Nucleic Acids Research, 2004, Vol. 32, Database issue ***D271–D272***
DOI: 10.1093/nar/gkh024

DEG: a database of essential genes

Ren Zhang, Hong-Yu Ou[1] and Chun-Ting Zhang[1,*]

Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China and [1]Department of Physics, Tianjin University, Tianjin 300072, China

Received July 26, 2003; Revised August 16, 2003; Accepted September 3, 2003

ABSTRACT

Essential genes are genes that are indispensable to support cellular life. These genes constitute a minimal gene set required for a living cell. We have constructed a Database of Essential Genes (DEG), which contains all the essential genes that are currently available. The functions encoded by essential genes are considered a foundation of life and therefore are likely to be common to all cells. Users can BLAST the query sequences against DEG. If homologous genes are found, it is possible that the queried genes are also essential. Users can search for essential genes by their function or name. Users can also browse and extract all the records in DEG. Essential gene products comprise excellent targets for antibacterial drugs. Analysis of essential genes could help to answer the question of what are the basic functions necessary to support cellular life. DEG is freely accessible from the website http://tubic.tju.edu.cn/deg/.

INTRODUCTION

Essential genes are genes that are indispensable to support cellular life. These genes constitute a minimal gene set required for a living cell. Therefore, the functions encoded by this gene set are essential and could be considered as a foundation of life itself (1,2). The definition of the minimal gene set needed to sustain a living cell is of considerable interest not only because it represents a fundamental question in biology, but also because it has much significance in practical use. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for antibacterial drugs (3).

DATABASE DESCRIPTION

The determination of the minimal gene set for bacteria has only been possible with the advent of the completion of many whole genome sequencing projects and the genome-scale gene inactivation technology. Consequently, essential genes have been determined in a number of different organisms. Essential genes have been determined in *Staphylococcus aureus* by an antisense RNA technique (4), in *Mycoplasma genitalium* by transposon mutagenesis (5), in *Haemophilus influenzae* by high-density transposon mutagenesis (6), in *Vibrio cholerae* by a mariner-based transposon (3), in yeast by genetic footprinting (7), and in *M.genitalium* and *H.influenzae* by comparative genomics (8).

We have constructed a Database of Essential Genes (DEG) that contains all the essential genes currently available. These genes include the essential genes identified in the genomes of *M.genitalium* (5), *H.influenzae* (6), *V.cholerae* (3), *S.aureus* (4), *Escherichia coli* and *Saccharomyces cerevisiae*. The essential genes in the *E.coli* genome were extracted from the web site http://magpie.genome.wisc.edu/~chris/essential.html, in which the essential genes are collected from a large number of related references. The essential genes in yeast genome were extracted from the yeast genome database (http://www.mips.biochem.mpg.de/proj/yeast), which is maintained by the Munich Information Center for Protein Sequences (9).

Each entry of essential genes has a unique DEG identification number, gene reference number, gene function and sequence. All information is stored and operated by using an open-source database management system, MySQL. Users can browse and extract all the records of these entries. In addition, users can also search DEG by gene function or name. Furthermore, we have installed the BLAST program locally. Therefore, users can BLAST the query sequences against all the essential gene sequences in DEG.

One of the applications is the prediction of essential genes based on homologous sequence search against DEG. The functions encoded by essential genes are considered to be generally essential for all cells (1). It is even believed that some basic functions and principles are common to all cellular life on this planet (10). Therefore, if the query sequences compared using BLAST have homologous genes in DEG, it is likely that the queried genes are also essential. In addition, by performing the BLAST search against DEG for all the protein-coding genes in a genome, it is possible to define the putative essential genes for the proteomes of newly sequenced genomes. However, caution must be taken in interpreting the BLAST results, since many essential genes are essential only in given growth conditions, such as in rich or minimal medium.

Another application is that by analyzing all the essential genes in DEG, some principles or regulations could be found to answer the question of what are the basic functions necessary to support cellular life. Those principles could lead to the development of new algorithms to predict essential genes. Some functions encoded by essential genes are expected, such as DNA replication, gene transcription, protein

*To whom correspondence should be addressed. Tel: +86 22 27402987; Fax: +86 22 27402697; Email: ctzhang@tju.edu.cn

synthesis, energy production and cell division. Some essential genes, however, are somewhat unexpected, such as Embden–Meyerhof–Parnas pathway genes and a purine biosynthesis gene (1). Analysis of DEG, which has all essential genes among different organisms, could help to classify those 'unexpected' essential genes.

Currently some essential gene projects are still ongoing and the identification of more essential genes is expected. DEG will be updated periodically to include more entries upon the availability of newly identified essential genes. We plan to integrate more information about experimental methods for each entry. In the next version of DEG, we also plan to include the essential genes of vertebrates, such as mouse. We welcome users' comments, corrections and new information, which will be used for updating.

DEG is freely available at the web site http://tubic.tju.edu.cn/deg/, and should be cited with the present publication as reference.

ACKNOWLEDGEMENTS

We are indebted to Professor Jingchu Luo for advice on database construction. The present study was supported in part by the 973 Project of China (grant 1999075606).

REFERENCES

1. Kobayashi,K., Ehrlich,S.D., Albertini,A., Amati,G., Andersen,K.K., Arnaud,M., Asai,K., Ashikaga,S., Aymerich,S., Bessieres,P. *et al.* (2003) Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA*, **100**, 4678–4683.
2. Itaya,M. (1995) An estimation of minimal genome size required for life. *FEBS Lett.*, **362**, 257–260.
3. Judson,N. and Mekalanos,J.J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.*, **18**, 740–745.
4. Ji,Y., Zhang,B., Van,S.F., Horn, Warren,P., Woodnutt,G., Burnham,M.K. and Rosenberg,M. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, **293**, 2266–2269.
5. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169.
6. Akerley,B.J., Rubin,E.J., Novick,V.L., Amaya,K., Judson,N. and Mekalanos,J.J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA*, **99**, 966–971.
7. Smith,V., Chou,K.N., Lashkari,D., Botstein,D. and Brown,P.O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science*, **274**, 2069–2074.
8. Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
9. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
10. Peterson,S.N. and Fraser,C.M. (2001) The complexity of simplicity. *Genome Biol.*, **2**, COMMENT2002.