D168–D170 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh004

RPG: the Ribosomal Protein Gene database

Akihiro Nakao, Maki Yoshihama and Naoya Kenmochi*

Department of Biotechnology, Research Center for Frontier Bioscience, Miyazaki University, 5200 Kihara, Kiyotake, Miyazaki 889-1692, Japan

Received August 6, 2003; Accepted August 7, 2003

ABSTRACT

RPG (http://ribosome.miyazaki-med.ac.jp/) is a new database that provides detailed information about ribosomal protein (RP) genes. It contains data from humans and other organisms, including *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Methanococcus jannaschii* and *Escherichia coli*. Users can search the database by gene name and organism. Each record includes sequences (genomic, cDNA and amino acid sequences), intron/exon structures, genomic locations and information about orthologs. In addition, users can view and compare the gene structures of the above organisms and make multiple amino acid sequence alignments. RPG also provides information on small nucleolar RNAs (snoRNAs) that are encoded in the introns of RP genes.

INTRODUCTION

The ribosome is a universal and essential catalyst of protein synthesis in all organisms. Because of the fundamental role played by the ribosome, its structure and function have been significantly conserved during evolution. In eukaryotes, the ribosome is composed of four RNA molecules (rRNAs) and about 80 different proteins (RPs) (1), which are each present as a single copy with the exception of two proteins. The genes encoding rRNAs are clustered at a few sites in the eukaryotic genome, whereas the genes encoding RPs are widely dispersed (2). rRNA sequences have been extensively studied and data from thousands of organisms are now available (The Ribosomal Database Project; http://rdp.cme.msu.edu/) (3). On the other hand, little has been done for RP genes despite the acute need for a dedicated database.

In mammalian cells, each RP is encoded by a single gene but this gene generates a large number (10–20 copies) of silent, processed pseudogenes (4). This has hampered the mapping and sequence analysis of the functional RP genes. Through persistent efforts, however, we have completely mapped and sequenced most of these functional genes in the human genome (2,5,6). This has enabled us to compare their gene structures with those of other eukaryotes whose genome sequences have already been sequenced. Here we present a new database containing information about RP genes from various species, which we hope will constitute a valuable resource for biomedical research and a powerful tool for comparative studies of gene evolution.

DATA SOURCES AND FORMATS

Most of the human RP gene sequences in the database were determined in our previous study (6). Others were collected from the DDBJ/EMBL/GenBank databases. Sequences from other eukaryotes (*Drosophila melanogaster, Caenorhabditis elegans* and *Saccharomyces cerevisiae*) were obtained primarily by a BLAST search of the following databases using the human cDNA sequences as queries: FlyBase at http://flybase.org/ (7), WormBase at http://www.wormbase.org/ (8) and Saccharomyces Genome Database (SGD) at http://www.yeastgenome.org/ (9), respectively. The archaea (*Methanococcus jannaschii*) RP gene sequences were collected from the NCBI Entrez FTP site, and the bacterial (*Escherichia coli*) RP gene sequences were from GenoBase at http://ecoli.aist-nara.ac.jp/. In addition to these sequence data, we have integrated different types of information about RP genes, including chromosomal positions, accession numbers, gene and CDS sizes, orthologs, snoRNAs and links to other public databases. These data were automatically written in RPG unique file format by using a MS-Excel VBA script. Graphical data showing intron/exon structures, translation start and stop sites, and snoRNA gene locations were also formatted by using the VBA script and integrated into the database. Users can access these files through a web site running a CGI program. We employed Clustal W 1.82 to align amino acid sequences among orthologs (10). The Clustal W outputs were formatted into HTML files after color shading of the sequences with a Perl script to show amino acid similarities. The number of current entries from various organisms is summarized in Table 1.

STRUCTURE AND INTERFACE

The data in the RPG database can be accessed in a variety of ways. Each entry can be searched by gene name or organism (Fig. 1A). For human RP genes, each entry is linked to a gene table or a chromosomal map position (5). For other genes, each entry is linked to an orthologous gene classification table, which includes all of the RP gene entries in this database. The main page for each gene provides the following information (Fig. 1B); (i) a schematic view of intron/exon structure, the translation start and stop sites, and positions of snoRNA genes if available, (ii) general information including the source organism, the gene name, the chromosomal localization, the

*To whom correspondence should be addressed. Tel: +81 985 859665; Fax: +81 985 851514; Email: kenmochi@post.miyazaki-med.ac.jp

accession number, the gene and CDS length, and the number of exons, (iii) a schematic view of human chromosomes with an indication of map position that links to NCBI Entrez MapView, (iv) links to orthologous gene entries, comparative views of orthologous gene structures, and protein sequence alignments, (v) links to snoRNA gene entries, and (vi) links to other public databases, e.g. GenBank, LocusLink, and NCBI Entrez Mapview. For snoRNA genes, especially, the entry page provides sequence, information about the host gene, the accession number, the rRNA modification site, and the modification type, for example 'Box C/D 2′-*O*-methylation' or 'Box H/ACA pseudouridylation' (Fig. 1D).

Table 1. Number of current entries from six organisms

	Large subunit	Small subunit
H.sapiens	47	33
D.melanogaster	58	41
C.elegans	55	32
S.cerevisiae	82	56
M.jannaschii	37	25
E.coli	33	21

FUTURE DEVELOPMENT

RPG will continue to expand by collecting additional data sets from other eukaryotes, including mouse, *Fugu*, *Arabidopsis* and *Schizosaccharomyces*. In addition to current cytoplasmic RPs, we will also include data from organelles, namely, mitochondria and chloroplasts, as these organelles possess their own ribosomes which have evolved from those of ancient bacteria.

ACKNOWLEDGEMENTS

The RPG database is supported by Grants-in-Aid for Scientific Research (158118, 14035103 and 15310135), the 21st Century COE Program (Life Science) and a fund for Research for the Future Program from the Japan Society for the Promotion of Science (JSPS) and Ministry of Education, Culture, Sports, Science and Technology (MEXT).

REFERENCES

1. Wool,I.G., Chan,Y.L. and Glück,A. (1996) Mammalian ribosomes: The structure and the evolution of the proteins. In Hershey,J.W.B.,

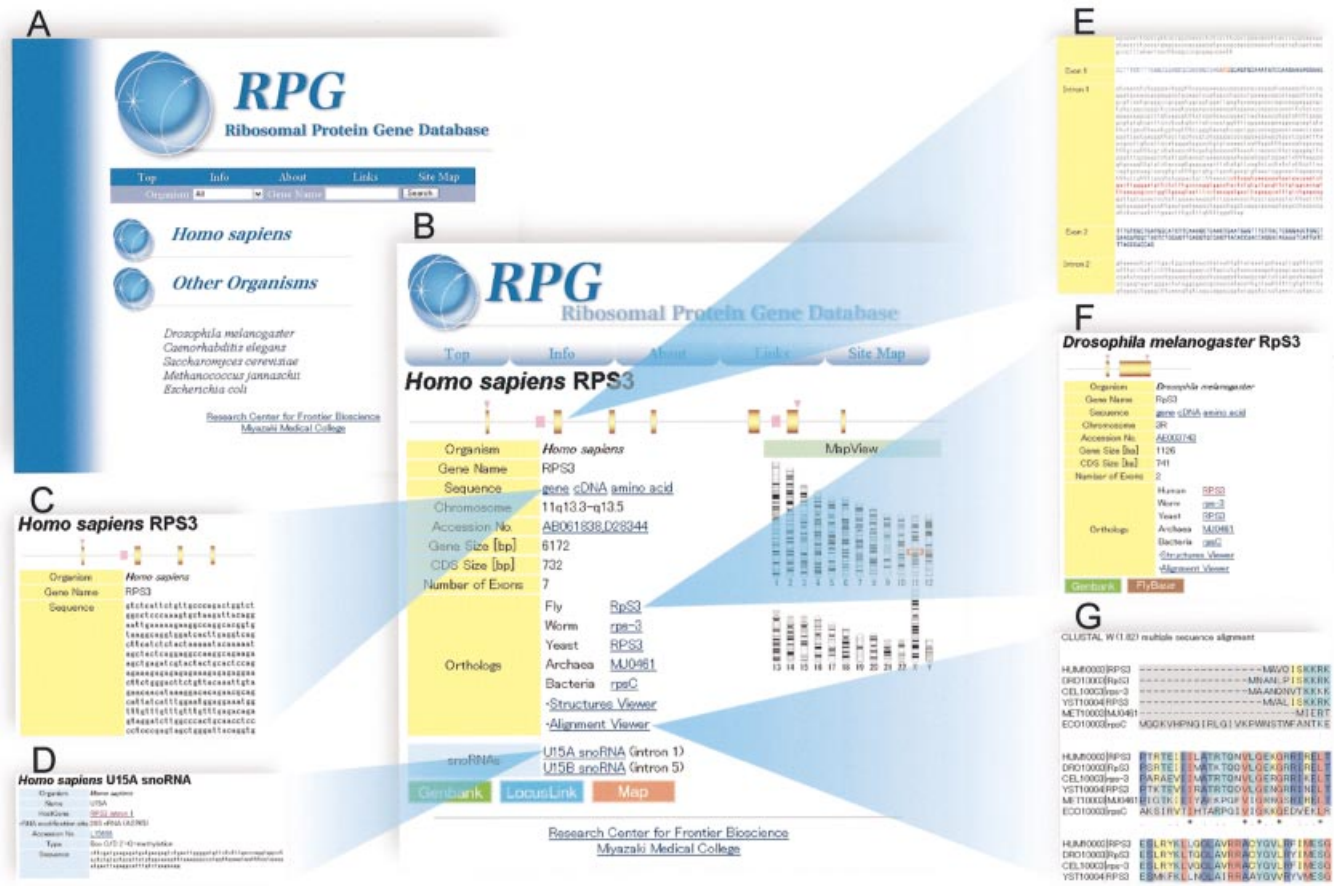

Figure 1. Examples of RPG data views. The top window (**A**) provides a search engine which users can use to search the database by gene name or organism and is linked to the human RP gene table (*Homo sapiens*) and to the orthologous gene table (*Other organisms*). The gene entry window (**B**) gives links to various sequences [genomic (**C**), cDNA and amino acid sequences], to the intron/exon sequence viewer (**E**), to GenBank entries, to orthologous gene entries (**F**), to Clustal W output (**G**) and to snoRNA gene entries (**D**).

Mathews,M.B. and Sonenberg,N. (eds), *Translational Control.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 685–732.
2. Kenmochi,N., Kawaguchi,T., Rozen,S., Davis,E., Goodman,N., Hudson,T.J., Tanaka,T. and Page,D.C. (1998) A map of 75 human ribosomal protein genes. *Genome Res.*, **8**, 509–523.
3. Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.*, **31**, 442–443.
4. Zhang,Z., Harrison,P. and Gerstein,M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.*, **12**, 1466–1482.
5. Uechi,T., Tanaka,T. and Kenmochi,N. (2001) A complete map of the human ribosomal protein genes: assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics*, **72**, 223–230.
6. Yoshihama,M., Uechi,T., Asakawa,S., Kawasaki,K., Kato,S., Higa,S., Maeda,N., Minoshima,S., Tanaka,T., Shimizu,N. and Kenmochi,N. (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.*, **12**, 379–390.
7. The FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
8. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al.* (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
9. Weng,S., Dong,Q., Balakrishnan,R., Christie,K., Costanzo,M., Dolinski,K., Dwight,S.S., Engel,S., Fisk,D.G., Hong,E. *et al.* (2003) *Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.
10. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.