Published online 23 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D489–D493***
doi:10.1093/nar/gkn689

*rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea

Zarraz May-Ping Lee, Carl Bussema III and Thomas M. Schmidt*

Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

Received September 14, 2008; Accepted September 24, 2008

ABSTRACT

A dramatic exception to the general pattern of single-copy genes in bacterial and archaeal genomes is the presence of 1–15 copies of each ribosomal RNA encoding gene. The original version of the Ribosomal RNA Database (*rrn*DB) cataloged estimates of the number of 16S rRNA-encoding genes; the database now includes the number of genes encoding each of the rRNAs (5S, 16S and 23S), an internally transcribed spacer region, and the number of tRNA genes. The *rrn*DB has been used largely by microbiologists to predict the relative rate at which microbial populations respond to favorable growth conditions, and to interpret 16S rRNA-based surveys of microbial communities. To expand the functionality of the *rrn*DB (http://ribosome.mmg.msu.edu/rrndb/index.php), the search engine has been redesigned to allow database searches based on 16S rRNA gene copy number, specific organisms or taxonomic subsets of organisms. The revamped database also computes average gene copy numbers for any collection of entries selected. Curation tools now permit rapid updates, resulting in an expansion of the database to include data for 785 bacterial and 69 archaeal strains. The *rrn*DB continues to serve as the authoritative, curated source that documents the phylogenetic distribution of rRNA and tRNA genes in microbial genomes.

INTRODUCTION

Ribosomes play a central role in every form of life by catalyzing the mRNA-dependent synthesis of proteins from amino acids. Crystal structures of this ribonucleoprotein complex reveal a catalytic center that consists primarily of ribosomal RNAs (rRNA) (1). Due to the conserved function of ribosomes, the 3D structure of the rRNAs is highly constrained, with regions of strong primary sequence conservation interspersed with variable regions. These characteristics make the molecule ideal for establishing the evolutionary relatedness of organisms, and for culture-independent molecular surveys of microbial communities.

As the applications of phylogenetic analyses and molecular surveys have expanded in microbiology, databases of aligned rRNA gene sequences including SILVA (2), the Ribosomal Database Project (3) and Greengenes (4) were created to assist in sequence analysis. However, these databases do not include information about a crucial characteristic of rRNA genes that influences molecular surveys: the number of rRNA genes per genome.

Genes encoding the 16S rRNA (*rrs*), 23S rRNA (*rrl*) and 5S rRNA (*rrf*) are typically arranged into an operon (*rrn* operon), with an internally transcribed spacer (ITS) between the 16S and 23S rRNA genes that is also used to discriminate amongst closely related organisms. The number of *rrn* operons ranges from one to 15 per genome. This redundancy must be considered in studies that measure the abundance of rRNA genes, especially techniques such as terminal restriction fragment length polymorphism (tRFLP), denaturing gradient gel electrophoresis (DGGE) and quantitative PCR (5). Due to redundancy of the rRNA genes in some organisms, the measured abundance of an rRNA gene might be attributed to few organisms with many rRNA genes or many organisms with few rRNA genes.

An additional benefit of knowing the *rrn* copy number of an organism is derived from the positive correlation between the number of rRNA genes in an organism's genome and the capacity of that organism to respond to favorable growth conditions (6,7). This relationship suggests that the number of rRNA genes copy number reveals the life history of an organism, where organisms with few *rrn* operons tend to be slow growing organisms that can utilize resources efficiently, while those with many *rrn* operons grow more rapidly in response to favorable growth conditions but with less efficient use of resources. In addition, microbes with few *rrn* operons tend to be oligotrophic, i.e. capable of growth in low-nutrient environments (8,9).

The value in linking life histories to the number of rRNA genes prompted the compilation of available

*To whom correspondence should be addressed. Tel: +1 517 884 5400; Fax: +517 353 8957; Email: tschmidt@msu.edu

information on 16S rRNA-encoding genes in the first iteration of the *rrn*DB (10). Microbiologists used the database as a reference for estimation of an organism's capacity to respond to favorable growth conditions, and to interpret abundance data from molecular surveys. In particular, the database has become critical for studies using quantitative PCR to enumerate bacteria and archaea in the environment (11).

Increasing use of the *rrn*DB and expansion of the regions of the *rrn* operon that are now included in molecular surveys motivated its redesign and expansion. The *rrn*DB is now based on a relational database that includes information on redundancy of all rRNA genes (*rrs*, *rrl*, *rrf)* and the number of ITS regions. The number of transfer RNA (tRNA) genes per genome has also been added to the database because it varies with *rrs* copy number. Users now have access to expanded queries that include dynamic calculation of average gene copy numbers for group of organisms selected, and additional search and sorting features. Curatorial tools have also been added to facilitate updating. As a result, the number of bacterial and archaeal strains included in the *rrn*DB has more than doubled, with 785 bacterial and 69 archaeal strains, and is now being updated regularly.

DATABASE DESCRIPTION

The redesigned website and database are accessible on the WWW at http://ribosome.mmg.msu.edu/rrndb. Entries in the database not only consist primarily of data from sequenced genomes, but also include data from strains in which the number of *rrn* genes has been determined through other methods. Among unique species in the database, 40% of bacteria have either one or two copies of the 16S rRNA gene (Figure 1). Bacterial species with eight or more 16S rRNA genes make up 11% of the unique entries, and they consist of bacteria in the phylum Firmicutes or the class γ-Proteobacteria; there is a single β-Proteobacteria, *Chromobacterium violaceum*, currently in this grouping. Two organisms with 15 *rrn* operons are know—*Clostridium paradoxum* and *Photobacterium profundum*. The range of *rrn* genes in archaea is smaller, with one to four copies of the 16S rRNA gene. More than half (57%) of sequenced archaeal genomes have a single copy of each of the *rrn* genes. Archaea with two or more 16S rRNA genes are all from the phylum Euryarchaeota.

There are multiple ways to access entries in the database: users can browse through the entire database; 'search by keyword', which is based on an organism's name or strain designation or simply by the number of 16S rRNA genes; 'search by taxonomy' allows users to select entries within a particular taxonomic level from a pull-down menu; or through combinations of these searches. The new search features also include dynamic calculation of average gene copy number for any subset of organisms selected in a search.

Results from database searches are presented in a table that appears below the search form (Figure 2). For each entry, the table presents the genus, species, strain designation and copy numbers for 16S, 23S, 5S rRNA, tRNA genes and the ITS. The majority of the entries will have the same number of *rrs, rrl, rrf* and the ITS, which might be expected because ribosomes are made up of a single transcript from each gene. However, 23.6% of genomic bacterial entries have unequal copies of the rRNA genes, due mainly to additional copies of the 5S rRNA gene. Other variations include Borrelia sp. which maintains two copies of the 23S–5S rRNA genes and one copy of the 16S rRNA gene encoded separately on the genome, and *Thermobispora bispora* which has four copies of 16S rRNA gene, three copies of 23S rRNA gene and only two copies of 5S rRNA gene (12,13).

For convenience, the entries in the result table can be sorted according to any column by clicking on the column

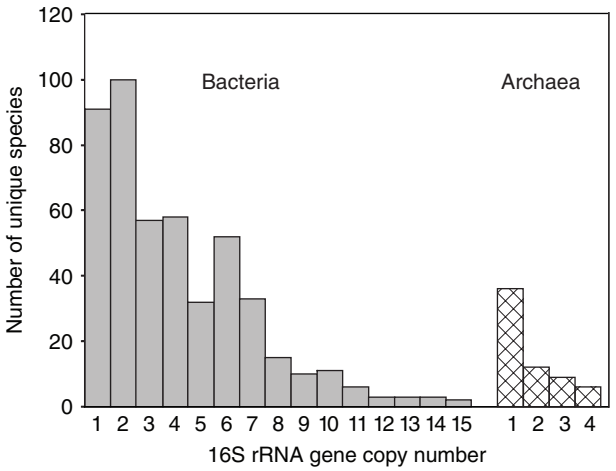

Figure 1. The number of 16S rRNA genes in bacterial and archaeal genomes. The analysis was performed on 476 bacterial species (gray bars) and 63 archaeal species (checkered bars).

Genus ↓	Species	Strain designation	16S	ITS	23S	5S	tRNA
Bdellovibrio	*bacteriovorus*	HD100	2	2	2	2	36
Desulfovibrio	*desulfuricans*	G20	4	4	4	4	66
Desulfovibrio	*vulgaris*	Hildenborough	5	5	5	6	68
Desulfovibrio	*vulgaris*	DP4	5	NA	5	6	68
Prosthecochloris	*vibrioformis*	DSM 265	1	1	1	1	45
Vibrio	*cholerae*	569B (Inaba)	NA	NA	7	NA	NA
Vibrio	*cholerae*	O395	8	8	8	9	96
Vibrio	*cholerae*	N16961	8	8	8	9	98
Vibrio	*fisheri*	ES114	12	12	12	12	119
Vibrio	*harveyi*	ATCC BAA-1116	11	10	10	11	121
Vibrio	*natriegens*	ATCC 14048	13	NA	NA	NA	NA
Vibrio	*parahaemolyticus*	RIMD 2210633	11	11	11	12	126
Vibrio	*vulnificus*	CMCP6	9	9	9	10	111
Vibrio	*vulnificus*	YJ106	9	9	9	10	112
Average			7.54	7.18	7	7.67	88.83

Figure 2. A screenshot of the result table from the *rrn*DB using 'search by keyword' for 'Vibrio'. The average gene copy number is presented at the end of the table. The dark gray highlighted column indicates that the table is sorted according to the genus name. NA indicates that information for the particular gene is 'not available'.

heading (Figure 2). A major addition to the database is the capacity for dynamic calculation of the average copy number for any collection of organisms listed in the result table: the arithmetic average is presented for each gene at the end of the table. This feature will be particularly useful for researchers using quantitative PCR to enumerate the abundance of a specific group of organisms. The number of 16S rRNA genes is typically constant among different strains of the same species, but in ~5% of bacterial species in the database, the number varies by one for different strains. For closely related species, the number of *rrn* genes per genome is often similar, but it is not entirely consistent with phylogenetic relationships. For instance, strains of both γ-proteobacteria and clostridia maintain from 1 to 15 copies.

Detailed information for each organism in the table can be viewed by clicking on the strain designation. Information will be presented below the results table. It includes organismal taxonomy, copy number for each gene and ITS, accession number for the gene entries in Genbank, genome size, genome accession number in Genbank, reference link to Entrez Pubmed and a comment section. For organisms with multiple chromosomes, the allocation of both rRNA and tRNA genes into each chromosome is described. The comment section also specifies the method used to determine the number of rRNA genes for entries not from genomic sequences. The two most common alternative methods for estimating the number of rRNA genes are Southern hybridization with rRNA gene or ITS-specific probes, and digestion with the restriction endonuclease ICeu1, which has recognition sites only in the 23S rRNA gene. The protocol for Southern hybridization method is provided in the website in the 'About *rrn*DB' section.

The new *rrn*DB also catalogues the number of internally transcribed spacer and tRNA genes per genome. The inclusion of the number of ITS helps capture organisms whose rRNA genes are not arranged in an operon, such as Leptospira sp., Thermoplasma sp. and *Nanoarchaeum equitans* (14–16). The rRNA genes of these organisms are separated on the chromosome and each under the control of their own promoter. The ITS region is increasingly used for diversity studies and since intragenomic heterogeneity increases with *rrn* operon copy number, the number of ITS region per genome will become more important in analyzing richness measures (17). Furthermore, organisms with multiple ITS can provide the heterogeneity required to differentiate organisms at subspecies level using restriction analysis (18),

Amino-acylated tRNAs are substrates for ribosome-mediated protein synthesis. When selection favors an increased number of *rrn* genes to synthesize ribosomes more quickly, the rate of protein synthesis can only be increased if there is a corresponding increase in the production of tRNAs. A positive correlation between the abundance or tRNA and rRNA genes has been documented previously (19). A significant positive correlation is maintained in an analysis that is expanded to include 590 bacterial genomes (Figure 3). As expected, *Photobacterium profundum*, which has the highest number of 16S rRNA genes also has the highest number of total tRNA genes (20).

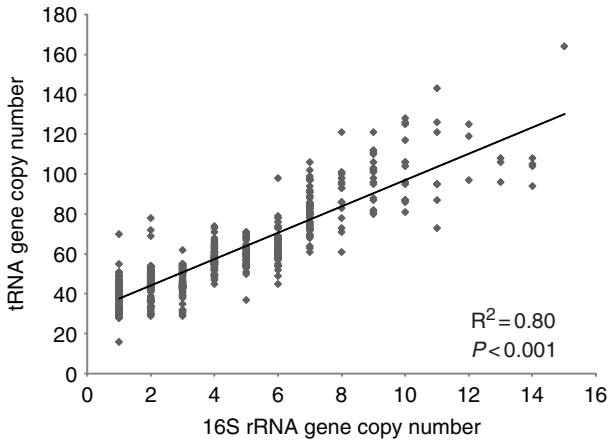

Figure 3. Correlation between the total number of tRNA genes and 16S rRNA genes in bacterial genomes. The data are gathered from sequenced genomes of 590 bacterial species.

DATA CURATION

The *rrn*DB is equipped with a password protected entry form that is accessible from the WWW, allowing curators to update the database online at anytime. The authors currently handle curation and maintenance of the database. Genomic data are obtained from the NCBI Microbial Genomes database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and the J. Craig Venter Institute Comprehensive Microbial Resource database (http://cmr.jcvi.org/tigr-scripts/CMR). Nongenomic data are obtained from literature searches, and these original references are maintained along with new genomic data and references as data from genome sequences are added. Taxonomic classifications are adopted from NCBI Entrez Taxonomy. The database is updated at least monthly.

The website has a 'Contact Us' form for users to alert the curators to new data for entry, ask questions about the site or provide suggestions for improving the *rrn*DB. The six most recent entries added to or updated in the database are listed in the left-side corner of the main page. Any changes in features are documented in the 'news and updates' section of the main page.

DATABASE MANAGEMENT SYSTEM

The *rrn*DB website is powered by PHP 5, MySQL 5, Apache 2 and runs on Mac OS X Server 10.4. The choice of programming language was made to facilitate ease of development and compatibility with available hardware. Except for the server operating system, all of these products are freely available under open source licenses, and have strong community support and a long history of integration.

The MySQL database is designed for speed and scalability: as new strains are added or updated through the administrative interface, the database size will grow, but by separating data out for as much normalization as possible, the growth will be reduced to a minimum.

Database indexes and table structure make it fast to search by keyword or taxonomy classification.

The front-end website for users is designed for ease of use and utilizes AJAX technologies to make dynamic searching possible, including multi-level taxonomy, and seamlessly filters a list of matching strains when criteria are entered. Sorting of the result table is done by client-side JavaScript with no additional load on the server.

Another key feature of the front-end website is XHTML and CSS design, which makes it well-suited to technologies such as screen readers and other products that help make the site accessible to persons with disabilities. When such technologies do not have the capability to use JavaScript, the site automatically falls back to a version that will work in any browser, with no additional steps required by the user.

For site administrators, a graphical interface facilitates curation. Drop-down lists with journal names, taxonomy classifications, the ability to add multiple citations or chromosomes at once, and other fields make it simple and fast to enter or update data, and changes are immediately live on the site with no need to leave the browser. When data for new strains are entered or existing data are updated, timestamp fields in the database are updated, making it easy to search for new or changed information.

Overall, the site design emphasizes ease of use while still providing useful options for end users and efficient data entry and maintenance for administrators. It was our goal in designing this site that it should be usable by researchers for many years to come without needing to involve IT personnel for more than routine maintenance, and in the year it has been operational since its redesign, fewer than 10 hours have been spent by any IT personnel, either programmers or systems administrators, giving us confidence that this site is sustainable.

FUTURE PLANS

One planned addition to the *rrn*DB is information on intragenomic heterogeneity of rRNA genes. Although intragenomic gene conversion amongst copies of *rrn* genes maintains nearly identical sequences (21), differences between copies of rRNA genes are known. The highest intragenomic heterogeneity currently documented is 7.2% sequence divergence between 16S rRNA genes found in *Thermobispora bispora* (12). Documenting this variability will help estimate the contribution of intragenomic variation to the microheterogeneity that is frequently observed in environmental clone libraries of rRNA genes (22–24). The motivation to develop the *rrn*DB is to understand the evolutionary implication of redundancy of *rrn* genes, and so we also plan to expand the database to include genomic characteristics (e.g. gene content, pathway preferences) that correlate with the number of rRNA and tRNA genes.

ACKNOWLEDGEMENT

We thank J.A. Klappenbach for the initial compilation of data.

FUNDING

National Science Foundation (IOS 0421900 to T.M.S.). Funding for open access charge: National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Schuwirth,B.S., Borovinskaya,M.A., Hau,C.W., Zhang,W., Vila-Sanjurjo,A., Holton,J.M. and Cate,J.H. (2005) Structures of the bacterial ribosome at 3.5 A resolution. *Science*, **310**, 827–834.
2. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W., Peplies,J. and Glockner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
3. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Bandela,A.M., Cardenas,E., Garrity,G.M. and Tiedje,J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.*, **35**, D169–D172.
4. Desantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
5. Crosby,L.D. and Criddle,C.S. (2003) Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques*, **34**, 790–794, 796, 798 passim.
6. Klappenbach,J.A., Dunbar,J.M. and Schmidt,T.M. (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, **66**, 1328–1333.
7. Stevenson,B.S. and Schmidt,T.M. (1998) Growth rate-dependent accumulation of RNA from plasmid-borne rRNA operons in Escherichia coli. *J. Bacteriol.*, **180**, 1970–1972.
8. Eichorst,S.A., Breznak,J.A. and Schmidt,T.M. (2007) Isolation and characterization of soil bacteria that define Terriglobus gen. nov., in the phylum Acidobacteria. *Appl. Environ. Microbiol.*, **73**, 2708–2717.
9. Cavicchioli,R., Ostrowski,M., Fegatella,F., Goodchild,A. and Guixa-Boixereu,N. (2003) Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of *Sphingopyxis alaskensis* (formerly *Sphingomonas alaskensis*). *Microb. Ecol.*, **45**, 203–217.
10. Klappenbach,J.A., Saxman,P.R., Cole,J.R. and Schmidt,T.M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res.*, **29**, 181–184.
11. Einen,J., Thorseth,I.H. and Ovreas,L. (2008) Enumeration of archaea and bacteria in seafloor basalt using real-time quantitative PCR and fluorescence microscopy. *FEMS Microbiol. Lett.*, **282**, 182–187.
12. Wang,Y., Zhang,Z. and Ramanan,N. (1997) The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. *J. Bacteriol.*, **179**, 3270–3276.
13. Ojaimi,C., Davidson,B.E., Saint Girons,I. and Old,I.G. (1994) Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of the Lyme disease spirochaetes *Borrelia burgdorferi*, *B. garinii* and *B. afzelii*. *Microbiology*, **140**(Pt 11), 2931–2940.
14. Fukunaga,M., Masuzawa,T., Okuzako,N., Mifuchi,I. and Yanagihara,Y. (1990) Linkage of ribosomal RNA genes in Leptospira. *Microbiol. Immunol.*, **34**, 565–573.
15. Ree,H.K. and Zimmermann,R.A. (1990) Organization and expression of the 16S, 23S and 5S ribosomal RNA genes from the archaebacterium *Thermoplasma acidophilum*. *Nucleic Acids Res.*, **18**, 4471–4478.
16. Waters,E., Hohn,M.J., Ahel,I., Graham,D.E., Adams,M.D., Barnstead,M., Beeson,K.Y., Bibbs,L., Bolanos,R., Keller,M. *et al.* (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl Acad. Sci. USA*, **100**, 12984–12988.
17. Stewart,F.J. and Cavanaugh,C.M. (2007) Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria. *J. Mol. Evol.*, **65**, 44–67.

18. Malimas,T., Yukphan,P., Takahashi,M., Potacharoen,W., Tanasupawat,S., Nakagawa,Y., Tanticharoen,M. and Yamada,Y. (2006) Heterogeneity of strains assigned to *Gluconobacter frateurii* Mason and Claus 1989 based on restriction analysis of 16S-23S rDNA internal transcribed spacer regions. *Biosci. Biotechnol. Biochem.*, **70**, 684–690.
19. Dethlefsen,L. and Schmidt,T.M. (2007) Performance of the translational apparatus varies with the ecological strategies of bacteria. *J. Bacteriol.*, **189**, 3237–3245.
20. Vezzi,A., Campanaro,S., D'Angelo,M., Simonato,F., Vitulo,N., Lauro,F.M., Cestaro,A., Malacrida,G., Simionati,B., Cannata,N. *et al.* (2005) Life at depth: *Photobacterium profundum* genome sequence and expression analysis. *Science*, **307**, 1459–1461.
21. Hashimoto,J.G., Stevenson,B.S. and Schmidt,T.M. (2003) Rates and consequences of recombination between rRNA operons. *J. Bacteriol.*, **185**, 966–972.
22. Shaw,A.K., Halpern,A.L., Beeson,K., Tran,B., Venter,J.C. and Martiny,J.B.H. (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ. Microbiol.*, **10**, 2200–2210.
23. Acinas,S.G., Marcelino,L.A., Klepac-Ceraj,V. and Polz,M.F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.*, **186**, 2629–2635.
24. Acinas,S.G., Klepac-Ceraj,V., Hunt,D.E., Pharino,C., Ceraj,I., Distel,D.L. and Polz,M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, **430**, 551–554.