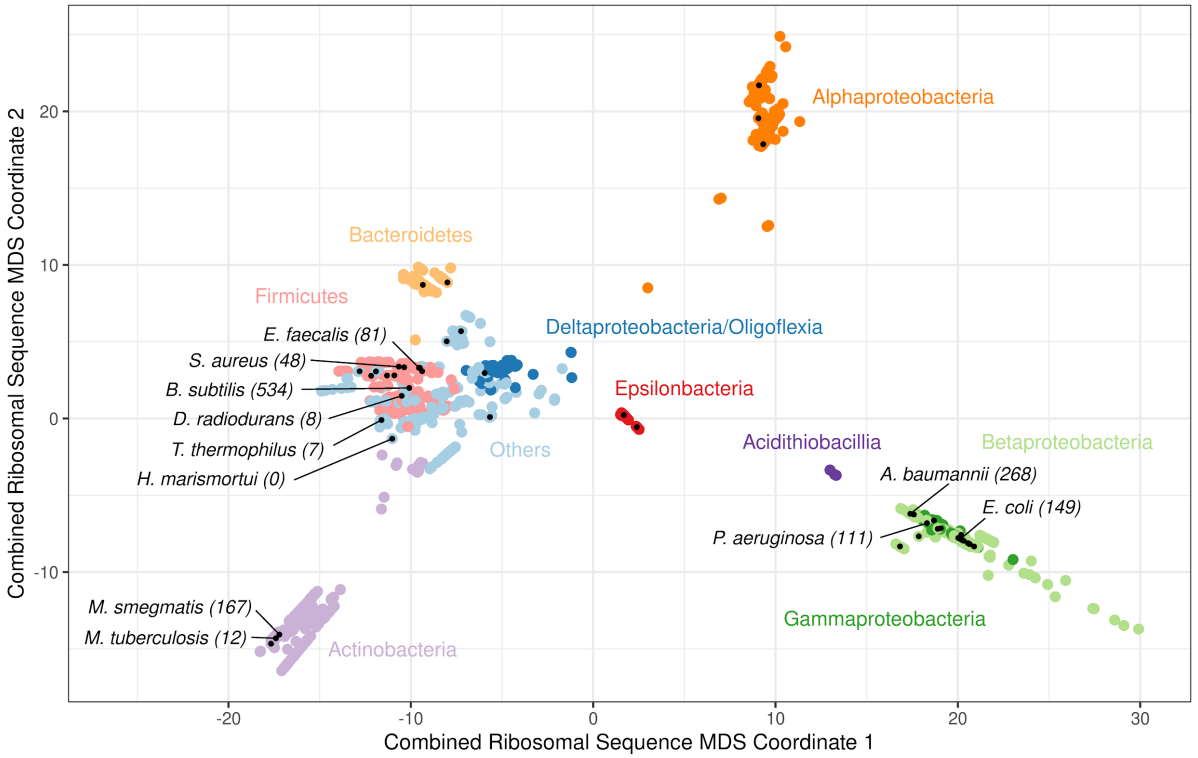


Fig. 1. MDS plot for the combined phylogenetic distances from 16S, 23S rRNA and 26 universally conserved ribosomal proteins (N=1,396). All Proteobacteria are coloured by class, and other larger phyla are coloured individually to highlight clustering. Species with a solved ribosome structure (N=11) or are known pathogenic bacteria without a structure (N=38) have been labelled 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 solving. 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 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 likely to be poorly represented by current structures. Available pathogens with a minimum distance above this threshold were prioritised due to their relevance in ribosome-targeting antibiotic 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 phyla is a reasonable 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, once 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 more representative non-pathogenic ribosome would be more beneficial for these phyla instead of a pathogen. *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 ribosome structure solved does not specifically need to be from a pathogenic strain, with either an attenuated lab strain (8, 9) or another species from the same genus (5, 7) being appropriate alternatives (Fig. 1).

Table 1. The ten highest priority ribosome structures to solve, based on the average minimum distances to a solved structure prior to the introduction of each proposed structure.

Proposed Structure	Phylum	Min Distance to Closest Structure	Avg Min Distance for Phylum*
Campylobacter jejuni	Epsilonproteobacteria	41.07 (*B. subtilis*)	42.24 (10.11)
Chlamydophila pneumoniae	Chlamydiae	40.29 (*B. subtilis*)	39.64 (8.05)
Singulisphaera acidiphila	Planctomycetes	37.81 (*B. subtilis*)	39.54 (21.48)
Borrelia recurrentis	Spirochaetes	38.00 (*B. subtilis*)	39.35 (20.86)
Chlorobium limicola	Chlorobi	38.31 (*B. subtilis*)	38.08 (5.64)
Brucella melitensis	Alphaproteobacteria	37.73 (*A. baumannii*)	38.78 (19.38)
Capnocytophaga canimorsus	Bacteroidetes	36.74 (*B. subtilis*)	38.54 (17.28)
Opitutaceae bacterium	Verrucomicrobia	38.24 (*B. subtilis*)	37.89 (16.00)
Ureaplasma urealyticum	Tenericutes	35.61 (*B. subtilis*)	33.88 (26.65)
Leptospirillum ferriphilum	Nitrospirae	33.96 (*B. subtilis*)	32.63 (16.34)

* 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

There are two limitations with this method of ribosome structure prioritisation: the bias towards species with available genomes, and the preference for pathogenic species rather than the most evolutionarily representative. An alternative method for removing these biases would be to prioritise species based only upon maximum phylogenetic diversity (21), allowing us to capture the maximum variation observed across bacterial ribosomes, regardless of whether the phylum hosts pathogenic species. However, the compromise of capturing the structural variation for outlier species is that they may not be informative of the phylum as a whole, limiting the applicability of these structures.

While having at least one representative ribosome structure per phyla is expected to capture more structural variation (Table 1), there is no guarantee that primary sequence differences will 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 (4), despite having no close relatives at the primary level (Fig. 1). Another consideration is that the structural variation observed may only be

species-specific, rather than at the phyla level, with species-specific differences having been observed between *P. aeruginosa* and *A. baumannii*, but no phyla-specific differences when compared to *E. coli* and other available structures (10, 11). 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 (11).

Data Availability

All data described was last accessed in December 2020. 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.

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,D.L., Pulcini,C., Kahlmeter,G., Kluytmans,J., Carmeli,Y., *et al.* (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,T.A., Polikanov,Y.S. and Gagnon,M.G. (2018) Ribosome-Targeting Antibiotics: Modes of Action, Mechanisms of Resistance, and Implications for Drug Design. *Annual Review of Biochemistry*, **87**, 451–478.

3. Wilson,D.N. (2014) Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat. Rev. Microbiol.*, **12**, 35–48.

4. Fox,G.E. (2010) Origin and evolution of the ribosome. *Cold Spring Harb. Perspect. Biol.*, **2**, a003483.

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

6. Sohmen,D., Chiba,S., Shimokawa-Chiba,N., Innis,C.A., Berninghausen,O., Beckmann,R., Ito,K. and Wilson,D.N. (2015) Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. *Nat. Commun.*, **6**, 6941.

7. Murphy,E.L., Singh,K.V., Avila,B., Kleffmann,T., Gregory,S.T., Murray,B.E., Krause,K.L., Khayat,R. and Jogl,G. (2020) Cryo-electron microscopy structure of the 70S ribosome from Enterococcus faecalis. *Sci. Rep.*, **10**, 16301.

8. Eyal,Z., Matzov,D., Krupkin,M., Wekselman,I., Paukner,S., Zimmerman,E., Rozenberg,H., Bashan,A. and 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.

9. Yang,K., Chang,J.-Y., Cui,Z., Li,X., Meng,R., Duan,L., Thongchol,J., Jakana,J., Huwe,C.M., Sacchettini,J.C., *et al.* (2017) Structural insights into species-specific features of the ribosome from the human pathogen Mycobacterium tuberculosis. *Nucleic Acids Res.*, **45**, 10884–10894.

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

11. Morgan,C.E., Huang,W., Rudin,S.D., Taylor,D.J., Kirby,J.E., Bonomo,R.A. and Yu,E.W. (2020) Cryo-electron Microscopy Structure of the Acinetobacter baumannii 70S Ribosome and Implications for New Antibiotic Development. *MBio*, **11**.

12. Amid,C., Alako,B.T.F., Balavenkataraman Kadhirvelu,V., Burdett,T., Burgin,J., Fan,J., Harrison,P.W., Holt,S., Hussein,A., Ivanov,E., *et al.* (2020) The European Nucleotide Archive in 2019. *Nucleic Acids Res.*, **48**, D70–D76.

13. Baliga,N.S., Bonneau,R., Facciotti,M.T., Pan,M., Glusman,G., Deutsch,E.W., Shannon,P., Chiu,Y., Weng,R.S., Gan,R.R., *et al.* (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. *Genome Res.*, **14**, 2221–2234.

14. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

15. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.

16. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

17. Felsenstein,J. (2009) PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. *http://www.evolution.gs.washington.edu/phylip.html*.

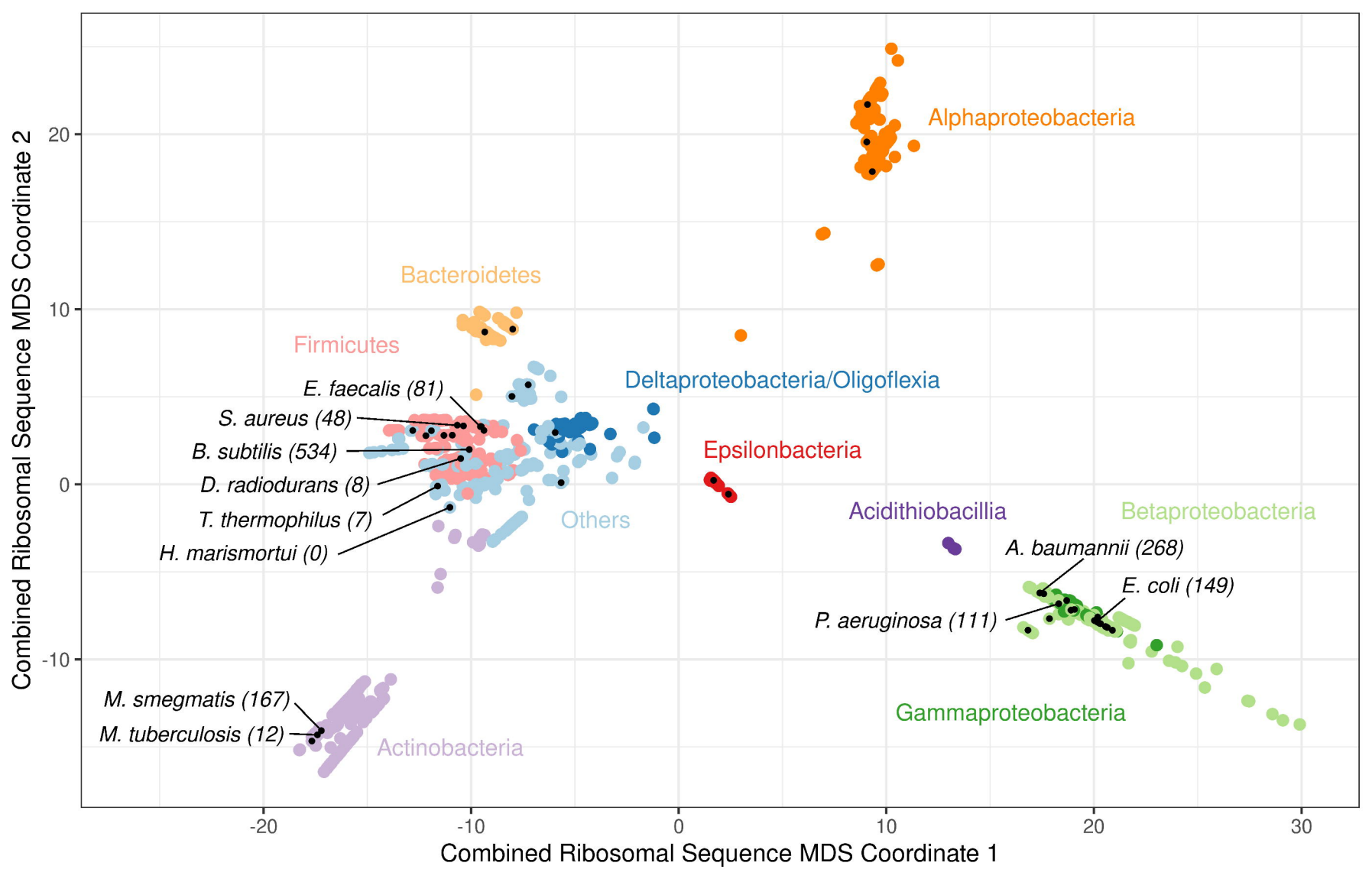
18. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.

19. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K., *et al.* (2016) A new view of the tree of life. *Nat Microbiol*, **1**, 16048.

20. Waite,D.W., Chuvochina,M., Pelikan,C., Parks,D.H., Yilmaz,P., Wagner,M., Loy,A.,

Naganuma,T., Nakai,R., Whitman,W.B., *et al.* (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.

21. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J., *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.


Alphaproteobacteria
Bacteroidetes
Firmicutes
Deltaproteobacteria/Oligoflexia
E. faecalis (81)
S. aureus (48)
B. subtilis (534)
D. radiodurans (8)
T. thermophilus (7)
H. marismortui (0)
Epsilonbacteria
Others
Acidithiobacillia
Betaproteobacteria
A. baumannii (268)
E. coli (149)
P. aeruginosa (111)
Gammaproteobacteria
M. smegmatis (167)
M. tuberculosis (12)
Actinobacteria
Combined Ribosomal Sequence MDS Coordinate 2
Combined Ribosomal Sequence MDS Coordinate 1
20
10
0
-10
-20
-10
0
10
20
30