D940–D944 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 30 October 2016*
doi: 10.1093/nar/gkw1013

OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines

Wei-Hua Chen[1,*,†], Guanting Lu[2,†], Xiao Chen[3], Xing-Ming Zhao[3] and Peer Bork[4,5,6,7]

[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology (HUST), 430074 Wuhan, Hubei, China, [2]Department of Blood Transfusion, Tangdu Hospital, the Fourth Military Medical University, No 1, Xinsi Road, Chanba District, 710000 Xi'an, China, [3]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China, [4]European molecular biology laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany, [5]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany, [6]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Straße 10, 13125 Berlin, Germany and [7]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Received August 23, 2016; Revised October 14, 2016; Editorial Decision October 15, 2016; Accepted October 18, 2016

ABSTRACT

***OGEE* is an Online GEne Essentiality database. To enhance our understanding of the essentiality of genes, in *OGEE* we collected experimentally tested essential and non-essential genes, as well as associated gene properties known to contribute to gene essentiality. We focus on large-scale experiments, and complement our data with text-mining results. We organized tested genes into data sets according to their sources, and tagged those with variable essentiality statuses across data sets as conditionally essential genes, intending to highlight the complex interplay between gene functions and environments/experimental perturbations. Developments since the last public release include increased numbers of species and gene essentiality data sets, inclusion of non-coding essential sequences and genes with intermediate essentiality statuses. In addition, we included 16 essentiality data sets from cancer cell lines, corresponding to 9 human cancers; with *OGEE*, users can easily explore the shared and differentially essential genes within and between cancer types. These genes, especially those derived from cell lines that are similar to tumor samples, could reveal the oncogenic drivers, paralogous gene expression pattern and chromosomal structure of the corresponding cancer types, and can be further screened to identify targets for cancer therapy and/or new drug development. *OGEE* is freely available at *http://ogee.medgenius.info*.**

INTRODUCTION

Essential genes are those genes of an organism that are critical for its survival; essential genes are of particular importance because of their theoretical and practical applications such as studying the robustness of a biological system (1), defining a minimal genome/organism (2,3) and identifying effective therapeutic targets in pathogens (4–6) and human cancers (7–11). In recent years, the technologies used for gene essentiality studies have been evolving rapidly, ranging from low-throughput single gene knockout experiment (12,13) to high-throughput mutagenesis (3), RNAi (7,8) and more recently CRISPR-based genome editing methods (14–18); recent studies showed that CRISPR technology outperformed other methods (14,19), featuring low noise and minimal off-target effects (19).

Being essential is not an intrinsic property of a gene; rather, it is highly dependent on a variety of factors including the function and expression pattern of the gene, the genetic background of the host, the environment and other settings. 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 (20). In addition, different experimental methods may generate different results. For example, CRISPR-based methods could identify more essential genes than siRNA-based methods (21), while cell lines generate lower propor-

*To whom correspondence should be addressed. Tel: +86 2787542127; Fax: +86 2787542527; Email: weihuachen@hust.edu.cn
†These authors contributed equally to the paper as first authors.

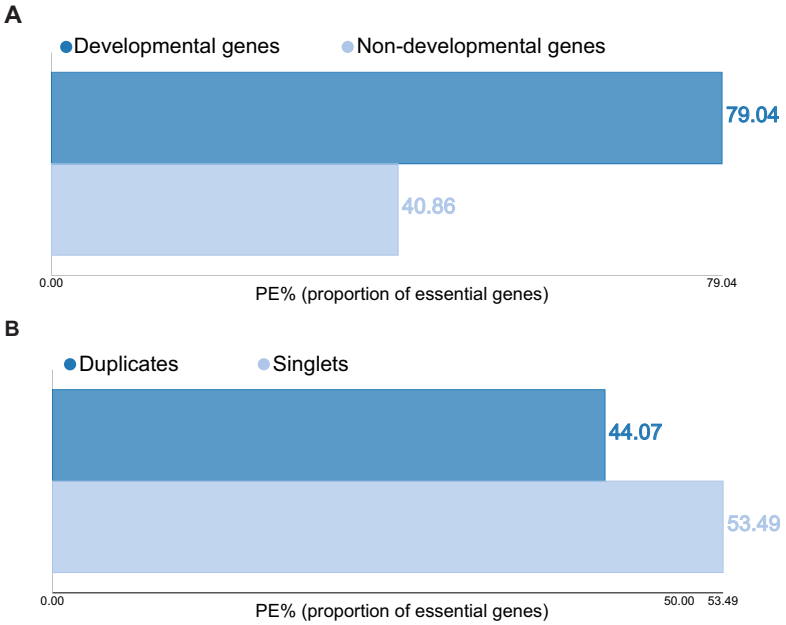

Figure 1. Screenshots taken from the 'Analyze' page. With integrated tools, users can easily analyze the collected data and visualize the results. Shown here are the proportion of essential genes (P_E) as a function of involvement in development (developmental versus non-developmental genes, panel (**A**) and duplication statuses (duplicates versus singlets, panel (**B**)) in mouse.

tion of essential genes than *in vivo* if the same multi-cellular organism is used (22).

Genes with variable essentiality statuses under different circumstances are referred to 'conditionally essential genes (CEGs)' or 'differentially essential genes (DEGs)' (14,22). CEG is a biologically meaningful and very important concept; e.g. genes that are essential in a cancer cell line but are non-essential in human tissues can reveal the oncogenic drivers, paralogous gene expression pattern and chromosomal structure of the corresponding cancer type (14).

In 2012, we introduced *OGEE* v1 (22) to promote the concept of 'conditional essentiality', which had not been widely adopted by existing essential gene databases at the time, and to advance our understanding on gene essentiality. We did so by including not only essential and non-essential genes, but also associated gene properties that are known to affect gene essentiality; we provided tools that allow users to compare gene essentiality among different gene groups, or compare properties of essential genes to non-essential genes. In addition, we organized experimentally tested genes into data sets according to their sources and tagged those with variable essentiality statuses across data sets as CEGs.

In this study we introduce an updated version of *OGEE*. In this new version we added new species and new data sets; we added genes with intermediate essentiality statuses (fitness genes) and non-coding essential genes. In addition, we re-organized the 16 gene essentiality data sets from human cancer cell lines corresponding to nine cancer types in order to help users to explore the shared and differentially essential genes within and between cancer types, because these genes, especially those derived from cell lines that are similar to tumor samples, could be further screened to identify targets for cancer therapy and/or new drug development.

DATA GENERATION

Collection and organization of genes tested for essentiality

We collected 99 large-scale gene essentiality experiments (data sets) for 48 species, including 34 data sets for 9 eukaryotes and 65 data sets for 39 prokaryotes. We added 1609 noncoding genes and 122 non-transcribed genomic regions from 10 species. In addition to essential and nonessential genes, we also included 1911 fitness genes from 10 species, and 37 growth-advanced genes from two species. Fitness genes are defined as those whose removal is not lethal but could result in significantly decreased fitness, while growth-advanced genes are defined as genes whose removal lead to significantly increased fitness. In the statistics below, fitness genes are counted as non-essential genes.

In sum, our database contains 167 799 genes tested for essentiality from 48 species, increased significantly from the 91 436 genes and 24 species respectively from the last version (22). In total 43 961 genes are covered by multiple data sets (in each species the text-mining results are considered as a data set), representing ∼26.2% of all collected genes; among which 13 397 genes are CEGs, accounting for ∼30.5% of those covered by multiple data sets. The proportion of conditionally essential genes (P_{CEG}) in species having more than 200 genes covered by two or more data sets ranges from 9.2% in *Staphylococcus aureus subsp. aureus NCTC 8325* to 41.6% in *Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344*, as shown in Table 1. The number of data sets does not seem to contribute significantly to P_{CEG} (Pearson's correlation coefficient = −0.24, P-value = 0.36).

Table 1. Statistics on conditionally essential genes in selected species with least 200 genes covered by multiple data sets in *OGEE*

Species	data sets	tested genes	essential genes	genes covered by multiple data sets	conditionally essential genes
Homo sapiens	18	21 556	7168	18 855	6985 (37.0%)
Schizosaccharomyces pombe	7	5509	1571	2522	279 (11.1%)
Drosophila melanogaster	2	13 781	408	437	141 (32.3%)
Pseudomonas aeruginosa UCBPP-PA14	4	5966	1842	5300	1455 (27.5%)
Escherichia coli K12	4	4322	740	4066	509 (12.5%)
Mycobacterium tuberculosis H37Rv	5	4008	1028	4002	1388 (34.7%)
Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344	4	3774	1514	2715	1130 (41.6%)
Staphylococcus aureus subsp. aureus NCTC 8325	2	2899	557	2713	250 (9.2%)
Haemophilus influenzae Rd KW20	4	1750	847	1634	617 (37.8%)
Mycoplasma pneumoniae M129	2	1203	508	1203	196 (16.3%)

Species having at least 200 genes covered by multiple data sets are listed here; the species are ordered first by the kingdom they are in (the 1st column) and then by the number of genes covered by multiple data sets (the 5th column). Essential genes are those that are essential in any collected data sets, i.e. genes that are essential in one data set but non-essential in others are also counted. The proportion of conditionally essential genes (P_{CEG}, percentage in parentheses of the last column) is calculated as the ratio between the 'conditionally essential genes' (the last column) and the 'genes covered by multiple data sets' (the 5th column). Please note that text-mining results, if available, will be counted as one data set in a species; please consult the 'Browse' page of the database for a complete and interactive version of the table.

Collection of gene properties influencing gene essentiality

We also collected several gene properties that are known to influence gene essentiality, including duplication status (23), the number of homologous genes (family size) in the same genome, connectivity in protein–protein interaction (PPI) networks (defined as the number of direct neighbors) (24), functional category of a gene (25) and the earliest expression stage during embryonic development. We used the BLAST tool (26) to search for duplicated genes within each genome using parameters and cutoffs described previously (27), and calculated the family size for each duplicated gens accordingly. We also calculated evolutionary measurements for each duplicated gene and its best BLAST hit, including Ka, Ks and Ka/Ks ratio using KaKs_Calculator (v2.0) (28). We obtained the PPI data from STRING v10.0 (29), the functional category data from Gene Ontology (30) and the expression data (for multi-cellular organisms only) from NCBI UniGene database (31). For more information, please consult the 'Help' page of the database.

BUILT-IN TOOLS FOR ANALYZING COLLECTED GENE PROPERTIES

We also provided integrated tools in the 'Analyze' page for users to analyze the impact of the collected gene properties on gene essentiality: users can divide genes into distinct groups according to one of the available properties, calculate the proportion of essential genes (P_E) in each group, and then plot the results as bar-chart. To illustrate this feature, we plotted in Figure 1 the P_E values of different groups of mouse genes as functions of their involvement in development (Figure 1A) and duplication status (Figure 1B). These results showed that that developmental genes are more essential than non-developmental genes, while singletons are more essential than duplicated genes, consisting to previous results (23,25); these trends are generally true in other species.

RE-ORGANIZATION OF ESSENTIAL GENES FROM HUMAN CANCER CELL LINES

In recent years, 'conditional essentiality' or 'differential essentiality' has been increasingly used as a tool for researchers to interrogate genes that are essential under specific conditions and search for genes required by the survival of human cancer cell lines (7,9–11,14,16). In *OGEE* we collected in total 16 such data sets and re-organized them into 9 groups according to their cancers of origin, including breast cancer, Burkitt's lymphoma, chronic myelogenous leukemia (CML), colon cancer, esophageal squamous carcinoma, glioblastoma (GBM), non-small cell lung cancer (NSCLC), ovarian cancer and pancreatic cancer. Shared and differentially essential genes within and between cancer types were pre-calculated. Shown in Table 2 are the brief summary on the nine cancers, including the number of data sets for each cancer, the number of total essential genes and the number of uniquely essential genes. Here, the 'uniquely essential genes' are defined as those that are non-essential in any other human data sets available in *OGEE*. An up-to-date version of this table and additional results can be found in the 'Cancer' page of our website.

Lineage-specific essential genes, i.e. those that are essential only in a particular cancer type, are important targets for cancer therapies; because they are likely the results of the unique mutational profile and subsequent functional consequences of the cell line, targeting these genes in cancer therapies will achieve high efficiency and specificity. Cancer cell lines are often used as models for cancer research. However, recent studies suggest that although some cell lines are indeed good models for cancers, some other cancer cell lines could have pronounced differences as compared to tumor samples of the same origin in terms of copy-number changes, key mutations and mRNA expression profiles, due in part to ambiguity in classification and annotation (32,33). Thus, in the future, we will exclude cell lines that are remarkably different from cancers of the same origin, and keep only the good ones, should such information are reliable and easily accessible.

DATA ACCESS

All data are freely accessible to all academic users. Users can download combined data from the 'Downloads' page. Users can also download individual data sets or combined data sets for individual species in the 'Browse' page.

Table 2. Summary of the 16 gene essentiality data sets from 9 human cancers collected in *OGEE*

Cancer	Data sets	Essential genes	Uniquely essential genes
breast	1	146	67
Burkitt's lymphoma	2	1897	198
CML	4	3210	1324
colon	2	1394	899
esophageal squamous	1	41	34
GBM	1	21	14
NSCLC	1	28	20
ovarian	2	130	87
pancreatic	2	199	126

'Essential genes' (the 3rd column) are genes that are essential in any of the data set(s) of a particular cancer type; 'Uniquely essential genes' (the last column) are genes that are subset of 'Essential genes' but non-essential in any other collected human data sets. An up-to-date version of this table can be found at http://ogee.medgenius.info/cancer/.

CONCLUSIONS

In this article, we introduced *OGEE* v2, an online gene essentiality database. Updates since the last updated version include increased numbers of species and gene essentiality data sets, inclusion of non-coding essential sequences and fitness genes. We also re-organize the essentiality data sets from nine human cancers so that our users can easily explore the shared and differentially essential genes within and between cancer types. As compared with existing gene essentiality databases such as DEG (34), *OGEE* provides several unique features. 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, genetic backgrounds and environments; (iii) *OGEE* lists a variety of gene properties known to influence gene essentiality; (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. In the future, we aim to update *OGEE* regularly in order to provide up-to-date contents to our users.

FUNDING

Funding for open access charge: startup funding for Chen's lab at the Huazhong University of Science and Technology, China

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. 3rd, Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 425–430.
3. Lluch-Senar,M., Delgado,J., Chen,W.H., Llorens-Rico,V., O'Reilly,F.J., Wodke,J.A., Unal,E.B., Yus,E., Martinez,S., Nichols,R.J. *et al.* (2015) Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.*, **11**, 780.
4. 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.
5. Lu,Y., Deng,J., Rhodes,J.C., Lu,H. and Lu,L.J. (2014) Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus. *Comput. Biol. Chem.*, **50**, 29–40.
6. Paul,M.L., Kaur,A., Geete,A. and Sobhia,M.E. (2014) Essential gene identification and drug target prioritization in Leishmania species. *Mol. Biosyst.*, **10**, 1184–1195.
7. Cowley,G.S., Weir,B.A., Vazquez,F., Tamayo,P., Scott,J.A., Rusin,S., East-Seletsky,A., Ali,L.D., Gerath,W.F., Pantel,S.E. *et al.* (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data*, **1**, 140035.
8. Luo,J., Emanuele,M.J., Li,D., Creighton,C.J., Schlabach,M.R., Westbrook,T.F., Wong,K.K. and Elledge,S.J. (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell*, **137**, 835–848.
9. Marcotte,R., Brown,K.R., Suarez,F., Sayad,A., Karamboulas,K., Krzyzanowski,P.M., Sircoulomb,F., Medrano,M., Fedyshyn,Y., Koh,J.L. *et al.* (2012) Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.*, **2**, 172–189.
10. Cheung,H.W., Cowley,G.S., Weir,B.A., Boehm,J.S., Rusin,S., Scott,J.A., East,A., Ali,L.D., Lizotte,P.H., Wong,T.C. *et al.* (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 12372–12377.
11. Luo,B., Cheung,H.W., Subramanian,A., Sharifnia,T., Okamoto,M., Yang,X., Hinkle,G., Boehm,J.S., Beroukhim,R., Weir,B.A. *et al.* (2008) Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 20380–20385.
12. Nagy,A. (2000) Cre recombinase: the universal reagent for genome tailoring. *Genesis*, **26**, 99–109.
13. Cho,A., Haruyama,N. and Kulkarni,A.B. (2009) Generation of transgenic mice. *Curr. Protoc. Cell Biol.*, **19**, doi:10.1002/0471143030.cb1911s42.
14. Wang,T., Birsoy,K., Hughes,N.W., Krupczak,K.M., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
15. Peters,J.M., Colavin,A., Shi,H., Czarny,T.L., Larson,M.H., Wong,S., Hawkins,J.S., Lu,C.H., Koo,B.M., Marta,E. *et al.* (2016) A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell*, **165**, 1493–1506.
16. Ma,H., Dang,Y., Wu,Y., Jia,G., Anaya,E., Zhang,J., Abraham,S., Choi,J.G., Shi,G., Qi,L. *et al.* (2015) A CRISPR-Based screen identifies genes essential for west-nile-virus-induced cell death. *Cell Rep.*, **12**, 673–683.
17. Chen,X., Li,M., Feng,X. and Guang,S. (2015) Targeted chromosomal translocations and essential gene knockout using CRISPR/Cas9 technology in caenorhabditis elegans. *Genetics*, **201**, 1295–1306.
18. Li,W., Xu,H., Xiao,T., Cong,L., Love,M.I., Zhang,F., Irizarry,R.A., Liu,J.S., Brown,M. and Liu,X.S. (2014) MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, **15**, 554.
19. Evers,B., Jastrzebski,K., Heijmans,J.P., Grernrum,W., Beijersbergen,R.L. and Bernards,R. (2016) CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat. Biotechnol.*, **34**, 631–633.
20. D'Elia,M.A., Pereira,M.P. and Brown,E.D. (2009) Are essential genes really essential? *Trends Microbiol.*, **17**, 433–438.

21. Hart,T., Chandrashekhar,M., Aregger,M., Steinhart,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S. *et al.* (2015) High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, **163**, 1515–1526.
22. Chen,W.H., Minguez,P., Lercher,M.J. and Bork,P. (2012) OGEE: an online gene essentiality database. *Nucleic Acids Res.*, **40**, D901–D906.
23. Chen,W.-H., Trachana,K., Lercher,M.J. and Bork,P. (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol. Biol. Evol.*, **29**, 1703–1706.
24. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
25. Makino,T., Hokamp,K. and McLysaght,A. (2009) The complex relationship of gene duplication and essentiality. *Trends Genet.*, **25**, 152–155.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Chen,W.-H., Zhao,X.-M., van Noort,V. and Bork,P. (2013) Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput. Biol.*, **9**, e1003073.
28. Wang,D., Zhang,Y., Zhang,Z., Zhu,J. and Yu,J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, **8**, 77–80.
29. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
30. 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.
31. Coordinators,N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
32. Zhao,N., Liu,Y., Wei,Y., Yan,Z., Zhang,Q., Wu,C., Chang,Z. and Xu,Y. (2016) Optimization of cell lines as tumour models by integrating multi-omics data. *Brief. Bioinform.*,
33. Domcke,S., Sinha,R., Levine,D.A., Sander,C. and Schultz,N. (2013) Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.*, **4**, 2126.
34. Luo,H., Lin,Y., Gao,F., Zhang,C.T. and Zhang,R. (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.