Database, 2023, 1–6
DOI: https://doi.org/10.1093/database/baad058
Original article

TRGdb: a universal resource for the exploration of taxonomically restricted genes in bacteria

Andrzej Zielezinski, Wojciech Dobrychlop and Wojciech M. Karlowski*

Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznanskiego 6, Poznan 61-614, Poland

*Corresponding author: Tel: +48 61 8295841; E-mail: wmk@amu.edu.pl

Citation details: Zielezinski, A., Dobrychlop, W. and Karlowski, W.M. TRGdb: a universal resource for the exploration of taxonomically restricted genes in bacteria. *Database* (2023) Vol. 2023: article ID baad058; DOI: https://doi.org/10.1093/database/baad058

Abstract

The TRGdb database is a resource dedicated to taxonomically restricted genes (TRGs) in bacteria. It provides a comprehensive collection of genes that are specific to different genera and species, according to the latest release of bacterial taxonomy. The user interface allows for easy browsing and searching as well as sequence similarity exploration. The website also provides information on each TRG protein sequence, including its level of disorder, complexity and tendency to aggregate. TRGdb is a valuable resource for gaining a deeper understanding of the TRG-associated, unique features, and characteristics of bacterial organisms.

Database URL: www.combio.pl/trgdb

Introduction

Taxonomically restricted genes (TRGs) are genes that are present only in a particular taxonomic unit and have no traceable evolutionary history (1). As the TR genes have no homology to genes from other organisms, they are fundamentally important for the study of the emergence of organismal unique traits such as morphological diversity, metabolic innovation and pathogenicity (2). Additionally, TRGs serve as taxon-specific diagnostic targets that can be used to reconstruct the evolutionary history of specific phylogenetically related groups of organisms (3, 4). TRGs that are present at the species and genus levels garner special interest because they are expected to perform a role in defining exclusive ecological adaptations of organisms to particular niches (4, 5).

With whole-genome sequencing further facilitated by next-generation technologies, species- and genus-specific TR genes continue to be discovered in every newly sequenced genome across prokaryotes (1, 5–12), eukaryotes (13–17) and viruses (18, 19). Although eukaryotic TRGs are important for understanding the diversity and complexity of organisms, studying bacterial TRGs is crucial for gaining insight into the properties and evolution of biotechnologically, biomedically and scientifically important microorganisms (1). Paradoxically, although information on bacterial TRGs is essential for advancing our understanding of bacterial biology and evolution, it is not readily accessible. Two pioneering databases of species-specific TRGs in bacteria, OrphanMine (5) and ORFanage (20), have been unavailable for many years now, which hinders attempts at a systematic analysis of TRGs in the domain of bacteria.

To reconcile the deficiency of the current surveys of TRGs, we present a comprehensive database of species- and genus-specific TRGs in bacteria. By applying the TRG identification scheme from our recent work (21), we analyzed available genomes of >80 000 bacterial species represented by ∼250 million proteins. We supply a user-friendly interface for the results of the predictions in the context of the most up-to-date bacterial taxonomy, allowing users to browse and search for TRGs across different taxonomic units of bacteria. We also provide information on every TRG protein, including its sequence properties (e.g. level of disorder and complexity, and tendency to aggregate), affiliation to the TR gene family and links to external databases.

Materials and methods

Data sources

The sequence data, information about the taxonomic classification and a phylogenetic tree of bacteria were obtained from the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org) release 08-RS214 (April 2023) (22). The data set includes 80 789 representative genomes, with one genome per species. The GTDB provides high-quality bacterial taxonomy based on phylogenetic analysis and carefully selects the best representative genome for each species. The representative genomes in the GTDB have the highest publically available assembly quality, the least amount of

Received 21 February 2023; Revised 30 June 2023; Accepted 31 July 2023

contamination in the sequence, and the most complete set of genes. It should be noted, however, that some of the genomes may still be represented by incomplete sequences.

Identification of TRGs

Our TRG identification procedure (21) includes several steps that allow high-quality TR gene predictions. However, this procedure can currently be applied only to a limited-sized data set (e.g. a single genus). Therefore, the approach employed in the construction of TRGdb was simplified and limited to three progressive steps. First, we used DIAMOND v2.0.15 (23) to perform an all-versus-all comparison between protein sequences ($n = 247\,617\,414$) from 80 789 bacterial species. Then, we removed from further analysis query proteins that had highly similar (homologous) sequence matches ($E \leq 10^{-3}$) belonging to bacterial species outside the genus of the query species. This step allowed us to reduce the number of candidate TRGs by ∼90%. The remaining sequences were classified as the candidate TR genes at the genus level. Next, we verified the TRG candidates by querying them with BLAST+ v2.13.1 (24). The number of reported hits per query (-max_target_seqs) was adjusted to accommodate for the number of all tested species in a given genus. Queries that did not show significant similarity ($E \leq 10^{-3}$) to any sequences from organisms outside the tested genus were identified as genus-specific TR genes. Finally, we extracted a list of bacterial species-specific sequences from the obtained TR gene list. These genes were defined as protein sequences that did not have homologs outside the query species and the genus of query species encompassed at least two species according to the GTDB taxonomy.

Sequence properties calculations

To determine the properties of the TRG protein sequences, we followed the methodology described in the work of Karlowski *et al.* (21). We used IUPred2 (25) to calculate the disorder score of the TRG proteins, which reflects the extent of the intrinsic disorder in the proteins. This score was calculated by dividing the number of residues with a disorder score >0.5 by the total length of the protein sequence. We used TANGO v2.3.1 (26) to determine the TRG protein's average aggregation, which was presented as the frequency of potential aggregating segments defined as hexapeptides with an aggregation score >5% of all amino acid residues. Finally, we used SciPy v1.21.4 (27) to calculate the sequence complexity of the TRG proteins, which was assessed as the Shannon entropy.

TRG protein families

We classified the TRG proteins into families, similar to how protein families are established in the Pfam-B database (28). Accordingly, we used MMseqs2 v14-7e284 (29) with the cluster option. We set the maximum *E*-value at 10^{-3} and used the bidirectional coverage mode (–cov-mode 0) with a minimum coverage of 0.8 (-c 0.8).

Protein annotations

We divided each line of the description from the National Center for Biotechnology Information (NCBI) protein record into individual words and manually removed common and meaningless terms (such as numbers and punctuation characters). Only words with >10 appearances were considered.

Table 1. A summary of the TRG identification results

Data type	Number
Bacteria species in total	80 789
Bacteria proteins in total	247 617 414
TRG proteins	10 737 409 (4.34%)
Genus-specific TRG proteins	5 514 533 (2.23%)
Species-specific TRG proteins	5 222 876 (2.11%)
Bacteria species with TRG(s)	80 778 (99.99%)

Calculation of isolation index of organisms

To determine the degree of phylogenetic isolation of individual bacterial taxa, we calculated the isolation index of organism (IIO) (6). In our study, the IIO parameter was calculated based on the distance in the phylogenetic tree of bacteria in the GTDB.

Database interface and application programming interface

The TRGdb web interface was developed in HTML5 with the Bootstrap framework (v5.3), JavaScript and Highcharts.js (v10.3.3). The database querying system was developed in Django (v4.0.0), Django REST framework (v3.13.1) and Python (v3.9.5) using the SQLite database as a data management system.

Results

Abundance and distribution of TRGs in bacteria

We searched for the TR genes, conserved at the genus or species levels, in genomes of 80 789 bacterial species by developing the progressive TRG identification pipeline (see Materials and methods) that accounts for up-to-date taxonomy of bacteria provided by the GTDB (see Materials and methods). We found ∼11 million TRG proteins, representing ∼4.3% of all proteins in bacteria (Table 1). Of those proteins, 2.1% were found exclusively in genomes of a single species (referred to as species-specific TRGs) and 2.2% were present in the genomes of different species within the same genus (referred to as genus-specific TRGs). Almost every tested bacterial species (99.99%; Table 1) contains at least one TR gene at the species or genus level. On average, genus- and species-specific genes account for 4.5% of all genes in each genome, which is in accordance with early estimates of TRGs in bacterial species (1–16%) (1).

To estimate the extent to which the separation of a species in a taxonomy tree of bacteria affects its TR gene count, we calculated an IIO, which represents the distance between a certain species or genus from the nearest other species or genus in a phylogenetic tree of bacteria (see Materials and methods). We observed a moderate positive correlation (Pearson's $r = 0.38$, $P \approx 0$) between the number of TRGs and the IIO distance for the genus-level TRGs (Figure 1a), indicating a minor influence ($r^2 = 14\%$) of available genomic data (not sufficient taxonomic saturation) on the TRG content. However, the number of species-level TR genes does not seem to be impacted by the IIO distance ($r = -0.01$; Figure 1b).

The highest number of genus-specific TRGs was identified in bacteria belonging to the *Eperythrozoon_B* genus. These bacteria are assigned to the *Mycoplasmoidaceae* family and are responsible for a rare, sporadic, non-contagious,

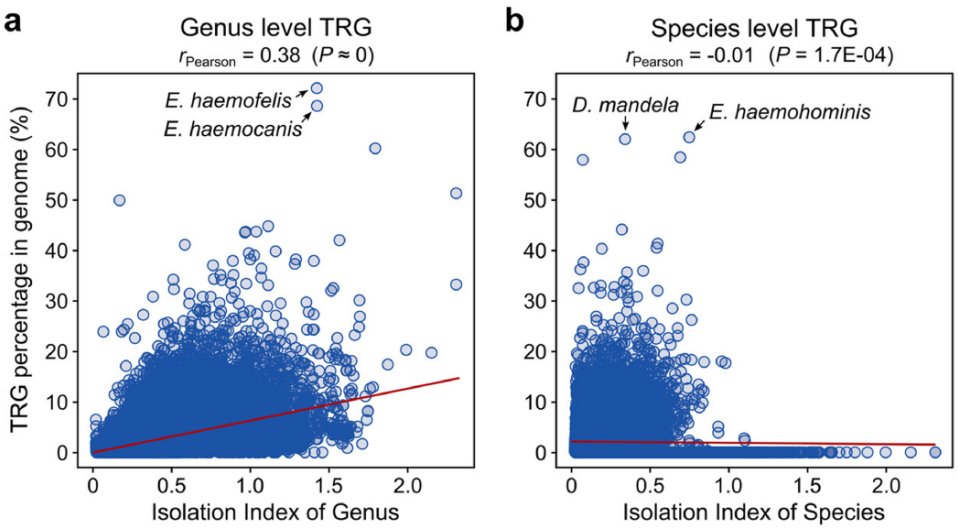

Figure 1. Correlation between the TRG number and IIO for (**a**) genus- and (**b**) species-specific genes.

blood-borne disease in ruminants (30). Two species, namely *E. haemofelis* (NCBI assembly accession: GCF_000200735) and *E. haemocanis* (GCF_000238995), contain 72.1% and 68.6% of genus-level TRGs, respectively (Figure 1a). The number of species-specific TRGs is relatively low for these species, with a maximum of 5.2% in *E. haemofelis*. Interestingly, among the four species classified in this genus, one (*E. haemohominis*) contains the largest proportion of species-specific genes from all other bacterial species (62.4%, Figure 1b). Following closely, the second species with the significant proportion of species-level TRGs is *Didemnitutus mandela* (GCA_002591725), a bacterial symbiont of marine tunicates. Despite the relatively low taxonomic isolation of *D. mandela* (IIO = 0.30), 62.0% of its genes (1898 out of 3060) were classified as species-specific (Figure 1b), and an additional 16 (0.5%) were shared only within the *Didemnitutus* genus. The exceptionally high number of species-level TRGs is supported by a recent study of the *D. mandela* genome (31), which highlighted an extremely high number of genes without detectable homologs, suggesting that the specific environment and distinct lifestyle of this bacterium resulted in this unusual accumulation of the species-specific genes. Although these species-specific sequences are annotated in the GTDB as protein-coding genes, it should be noted that some of them might potentially be pseudogenes, as previously suggested (31).

We did not find any TR genes at the species or genus levels in the genomes of 11 species. Eight of these species belong to the *Enterobacteriaceae* family and mostly represent bacteria from unassigned genus GCA_012562765, and two species of aphids (*Serratia symbiotica*) and blood sucking fly *Lipoptena cervi* (Candidatus *Arsenophonus lipoptenae*). Three other bacterial species are classified as uncultured *Pelagibacteraceae* bacteria and belong to the *Pelagibacter_A* and *MED727* genera. Since the species isolation index IIO for these genomes varies greatly (between 0.06 and 0.41 for species and from 0.31 to 0.99 for genus), it is very difficult to assign the effect of these predictions to adequate taxonomic saturation.

Properties of the TRG protein sequences

Since the protein sequences in the GTDB lack annotation descriptions, we mapped the sequences of the predicted TRGs to the corresponding NCBI protein records. Due to missing protein sequence data for some NCBI genome assemblies, only 32% of the predicted TRGs could be mapped to the identical sequences deposited at NCBI. To gain insight into the functional annotations of these proteins, we used a word frequency strategy because the information is not available in any formal annotation system like Gene Ontology or Kyoto Encyclopedia of Genes and Genomes. As expected, the most repeated descriptors for TRGs present in 49% TRG proteins include 'hypothetical', 'uncharacterized', 'partial', 'unknown' and 'putative' (Supplementary Table S1). The second layer of functional characterization includes general, non-specific descriptors (1%) such as 'plasmid', 'domain-containing', 'conserved' or 'membrane'. Finally, the less frequent descriptors suggest TRGs may be involved in processes such as 'regulation', 'transcription', 'export' and 'transport'. Together, these annotations align well with the postulated role of TRGs in adaptation and specific organismal functions. Notably, one of the middle scoring description words of the TRG records includes the keyword 'orfan', highlighting the taxonomically restricted nature of the proteins. Together, these annotations align well with the postulated role of TRGs in adaptation and specific organismal functions.

To further characterize the TRG proteins, we analyzed four common properties of these sequences: length, disorder and complexity levels and aggregation potential. When compared to the properties of a randomly selected group of similar size from all non-TRG bacterial proteins, we observed that the TRG sequences are on average of smaller size (with a median of 66 and 86 for species- and genus-specific genes, respectively, compared to 281 for all other proteins; Figure 2a), are less complex (median ~3.8 for TRGs versus ~4.0 for background proteins; Figure 2b), show higher disorder levels (median 0.16 and 0.18 for genus- and species-specific TRGs, respectively, and 0.06 for randomly selected proteins; Figure 2c) and show a slightly higher aggregation potential for genus-specific TRGs

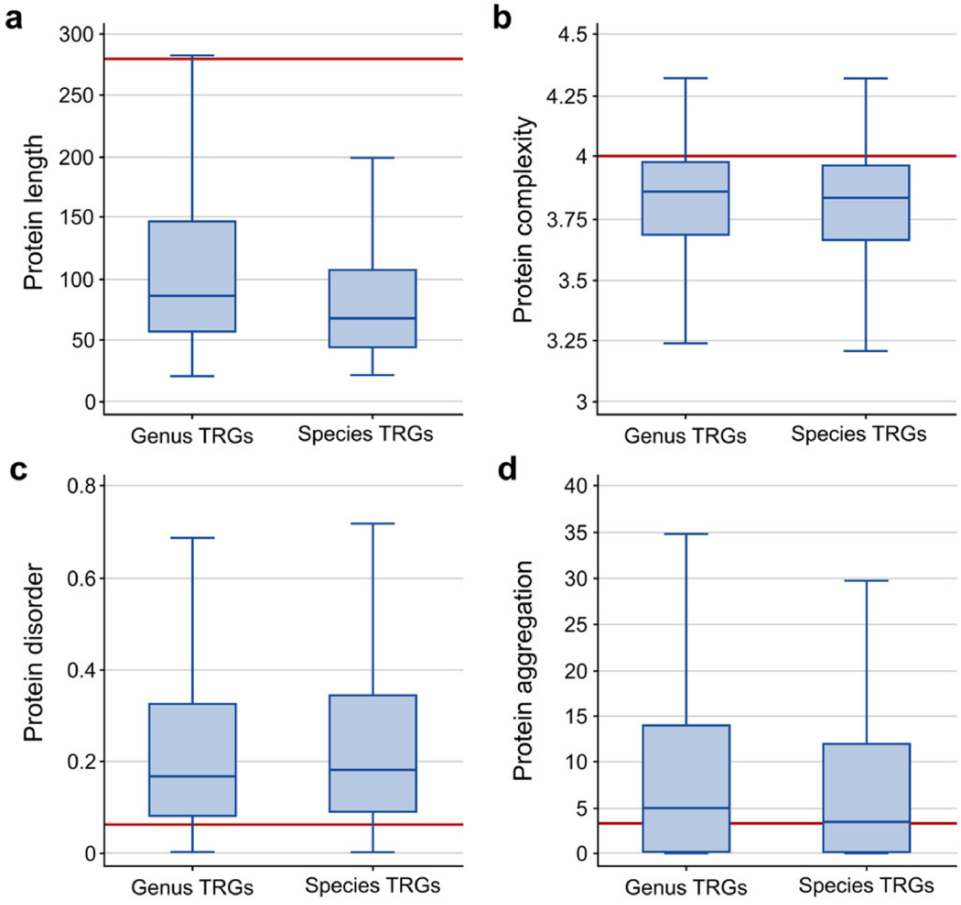

Figure 2. Comparison of protein sequence properties between the TR genes at the genus and species levels: (a) length, (b) complexity, (c) disorder, and (d) aggregation. The horizontal line indicates the median value calculated from a random sample of the non-TRG bacterial proteins.

(median 4.8 versus 3.3 measured as background; Figure 2d). The high aggregation potential of the genus-specific TRG sequences is a phenomenon noticed previously in *Bacillus* (21) and it may indicate that this group of genes has a unique evolutionary history.

Finally, we provide insights into the evolution of the TR genes in bacteria by clustering their sequences based on similarity (see Materials and methods). Although the majority of the TRG proteins (79%) do not form groups of evolutionary-related sequences (singleton clusters), we observed orthologs and paralogs on the genus level and paralogs on the species (single-genome) level. The largest cluster was found in genomes belonging to the *Pseudomonas_E* genus and it covers 2932 genus-level proteins. This cluster is composed of sequences coming from different species (orthologs) as well as encoded in the same genome (paralogs). Surprisingly, at the species level, we identified 283 sequences forming a group of homologous sequences in *Arsenophonus sp000757905*. It has to be noted that in this case, the sequences forming the cluster are very short, mainly 35 amino acids long.

Database interface and functionality

The landing page of the TRGdb database provides the summary statistics concerning the number of predicted TRGs and their general characteristics (Table 1). The database interface offers simple and fast exploration capabilities, including browsing by taxonomy and searching by keyword and/or the TRG sequence.

The keyword search interface offers initial access to a table with all the bacterial species, allowing users to look for bacteria taxa of interest. For convenience, the search box features an autocomplete functionality that suggests terms matching the user query. Selecting a taxon name results in filtering the table rows. This procedure can be progressively applied to the previous search results.

The browse view provides a hierarchical exploration of the TR genes through bacterial taxonomy based on the GTDB classification. The interactive interface allows users to expand branches of the bacteria tree and view the number of genus- and species-level TRGs associated with each node. It begins with the superkingdom bacteria and selecting taxon names limits the range of selected records. Direct access to the list of genomes (species) is provided by choosing the 'details' button (three dots icon). From that view, a user can access the genome-related statistics, including the list of the TR genes.

The TRG record is a collection of information about a selected sequence including external links to the information deposited at NCBI. However, since the TRG sequences are often poorly annotated and accompanied by limited information, the TRG records are constantly updated with new calculations using available tools and datasets. There are two ongoing projects that provide extra information about

the quality of predictions and biological properties of the TRG sequences deposited in TRGdb. One project focuses on the high-quality annotation of the TRG sequences in the *Bacillus* genus (21), and these superior records are labeled as 'high confidence' in the database. The other project is a high-throughput effort to annotate the expression status of the TR genes, only considering high-quality predictions. This project uses publicly available RNA-seq data sets that are specific to species with the TR genes and labels the TRG records with a positive expression status.

To enhance the search experience of TRGdb, we also provide the BLAST feature to search for the TR genes based on sequence similarity. This option is useful for users who want to quickly check if their proteins of interest are classified as TRGs.

Finally, for customizable access to TRGdb, we provide an application programming interface (API) allowing users to programmatically search and browse the database content and retrieve statistics on genome and the TRG records. The API of TRGdb has a dedicated help page listing all the functionalities. In addition, all the data that were used to construct TRGdb are also available for download including CSV files containing information on each bacterial genome and a TRG protein, and sequences in FASTA format of all TRG proteins at genus and species levels.

Discussion

TRGdb represents the most comprehensive and up-to-date database of TRGs in bacteria. We aim to further identify and characterize these genes, and TRGdb serves as a valuable starting point for this journey. The database provides updated information on marker sequences in the context of the latest bacterial taxonomy, including analysis of their sequence properties and access to available annotations. Additionally, we continuously update TRGdb with data from RNA-seq analysis to improve the prediction quality and functional characterization of these genes.

The identification of TRGs in bacteria is a complex task that presents several challenges. One of the main challenges is the quality of the taxonomic classification. The classification of bacteria is a very dynamic field with a variety of different phylogenetic approaches (32–35). Surprisingly, the advent of molecular and genomic data seems to have increased the variety of classifications rather than reducing the problem (36). Since the TRG identification solely depends on the taxonomic classification of analyzed organisms, we decided to adopt the well-accepted and standardized solution provided in the GTDB. Another challenge of TRG identification is the computational cost of searching for TRGs. Even when limiting the available data to one representative genome for one species, the current data volume of available proteins is counted in hundreds of millions. The well-accepted solution to search for sequence similarity during the TRG annotation is BLAST. However, even when executed on high computational power clusters, searches with BLAST would take months of computational time. To ensure the up-to-date status of our database, we have designed a simplified mode of the TRG annotation pipeline. The predictions are processed in three steps, where the first is designed to limit the search space for the application of the BLAST search. It should be noted that such a simplified identification procedure does not explore transitive evolutionary relationships, which may negatively influence the quality of some of the predictions.

The increasing number of TRGs is one of the greatest surprises in the field of bacterial genome sequencing (5). We also observe this trend in our database—the comparison of the first publicly available release of the TRGdb database with the previous version (based on GTDB release 207 from 8 April 2022) shows that the number of predicted TRG sequences has increased by 25% (the current 10.7 versus previous 8.5 million sequences). It has to be noted, however, that the previous release of the GTDB covered only 62 291 species, while the current version contains 80 789 species, and the mean number of TRGs per species between TRGdb releases seems constant (∼4%). This comparison demonstrates the dynamic nature of TRGs and the importance of keeping a regularly updated database with the most current information on bacterial TRGs.

Supplementary material

Supplementary material is available at *Database* online

Data availability

The TRGdb database along with documentation is available at http://combio.pl/trgdb.

Author contributions

A.Z. performed the calculations, analyzed the data and implemented the database and web interface. W.D. implemented the API and provided data from previous database releases. W.M.K. designed the study, analyzed data and wrote the manuscript.

Funding

W.M.K. is supported by National Science Center (2017/25/B/NZ2/00187).

Conflict of interest

None declared.

Acknowledgements

The computations were performed at the Poznan Supercomputing and Networking Center within grant numbers 312 and 528.

References

1. Wilson,G.A., Bertrand,N., Patel,Y. *et al.* (2005) Orphans as taxonomically restricted and ecologically important genes. *Microbiology*, **151**, 2499–2501.
2. Chen,S., Krinsky,B.H. and Long,M. (2013) New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.*, **14**, 645–660.
3. Gupta,A. and Sharma,V.K. (2015) Using the taxon-specific genes for the taxonomic classification of bacterial genomes. *BMC Genom.*, **16**, 1–9.
4. Khalturin,K., Hemmrich,G., Fraune,S. *et al.* (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.*, **25**, 404–413.

5. Wilson,G.A., Feil,E.J., Lilley,A.K. *et al.* (2007) Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. *PLoS One*, **2**, e324.
6. Fukuchi,S. and Nishikawa,K. (2004) Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res.*, **11**, 219–31, 311–313.
7. Jordan,I.K., Makarova,K.S., Spouge,J.L. *et al.* (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.*, **11**, 555–565.
8. Siew,N. and Fischer,D. (2003) Unravelling the ORFan Puzzle. *Comp. Funct. Genomics*, **4**, 432–441.
9. Siew,N. and Fischer,D. (2003) Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, **53**, 241–251.
10. Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.
11. Daubin,V. and Ochman,H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. Coli*. *Genome Res.*, **14**, 1036–1042.
12. Chin,K.-H., Ruan,S.-K., Wang,A.H.-J. *et al.* (2007) XC5848, an ORFan protein from *Xanthomonas campestris*, adopts a novel variant of Sm-like motif. *Proteins*, **68**, 1006–1010.
13. Zhou,K., Huang,B., Zou,M. *et al.* (2015) Genome-wide identification of lineage-specific genes within *Caenorhabditis elegans*. *Genomics*, **106**, 242–248.
14. Johnson,B.R. and Tsutsui,N.D. (2011) Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genom.*, **12**, 164.
15. Zile,K., Dessimoz,C., Wurm,Y. *et al.* (2020) Only a single taxonomically restricted gene family in the drosophila melanogaster subgroup can be identified with high confidence. *Genome Biol. Evol.*, **12**, 1355–1366.
16. Sollars,E.S.A., Harper,A.L., Kelly,L.J. *et al.* (2017) Genome sequence and genetic diversity of European ash trees. *Nature*, **541**, 212–216.
17. Toll-Riera,M., Castelo,R., Bellora,N. *et al.* (2009) Evolution of primate orphan proteins. *Biochem. Soc. Trans.*, **37**, 778–782.
18. Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.*, **3**, 504–510.
19. Prangishvili,D., Garrett,R.A. and Koonin,E.V. (2006) Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res.*, **117**, 52–67.
20. Siew,N., Azaria,Y. and Fischer,D. (2004) The ORFanage: an ORFan database. *Nucleic Acids Res.*, **32**, D281–D283.
21. Karlowski,W.M., Varshney,D. and Zielezinski,A. (2023) Taxonomically restricted genes in *Bacillus* may form clusters of homologs and can be traced to a large reservoir of noncoding sequences. *Genome Biol. Evol*, **15**, evad023.
22. Parks,D.H., Chuvochina,M., Rinke,C. *et al.* (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
23. Buchfink,B., Reuter,K. and Drost,H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
24. Camacho,C., Coulouris,G., Avagyan,V. *et al.* (2009) BLAST: architecture and applications. *BMC Bioinform.*, **10**, 1–9.
25. Mészáros,B., Erdos,G. and Dosztányi,Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.
26. Fernandez-Escamilla,A.-M., Rousseau,F., Schymkowitz,J. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
27. Virtanen,P., Gommers,R., Oliphant,T.E. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
28. Mistry,J., Chuguransky,S., Williams,L. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
29. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
30. Delano,M.L., Mischler,S.A. and Underwood,W.J. (2002) Biology and diseases of ruminants: sheep, goats, and cattle. In: Fox J.G (ed.) *Laboratory Animal Medicine*. Academic Press (Elsevier), Cambridge, Massachusetts, United States, pp. 519–614.
31. Lopera,J., Miller,I.J., McPhail,K.L. *et al.* (2017) Increased biosynthetic gene dosage in a genome-reduced defensive bacterial symbiont. *Msystems*, **2**, e00096-17.
32. Martinez-Gutierrez,C.A., Aylward,F.O. and Battistuzzi,F.U. (2021) Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol. Biol. Evol.*, **38**, 5514–5527.
33. Doolittle,W.F. and Papke,R.T. (2006) Genomics and the bacterial species problem. *Genome Biol.*, **7**, 116.
34. Tsang,A.K.L., Lee,H.H., Yiu,S.-M. *et al.* (2017) Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Sci. Rep.*, **7**, 4536.
35. Gupta,R.S. and Griffiths,E. (2002) Critical issues in bacterial phylogeny. *Theor. Popul. Biol.*, **61**, 423–434.
36. Som,A. (2015) Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.*, **16**, 536–548.