Published online 30 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D455–D458***
doi:10.1093/nar/gkn858

DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes

Ren Zhang[1,*] and Yan Lin[2]

[1]Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060 and [2]Department of Physics, Tianjin University, Tianjin 300072, China

Received September 15, 2008; Revised October 14, 2008; Accepted October 16, 2008

ABSTRACT

Essential genes are those indispensable for the survival of an organism, and their functions are therefore considered a foundation of life. Determination of a minimal gene set needed to sustain a life form, a fundamental question in biology, plays a key role in the emerging field, synthetic biology. Five years after we constructed DEG, a database of essential genes, DEG 5.0 has significant advances over the 2004 version in both the number of essential genes and the number of organisms in which these genes are determined. The number of prokaryotic essential genes in DEG has increased about 10-fold, mainly owing to genome-wide gene essentiality screens performed in a wide range of bacteria. The number of eukaryotic essential genes has increased more than 5-fold, because DEG 1.0 only had yeast ones, but DEG 5.0 also has those in humans, mice, worms, fruit flies, zebrafish and the plant *Arabidopsis thaliana*. These updates not only represent significant advances of DEG, but also represent the rapid progress of the essential-gene field. DEG is freely available at the website http://tubic.tju.edu.cn/deg or http://www.essentialgene.org.

INTRODUCTION

Essential genes are those indispensable for the survival of an organism under certain conditions, and the functions they encode are therefore considered a foundation of life. Essential genes of an organism constitute its minimal gene set, which is the smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions (1–3). Determination of the minimal gene set for an organism addresses a conceptually important question: what are the basic functions needed to sustain a life form, and therefore the minimal-gene-set concept plays a key role in the emerging field, synthetic biology (4). Essential-gene studies are of interest for practical reasons as well. For instance, essential genes, because of lethality from their disruptions, are attractive targets of antibiotics (5). Some essential genes that are conserved across species are candidates for broad-spectrum drug targets, whereas those specific for one bacterium are candidates for species-specific ones.

In 2004, we constructed DEG 1.0, a database of essential genes (6). In the past five years, fueled by the accumulation of sequenced genomes, sophisticated genome-wide mutagenesis techniques (7), and the burgeoning field of synthetic biology (8–10), significant advances have been made in determining essential genes in a wide range of organisms. This paper represents an update, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes.

SUMMARY OF DATABASE UPDATES

In parallel to the rapid progress of the essential-gene field, DEG 5.0 has significant advances over DEG 1.0 by the following changes:

(1) The number of prokaryotic essential genes has increased about 10-fold, from 543 to 5260 (Table 1). (i) In DEG 1.0, some essential genes, e.g. those in *Escherichia coli*, were collected from literature searches, but in DEG 5.0 these records were replaced by those determined by genome-wide studies using the genetic footprinting technique (11) and systematic gene knockout experiments (12), (ii) In DEG 1.0, some essential genes, e.g. those in *Haemophilus influenzae* were determined by theoretical prediction from comparative genomics studies (13), but in DEG 5.0 these records were replaced by those determined by genome-wide studies using global transposon mutagenesis (14) and (iii) In 2004, only two genome-wide studies in identifying bacterial essential genes were done, but now 12 have been finished.
(2) The number of essential genes in eukaryotes has increased more than 5-fold, from 878 to 4808, because DEG 1.0 only had yeast essential genes,

*To whom correspondence should be addressed. Tel: +86 22 2740 2987; Fax: +86 22 2335 8329; Email: rzhang.cn@gmail.com

Table 1. Contents of DEG version 5.0[a]

No.	Kingdom	Organism	Essential gene no.	Method	Saturated or near saturated	References
1	Prokaryote	*Acinetobacter baylyi*	499	Single-gene deletions	Y	(23)
2	Prokaryote	*Bacillus subtilis*	271	Single-gene deletions	Y	(24)
3	Prokaryote	*Escherichia coli*	712	Genetic footprinting and single-gene deletions	Y	(11,12)
4	Prokaryote	*Francisella novicida*	392	Transposon mutagenesis	Y	(25)
5	Prokaryote	*Haemophilus influenzae*	642	Transposon mutagenesis	Y	(14)
6	Prokaryote	*Helicobacter pylori*	323	Transposon mutagenesis	Y	(26)
7	Prokaryote	*Mycobacterium tuberculosis*	614	Transposon mutagenesis	Y	(27)
8	Prokaryote	*Mycoplasma genitalium*	381	Transposon mutagenesis	Y	(18)
9	Prokaryote	*Mycoplasma pulmonis*	310	Transposon mutagenesis	Y	(28)
10	Prokaryote	*Pseudomonas aeruginosa*	335	Transposon mutagenesis	Y	(29)
11	Prokaryote	*Salmonella typhimurium*	230	Insertional-duplication mutagenesis	Y	(31)
12	Prokaryote	*Staphylococcus aureus*	302	Antisense RNA	N	(21,22)
13	Prokaryote	*Streptococcus pneumoniae*	244	Single-gene deletions and allelic replacement mutagenesis	N	(19,20)
14	Prokaryote	*Vibrio cholerae*	5	Transposon mutagenesis	N	(40)
15	Eukaryote	*Arabidopsis thaliana*	777	T-DNA insertion	N	(36)
16	Eukaryote	*Caenorhabditis elegans*	294	RNA interference	N	(34)
17	Eukaryote	*Danio rerio*	288	Insertion mutagenesis	N	(35)
18	Eukaryote	*Drosophila melanogaster*	339	P-element insertion mutagenesis	N	(33)
19	Eukaryote	*Homo sapiens*	118	Literature search	N	(38)
20	Eukaryote	*Mus musculus*	2114	Literature search	N	(37)
21	Eukaryote	*Saccharomyces cerevisiae*	878	Single-gene deletions	Y	(32)

[a]In some cases, there is a slight difference between the essential-gene number reported and that in DEG. This is mainly due to annotation changes in the latest versions of genomes such that some records could not be found, or because reported results contain identical records or are only partially published.

but DEG 5.0 also has those in humans, mice, worms, fruit flies, zebrafish and the plant *Arabidopsis thaliana*.

DATABASE DESCRIPTION

Essential genes in prokaryotes

Determination of a minimal gene set for cellular life was made possible by the availability of the first two completely sequenced genomes from the bacteria *Mycoplasma genitalium* (15) and *H. influenzae* (16). An attempt to determine a minimal gene set was pioneered by Koonin and coworkers by comparing these two sequenced genomes that belong to two ancient bacterial lineages, based on a notion that genes that are conserved between them are likely essential for cellular functions (13).

In 1999, Venter's group performed the first global transposon mutagenesis in *M. genitalium* to experimentally address the question of what is the minimal gene set for a living organism (17), and about 300 genes were estimated to be essential and were included in DEG 1.0. However, the concept of global transposon mutagenesis is in fact based on the identification of non-essential genes, i.e. those disrupted by transposons are identified, and those not disrupted are considered essential. Therefore, to gain the proof of gene dispensability in *M. genitalium*, Venter's group isolated and characterized every Tn4001 insertion mutants that were present in individual colonies picked from agar plates (18). Consequently, 382 genes were demonstrated to be essential, and these genes were included in DEG 5.0 by replacing those in version 1.0. A high-density transposon mutagenesis strategy was also applied to *H. influenzae* (14), and the essential genes so obtained replaced corresponding records in DEG 1.0, which were determined by comparative genomics (13).

In DEG 1.0, essential genes of *E. coli* were collected from http://magpie.genome.wisc.edu/~chris/essential.html, in which essential genes were obtained by searching related literatures. Using a genetic footprinting technique, Gerdes *et al.* (11) conducted a genome-wide, comprehensive experimental assessment of the *E. coli* genes necessary for robust aerobic growth, and consequently, 620 genes were identified to be essential. In addition, the Keio collection contains 303 essential genes that were determined by systematic single-gene knockout experiments (12). Therefore, in DEG 5.0, essential gene records obtained by literature search were replaced by those obtained through both genome-wide mutagenesis studies (11) and systematic single-gene knockout experiments (12), except that only one copy is retained for the 205 genes that overlap between the two studies.

About 100 *Streptococcus pneumoniae* essential genes were determined by a high-throughput gene disruption system (19). Later, 133 essential genes were determined by allelic replacement mutagenesis (20). In DEG 5.0, the two results were combined by removing redundant records, resulting in 244 essential genes in *S. pneumoniae*. DEG 1.0 contained 65 *Staphylococcus aureus* essential genes determined by using antisense RNA technique (21), and DEG 5.0 now contains 302 *S. aureus* essential genes by combining with results from the studies using the rapid shotgun antisense RNA method (22).

In the past several years, many genome-wide mutagenesis studies have been performed in a wide range of bacteria. In addition to those mentioned above, DEG 5.0 contains essential genes determined by large-scale

single-gene deletion studies in *Acinetobacter baylyi* (23) and *Bacillus subtilis* (24), those determined by global transposon mutagenesis in *Francisella novicida* (25), *Helicobacter pylori* (26), *Mycobacterium tuberculosis* (27), *Mycoplasma pulmonis* (28) and *Pseudomonas aeruginosa* (29,30), and those determined by trapping lethal insertions in *Salmonella typhimurium* (31).

Essential genes in eukaryotes

Another major improvement in DEG 5.0 is the inclusion of essential genes of many eukaryotes, including animals and the plant *A. thaliana*, whereas the only eukaryotic species in DEG 1.0 was *Saccharomyces cerevisiae* (32). The goal of determining bacterial minimal gene set also applies to eukaryotes, i.e. to define a minimal gene set needed to produce a living multicellular organism or a viable plant. Although this goal is obviously too ambitious at the current stage, much effort has already been devoted in the identification of essential genes in eukaryotes.

In the Drosophila genome, about 25% of genes were disrupted by P-element insertions by The Berkeley Drosophila Genome Project (33), and those genes whose disruption had lethal phenotypes were collected in DEG 5.0. In the *Caenorhabditis elegans* genome, using the RNA interference, Kamath *et al.* (34) inhibited the activity of about 86% of all genes, and characterized their phenotypes, and genes whose inhibition were lethal were included in DEG 5.0. Hopkins and coworkers conducted a large-scale insertional mutagenesis in zebrafish to identify genes essential for embryonic and early larval development (35), and the identified essential genes were collected in DEG 5.0. The first large-scale identification of essential genes in a flowering plant was performed by Meinke and coworker in *A. thaliana* by characterizing a large number of T-DNA insertion lines (36), and the identified essential genes were collected in DEG 5.0.

Large-scale gene inactivation studies have not been performed in mice, likely due to technical difficulties and labor intensiveness, however, because mice are probably the most important model organism, a large number of genes have already been inactivated by individual laboratories. In a study comparing essentiality between duplicate genes and singleton genes, Liao and Zhang (37) analyzed nearly 3900 individually inactivated mouse genes, and found that about 55% were essential in both singletons and duplicates. The essential genes analyzed in this study were collected in DEG 5.0. In another study comparing human and mouse essential genes, Liao and Zhang (38) extensively reviewed literatures to find genes whose null mutations in humans are lethal, and these human essential genes were also collected in DEG 5.0.

User interface and data access

The whole database is divided into two subdatabases, those of prokaryotic and eukaryotic essential genes. Each entry has a unique DEG identification number, gene name, gene reference number, gene function, and DNA and protein sequences. For prokaryotic essential genes, a link to the COG information (39) is also provided. All information is stored and operated by an open-source database management system, MySQL, which allows rapid data retrieval. There are several ways by which users can have access to the data. Users can browse the essential gene records, and can also search for essential genes by their names, functions, accession numbers and organisms. In addition, users can also perform BLAST searches against DEG for query DNA or protein sequences. Because the database is composed of two subdatabases, i.e. those for prokaryotes and eukaryotes, users need to perform the functions of Browse, Search and BLAST in individual databases. In addition, the whole database can also be downloaded upon request.

CONCLUSION AND FUTURE DEVELOPMENT

DEG 5.0 has significant advances over DEG 1.0 in both the number of essential genes and the number of organisms in which these genes are determined. These updates not only represent significant advances over the 2004 version of DEG, but also represent the rapid progress of the essential-gene field. In future, in prokaryotes, fueled by the availability of more and more complete genomes and the emerging field, synthetic biology, it is expected that the increase in the essential gene number will accelerate, whereas in eukaryotic model organisms, because most gene essentiality screens are far from saturated, the number of essential genes is also expected to grow. These advances will be reflected timely by DEG future updates. We welcome users' comments, corrections and new information, which will be used for updating. DEG is freely available at the website http://tubic.tju.edu.cn/deg or http://www.essentialgene.org.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referees for their constructive comments.

FUNDING

The present work was supported in part by the National Natural Science Foundation of China (NNSF 90408028). Funding for open access charge: NNSF 90408028.

Conflict of interest statement. None declared.

REFERENCES

1. Koonin,E.V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.*, **1**, 99–116.
2. Koonin,E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, **1**, 127–136.
3. Gerdes,S., Edwards,R., Kubal,M., Fonstein,M., Stevens,R. and Osterman,A. (2006) Essential genes on metabolic maps. *Curr. Opin. Biotechnol.*, **17**, 448–456.
4. Lartigue,C., Glass,J.I., Alperovich,N., Pieper,R., Parmar,P.P., Hutchison,C.A. 3rd, Smith,H.O. and Venter,J.C. (2007) Genome transplantation in bacteria: changing one species to another. *Science*, **317**, 632–638.
5. Galperin,M.Y. and Koonin,E.V. (1999) Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.*, **10**, 571–578.

6. Zhang,R., Ou,H.Y. and Zhang,C.T. (2004) DEG: a database of essential genes. *Nucleic Acids Res.*, **32**, D271–D272.
7. Judson,N. and Mekalanos,J.J. (2000) Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol.*, **8**, 521–526.
8. Benner,S.A. and Sismour,A.M. (2005) Synthetic biology. *Nat. Rev. Genet.*, **6**, 533–543.
9. Andrianantoandro,E., Basu,S., Karig,D.K. and Weiss,R. (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, **2**, 2006 0028.
10. Galperin,M.Y. (2008) The dawn of synthetic genomics. *Environ. Microbiol.*, **10**, 821–825.
11. Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S. *et al.* (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J. Bacteriol.*, **185**, 5673–5684.
12. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006 0008.
13. 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.
14. 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.
15. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of Mycoplasma genitalium. *Science*, **270**, 397–403.
16. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
17. 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.
18. Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M., Hutchison,C.A. 3rd, Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA*, **103**, 425–430.
19. Thanassi,J.A., Hartman-Neumann,S.L., Dougherty,T.J., Dougherty,B.A. and Pucci,M.J. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. *Nucleic Acids Res.*, **30**, 3152–3162.
20. Song,J.H., Ko,K.S., Lee,J.Y., Baek,J.Y., Oh,W.S., Yoon,H.S., Jeong,J.Y. and Chun,J. (2005) Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. *Mol. Cells*, **19**, 365–374.
21. 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.
22. Forsyth,R.A., Haselbeck,R.J., Ohlsen,K.L., Yamamoto,R.T., Xu,H., Trawick,J.D., Wall,D., Wang,L., Brown-Driver,V., Froelich,J.M. *et al.* (2002) A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Mol. Microbiol.*, **43**, 1387–1400.
23. de Berardinis,V., Vallenet,D., Castelli,V., Besnard,M., Pinet,A., Cruaud,C., Samair,S., Lechaplais,C., Gyapay,G., Richez,C. *et al.* (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Mol. Syst. Biol.*, **4**, 174.
24. 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.
25. Gallagher,L.A., Ramage,E., Jacobs,M.A., Kaul,R., Brittnacher,M. and Manoil,C. (2007) A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proc. Natl Acad. Sci. USA*, **104**, 1009–1014.
26. Salama,N.R., Shepherd,B. and Falkow,S. (2004) Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *J. Bacteriol.*, **186**, 7926–7935.
27. Sassetti,C.M., Boyd,D.H. and Rubin,E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.*, **48**, 77–84.
28. French,C.T., Lao,P., Loraine,A.E., Matthews,B.T., Yu,H. and Dybvig,K. (2008) Large-scale transposon mutagenesis of Mycoplasma pulmonis. *Mol. Microbiol.*, **69**, 67–76.
29. Liberati,N.T., Urbach,J.M., Miyata,S., Lee,D.G., Drenkard,E., Wu,G., Villanueva,J., Wei,T. and Ausubel,F.M. (2006) An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proc. Natl Acad. Sci. USA*, **103**, 2833–2838.
30. Jacobs,M.A., Alwood,A., Thaipisuttikul,I., Spencer,D., Haugen,E., Ernst,S., Will,O., Kaul,R., Raymond,C., Levy,R. *et al.* (2003) Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proc. Natl Acad. Sci. USA*, **100**, 14339–14344.
31. Knuth,K., Niesalla,H., Hueck,C.J. and Fuchs,T.M. (2004) Large-scale identification of essential Salmonella genes by trapping lethal insertions. *Mol. Microbiol.*, **51**, 1729–1744.
32. 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.
33. Spradling,A.C., Stern,D., Beaton,A., Rhem,E.J., Laverty,T., Mozden,N., Misra,S. and Rubin,G.M. (1999) The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics*, **153**, 135–177.
34. Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al.* (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231–237.
35. Amsterdam,A., Nissen,R.M., Sun,Z., Swindell,E.C., Farrington,S. and Hopkins,N. (2004) Identification of 315 genes essential for early zebrafish development. *Proc. Natl Acad. Sci. USA*, **101**, 12792–12797.
36. Tzafrir,I., Pena-Muralla,R., Dickerman,A., Berg,M., Rogers,R., Hutchens,S., Sweeney,T.C., McElver,J., Aux,G., Patton,D. *et al.* (2004) Identification of genes required for embryo development in Arabidopsis. *Plant Physiol.*, **135**, 1206–1220.
37. Liao,B.Y. and Zhang,J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.*, **23**, 378–381.
38. Liao,B.Y. and Zhang,J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl Acad. Sci. USA*, **105**, 6987–6992.
39. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
40. Judson,N. and Mekalanos,J.J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.*, **18**, 740–745.