*Published online 10 November 2011* *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D901–D906***
*doi:10.1093/nar/gkr986*

# OGEE: an online gene essentiality database

**Wei-Hua Chen[1], Pablo Minguez[1], Martin J. Lercher[2] and Peer Bork[1,3,*]**

[1]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, [2]Institute for Computer Science, Heinrich-Heine-University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf and [3]Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch Robert-Rössle-Str. 10, 13092 Berlin, Germany

Received August 14, 2011; Revised September 28, 2011; Accepted October 17, 2011

## ABSTRACT

***OGEE* is an Online GEne Essentiality database. Its main purpose is to enhance our understanding of the essentiality of genes. This is achieved by collecting not only experimentally tested essential and non-essential genes, but also associated gene features such as expression profiles, duplication status, conservation across species, evolutionary origins and involvement in embryonic development. We focus on large-scale experiments and complement our data with text-mining results. Genes are organized into data sets according to their sources. Genes with variable essentiality status across data sets are tagged as conditionally essential, highlighting the complex interplay between gene functions and environments. Linked tools allow the user to compare gene essentiality among different gene groups, or compare features of essential genes to non-essential genes, and visualize the results. *OGEE* is freely available at http://ogeedb.embl.de.**

## INTRODUCTION

Large-scale efforts to link genotypes to phenotypes belong to the most important and challenging tasks in the post-omics era. Essential genes, whose removal results in inviability or infertility, are of particular interests because of their theoretical and practical applications, for example, in studying the robustness of a biological system (1), defining a minimal set of genes for a free living organism (2) and identifying effective drug targets (3).

Essentiality often depends on the environment (4), especially for bacterial genes, or for eukaryotic genes that were tested in cell lines. For example, genes coding for proteins involved in the biosynthesis of amino acids, nucleic acids and vitamins are essential for cell survival in minimal media, but not in rich media where the corresponding metabolites are supplied (4). However, so far the concept of 'conditional essentiality' has not been widely adopted by existing essential gene databases.

Gene essentiality does not only depend on individual gene functions, but can also be affected by global factors. Duplicated genes are typically less essential than the genomic average because they often overlap in gene function and expression profile; genes forming hubs in PPI networks (those connected to many direct neighbors) are more often essential (5); and genes involved in development and tissue differentiation in higher eukaryotes are also more likely to be essential (6). However, given the complex nature of biological systems, gene essentiality is often affected by multiple factors simultaneously; studying one factor at a time may generate conflicting results among species. For example, in a biased data set, mouse duplicates and singletons were reported to be equally essential (7), which disagreed with theoretical expectations and experimental findings in yeast (8). Experimental biases could only partially explain the contradiction (6). In a previous study, we showed that considering both the duplication status of genes and their evolutionary origins could solve the discrepancies (Chen, W.-H., Trachana, K., Lercher, M.J., and Bork, P., unpublished data).

Our understanding of gene essentiality is still limited. Progress can be enhanced by collecting the following information into a central database: (i) tested essential and non-essential genes, allowing comparisons between the two groups; (ii) essentiality information obtained from large-scale studies, facilitating genome-wide analyses, as well as more precise information from small-scale studies, more suited for gene-centered biological research; (iii) additional gene features that are either known or hypothesized to influence gene essentiality. Ideally, such a database should come with a set of tools that allow the user to systematically explore and analyze the raw data.

Existing essential gene databases either only include data for a specific species (9) or contain only essential genes (10). This provided the motivation to develop *OGEE*, an online gene essentiality database that combines points a–d with a set of tools for large-scale data analysis. This should make *OGEE* useful to both biologists and bioinformaticians.

*To whom correspondence should be addressed. Tel: +49 6221 387 852; Fax: +49 6221 387 517; Email: bork@embl.de

## DATA GENERATION

### Collection and organization of genes tested for essentiality

We collected 91 436 protein-coding genes from 8 eukaryotic and 16 prokaryotic organisms tested for essentiality in genome-wide studies (2,3,9,11–37). For data sets that both essential and non-essential genes are publicly available, the genomic proportion of essential genes ($P_E$) ranges from ~2% [data set 347 (11) of *Drosophila melanogaster*] to 66.04% [*Aspergillus fumigatus Af293*, data set 361 (3)] in eukaryotes and from 5.46% [*Bacilus subtilis 168*, data set 352 (17)] to 80% [*Mycoplasma genitalium G37*, data set 357 (2)] in prokaryotes. It seems that overall $P_E$ in eukaryotes is most strongly influenced by organism complexity and by the methods employed for testing, in particular by the experimental conditions surveyed. Gene knockout techniques [data sets 349 (14) and 350 (37) of *Mus musculus* and *Saccharomyces cerevisiae*, respectively] generate higher $P_E$ than siRNA-based methods [data sets 348 (12) and 347 (11) of *Homo sapiens* and *D. melanogaster*, respectively]. Multi-cellular organisms have higher $P_E$ than single-celled eukaryotes (*M. musculus* versus *S. cerevisiae*) if similar techniques were used. Cell lines generate lower $P_E$ than *in vivo* if the same multi-cellular organism is used [data sets 347 (11) and 363 (25) of *D. melanogaster*]. In prokaryotes, overall $P_E$ is affected by details of the survey technology as well as by genome size and life style (free living versus parasitic).

In addition to the collection of large-scale data, we also employed text-mining to obtain 3543 genes from 38 species that were tested in small-scale studies. We applied a customized text-mining pipeline based on the one used for data collection by the STRING database (38). We searched for a set of terms related to essentiality (Supplementary Table S1) in PubMed abstracts (as published February 2011) and manually checked the results and removed some false positives. We divided identified genes into essential and non-essential genes according to their associated terms. Due to a strong reporting bias, most genes identified in this way were essential. Among those, 3168 (89.4% of 3543) genes overlapped with those tested in genome-wide studies. Please note that although substantial efforts have been made to improve the quality

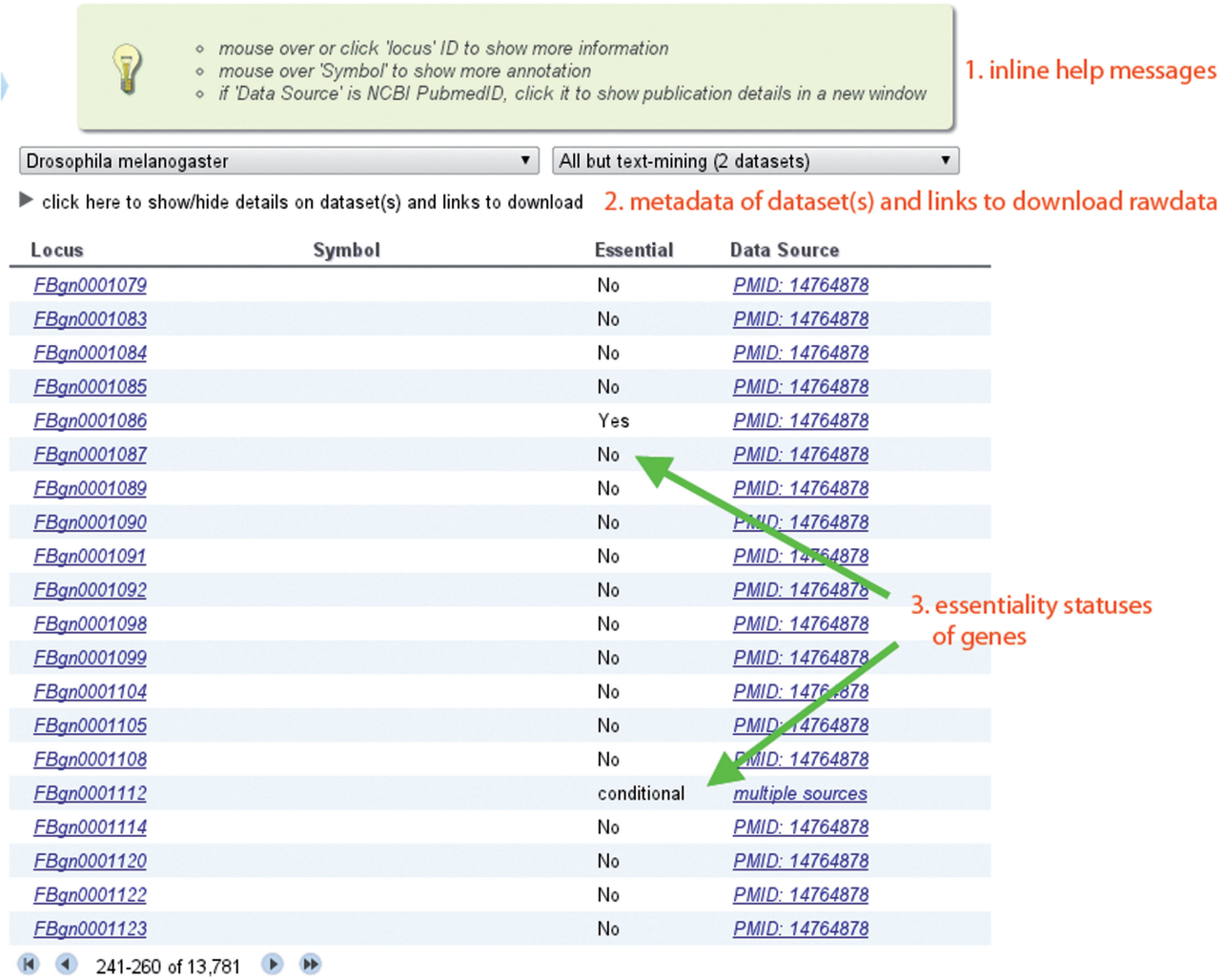


**Figure 1.** Interface of the 'Browse' module.

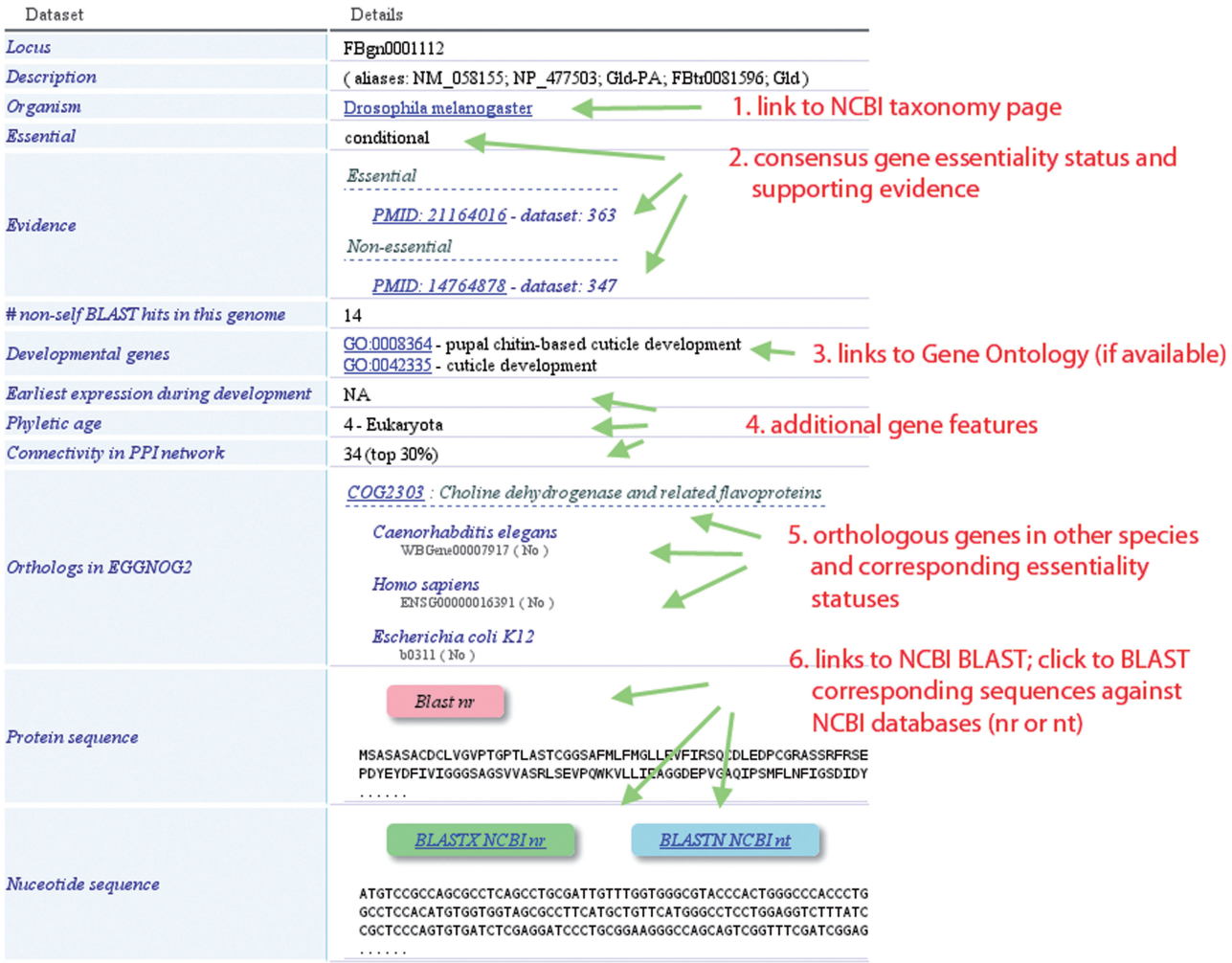


**Figure 2.** Extra gene features shown in a popup window. This window will show up when clicking locus IDs in the 'Browse' or 'Search' modules.

of the text-mining data, there might still be significant fraction of false-positive results; please use with caution.

We organized genes in each organism into distinct data sets according to the data source; a gene can have multiple entries within a data set or in different data sets. Two entries of a gene would be included in two distinct data sets if the gene was tested in a large-scale study as well as in a small-scale study; if a gene was tested by several small-scale studies, multiple entries of this gene would be included in the text-mining data set, with each entry corresponding to a distinct PubMed record. A gene was marked as 'conditionally essential' if multiple entries for this gene exist in *OGEE* but essentiality status varies among entries (see, e.g. the essentiality status of gene 'FBgn0001112' in Figure 1 and the supporting evidence in Figure 2).

### Collection of gene features influencing gene essentiality

We collected several gene features that are known to influence gene essentiality, encompassing duplication status, connectivity in protein–protein interaction (PPI) networks (defined as the number of direct neighbors) (5) and evolutionary origins of genes (defined as the age of the evolutionarily most distant species group where homologs can be found (39); see the web Q&As for more details).

We also collected several extra features that might influence gene essentiality, including the number of homologous genes (family size) in the same genome, and the earliest expression stage during embryonic development [for multi-cellular organisms only; data was obtained from the NCBI UniGene database (40)]. It is known that duplicates are often less essential than singletons. This may be due to a range of factors, including the ability of duplicates to provide a functional backup for each other, lower expression abundances of duplicates (41,42), or a lower duplicability of the genes in certain important functional classes (43). It is thus conceivable that duplicates in large gene families are even less likely to be essential than duplicates in smaller families. In multicellular organisms, embryonic development is a tightly regulated chain of events. Disruption of genes expressed earlier may affect all subsequent events, thereby causing more severe phenotypes in the host. Both gene family size

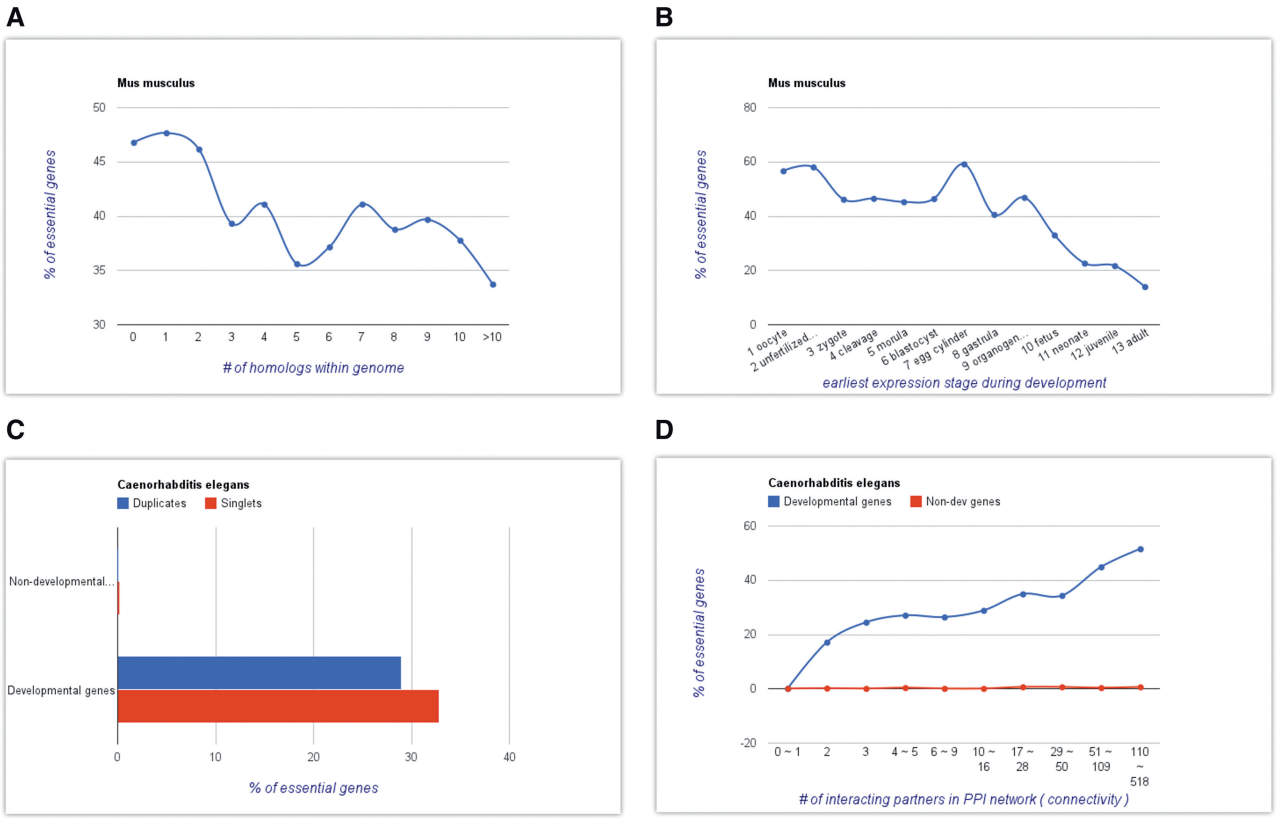


**Figure 3.** Screen shots taken from the 'Analyze' module. With integrated tools, the user can easily explore and analyze the collected data, including the visualization of results. Shown here are the results of the following analyses: (**A**) the proportion of essential genes ($P_E$) as a function of family size (number of homologous genes within the genome) in mouse, (**B**) $P_E$ as a function of the earliest expression stage during mouse development, (**C**) the effects of gene duplication status and involvement in development on gene essentiality in *Caenorhabditis elegans* and (**D**) the effects of gene connectivity and involvement in development on gene essentiality in *C. elegans*.

and earliest expression in development are indeed correlated with $P_E$ in mouse (Figure 3A and B).

## USAGE OF OGEE

The functionalities of *OGEE* have been divided into six different modules (tabs): 'Summary', 'Browse', 'Search', 'Analyze', 'Download' and 'Q&As'. We provide inline help messages displayed as 'tooltips' within each module; we also provide detailed help contents and answers to frequently asked questions in 'Q&As'. Below, we introduce several of the most interesting features of *OGEE*.

### Viewing details of individual genes

In the 'Browse' and 'Search' modules, by default only some gene features such as essentiality, duplication status and data sources will be displayed (Figure 1). To view more details of individual genes, the user can simply mouse over or click the locus names; a popup window containing all available information for the corresponding gene will appear. As shown in Figure 2, extra information including gene description, type of evidence for gene essentiality and corresponding links to original data sources, involvement in development, evolutionary origin (phyletic age), connectivity in the PPI network, as well as nucleotide and protein sequences are available. Links to other databases, including Gene Ontology (44), EGGNOG2 (45), NCBI taxonomy, as well as NCBI BLAST (40) are also integrated (Figure 2). For example, if the gene of interests is involved in development, several corresponding GO IDs and terms will be shown; clicking each GO ID, the user will be redirected to the corresponding page at the Gene Ontology website. Similarly, the user will be redirected to the corresponding NCBI taxonomy page if clicking on the organism name. The NCBI BLAST website will be opened in a new window if clicking on the BLAST NCBI links.

The popup window also features in-site data integration. For example, if a query gene has orthologs in other species collected by *OGEE*, not only the corresponding orthologs [based on EGGNOG2 (45)], but also their essentiality status will be shown (Figure 2). This way, the conservation of a gene as well as the conservation of its essentiality across species can be checked easily.

### Analyzing collected gene features using linked tools

One of the most interesting features of *OGEE* is that users can analyze the data systematically and visualize the results with integrated tools from the 'Analyze' module. With 'Analyze', the user can divide genes into distinct groups according to one of the available features, calculate the proportion of essential genes ($P_E$) in each group and

then plot the results as either a bar-chart or line plot. To illustrate this feature, Figures 3A and B show average mouse $P_E$ values as functions of the earliest expression stage during development and gene family size, respectively; both factors affect $P_E$ values globally.

Users can also investigate two gene features simultaneously to study their effects individually or in combination. For example, the user can divide genes first into developmental and non-developmental genes, and then further divide each group into duplicates and singletons (Figure 3C). Similarly, on could first divide genes according to the connectivity in PPI network and then according to their involvement in development (Figure 3D).

By default, predefined breaks by which genes can be divided into distinct groups and matching labels are used. However, if desired, the user can change the default settings by providing customized breaks and labels.

### Open access to all data contained in OGEE

Our data are freely accessible to all academic users. We provide an SQL-dump file of the whole database as well as several selected data sections as tab-delimited flat files in the 'Download' module. Users can also download individual gene essentiality data sets for a selected species in 'Browse' and raw data used in data analysis in 'Analyze'.

## CONCLUSIONS

*OGEE* introduces several unique and novel features compared with existing gene essentiality databases. For example (i) *OGEE* provides both essential and non-essential genes from large-scale as well as small-scale studies; (ii) *OGEE* introduces 'conditional essentiality' to reflect the complexity of biological systems and the interplay between gene functions and environments; (iii) *OGEE* lists a variety of gene features known or suspected to influence gene essentiality; and (iv) *OGEE* provides a set of online tools to explore and analyze the data and to visualize the results. We thus believe that *OGEE* should be highly useful to biologists and bioinformaticians studying gene essentiality, whether focusing on individual genes or on genome-wide analyses.

## FUTURE DIRECTIONS

Future development of *OGEE* will include the incorporation of essential non-coding genes, and the possibility for users to submit additional essentiality data.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table S1. Key words used to search for essential and non-essential genes in PubMed abstracts.

## FUNDING

Funding for open access charge: BMBF (Bundesministerium für Bildung und Forschung) MedSys grant #0315450C to Peer Bork.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Keller,P.J. and Knop,M. (2009) Evolution of mutational robustness in the yeast genome: a link to essential genes and meiotic recombination hotspots. *PLoS Genet*., **5**, e1000533.
2. Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M., Hutchison,C.A. III, Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA*, **103**, 425–430.
3. Hu,W., Sillaots,S., Lemieux,S., Davison,J., Kauffman,S., Breton,A., Linteau,A., Xin,C., Bowman,J., Becker,J. *et al.* (2007) Essential gene identification and drug target prioritization in Aspergillus fumigatus. *PLoS Pathog*., **3**, e24.
4. D'Elia,M.A., Pereira,M.P. and Brown,E.D. (2009) Are essential genes really essential? *Trends Microbiol*., **17**, 433–438.
5. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
6. Makino,T., Hokamp,K. and McLysaght,A. (2009) The complex relationship of gene duplication and essentiality. *Trends Genet*., **25**, 152–155.
7. Liao,B.-Y. and Zhang,J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet*., **23**, 378–381.
8. Gu,Z., Steinmetz,L.M., Gu,X., Scharfe,C., Davis,R.W. and Li,W.H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.
9. Hashimoto,M., Ichimura,T., Mizoguchi,H., Tanaka,K., Fujimitsu,K., Keyamura,K., Ote,T., Yamakawa,T., Yamazaki,Y., Mori,H. *et al.* (2005) Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Mol. Microbiol*., **55**, 137–149.
10. Zhang,R. and Lin,Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res*., **37**, D455–D458.
11. Boutros,M., Kiger,A.A., Armknecht,S., Kerr,K., Hild,M., Koch,B., Haas,S.A., Paro,R. and Perrimon,N. (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science*, **303**, 832–835.
12. Silva,J.M., Marran,K., Parker,J.S., Silva,J., Golding,M., Schlabach,M.R., Elledge,S.J., Hannon,G.J. and Chang,K. (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, **319**, 617–620.
13. 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.
14. Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E., Eppig,J.T. and Mouse Genome Database,G. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*., **39**, D842–D848.
15. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*., **30**, 69–72.
16. 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.
17. 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.
18. Langridge,G.C., Phan,M.D., Turner,D.J., Perkins,T.T., Parts,L., Haase,J., Charles,I., Maskell,D.J., Peters,S.E., Dougan,G. *et al.* (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res*., **19**, 2308–2316.
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. Chaudhuri,R.R., Allen,A.G., Owen,P.J., Shalom,G., Stone,K., Harrison,M., Burgis,T.A., Lockyer,M., Garcia-Lara,J., Foster,S.J. *et al.* (2009) Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics*, **10**, 291.
21. 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.
22. Salama,N.R., Shepherd,B. and Falkow,S. (2004) Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *J. Bacteriol.*, **186**, 7926–7935.
23. Trepod,C.M. and Mott,J.E. (2005) Elucidation of essential and nonessential genes in the Haemophilus influenzae Rd cell wall biosynthetic pathway by targeted gene disruption. *Antimicrob. Agents Chemother.*, **49**, 824–826.
24. 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.
25. Chen,S., Zhang,Y.E. and Long,M. (2010) New genes in Drosophila quickly become essential. *Science*, **330**, 1682–1685.
26. Meinke,D., Muralla,R., Sweeney,C. and Dickerman,A. (2008) Identifying essential genes in Arabidopsis thaliana. *Trends Plant Sci.*, **13**, 483–491.
27. 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.
28. 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.
29. 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.
30. Kraemer,P.S., Mitchell,A., Pelletier,M.R., Gallagher,L.A., Wasnick,M., Rohmer,L., Brittnacher,M.J., Manoil,C., Skerett,S.J. and Salama,N.R. (2009) Genome-wide screen in Francisella novicida for genes required for pulmonary and systemic infection in mice. *Infect. Immun.*, **77**, 232–244.
31. 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.
32. Chalker,A.F., Minehart,H.W., Hughes,N.J., Koretke,K.K., Lonetto,M.A., Brinkman,K.K., Warren,P.V., Lupas,A., Stanhope,M.J., Brown,J.R. *et al.* (2001) Systematic identification of selective essential genes in Helicobacter pylori by genome prioritization and allelic replacement mutagenesis. *J. Bacteriol.*, **183**, 1259–1268.
33. Sassetti,C.M. and Rubin,E.J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl Acad. Sci. USA*, **100**, 12989–12994.
34. 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.
35. 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.
36. Lamichhane,G., Zignol,M., Blades,N.J., Geiman,D.E., Dougherty,A., Grosset,J., Broman,K.W. and Bishai,W.R. (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. *Proc. Natl Acad. Sci. USA*, **100**, 7213–7218.
37. Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. *et al.* (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67–73.
38. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
39. Wolf,Y.I., Novichkov,P.S., Karev,G.P., Koonin,E.V. and Lipman,D.J. (2009) Inaugural article: the universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl Acad. Sci. USA*, **106**, 7273–7280.
40. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
41. Qian,W., Liao,B.-Y., Chang,A.Y.-F. and Zhang,J. (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.*, **26**, 425–430.
42. Schrimpf,S.P., Weiss,M., Reiter,L., Ahrens,C.H., Jovanovic,M., Malmstrom,J., Brunner,E., Mohanty,S., Lercher,M.J., Hunziker,P.E. *et al.* (2009) Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol.*, **7**, e48.
43. He,X. and Zhang,J. (2006) Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.*, **23**, 144–151.
44. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
45. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.