GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Predicting bacterial essential genes using only sequence composition information

L.W. Ning, H. Lin, H. Ding, J. Huang, N. Rao and F.B. Guo

Center of Bioinformatics and Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: F.B. Guo
E-mail: fbguo@uestc.edu.cn

Genet. Mol. Res. 13 (2): 4564-4572 (2014)
Received May 29, 2013
Accepted September 19, 2013
Published June 17, 2014
DOI http://dx.doi.org/10.4238/2014.June.17.8

ABSTRACT. Essential genes are those genes that are needed by organisms at any time and under any conditions. It is very important for us to identify essential genes from bacterial genomes because of their vital role in synthetic biology and biomedical practices. In this paper, we developed a support vector machine (SVM)-based method to predict essential genes of bacterial genomes using only compositional features. These features are all derived from the primary sequences, i.e., nucleotide sequences and protein sequences. After training on the multiple samplings of the labeled (essential or not essential) features using a library for SVM, we obtained an average area under the ROC curve (AUC) of about 0.82 in a 5-fold cross-validation for *Escherichia coli* and about 0.74 for *Mycoplasma pulmonis*. We further evaluated the performance of the method proposed using the dataset consisting of 16 bacterial genomes, and an average AUC of 0.76 was achieved. Based on this training dataset, a model for essential gene prediction was established. Another two independent genomes, *Shewanella oneidensis* RW1 and *Salmonella enterica* serovar Typhimurium SL1344 were used to evalutate the model. Results showed that the AUC sores were 0.77 and 0.81, respectively. For the convenience of the vast majority

of experimental scientists, a web server has been constructed, which is freely available at http://cefg.uestc.edu.cn:9999/egp.

Key words: Essential gene prediction; Compositional features; Library for support vector machine; Bacteria

INTRODUCTION

These days, studies on essential genes are becoming increasingly important with the emergence and development of synthetic biology. A gene is considered to be essential only if it is essential under any growth conditions. Under this definition, genes that are needed for some specific conditions such as aerobic or anaerobic conditions (Sassetti et al., 2001) should be considered non-essential genes. Usually, essential genes are identified in the most optimal conditions, where the minimal functional gene set would be expected. Essential genes are crucial for several reasons. First, essential genes are perfect promising targets of drugs against pathogenic bacteria (Judson and Mekalanos, 2000; Juhas et al., 2012). Second, they are the foundation for revealing the minimal gene set for living organisms, and this is very important in synthetic biology (Juhas et al. 2011). Third, this kind of genes is important for understanding the origin of life and evolutionary relationships among species (Jordan et al., 2002).

Several experimental methods can be used to identify essential genes of bacteria, such as systematic random mutation, RNA interference, single gene knockout, conditional knockout, and transposon mutagenesis (Christen et al., 2011). Nevertheless, all of these are laborious, time consuming, and costly processes. Therefore, to date, the essential genes of less than 20 among thousands of sequenced bacterial genomes have been systematically experimentally identified. These data have been carefully collected and combined into databases (Zhang and Lin, 2009; Chen et al., 2012), greatly facilitating related research. With more research groups putting efforts into the systematic identification of essential genes, information about the essential genes of more bacterial strains will be available in the coming days. However, it could still not satisfy the great desire to identify essential genes in most bacteria with medical or industrial importance. Besides, some intrinsic complications exist in the experimental ways. When considering much more uncultured microorganisms (Chitsaz et al., 2011), the experimental methods for essential gene identification would further show its limitation (Acenico and Lemke, 2009; Holman et al., 2009). Considering the importance of essential genes and the limitation of experimental methods, predicting essential genes *in silico* is of paramount importance.

In 2005, Chen and Xu predicted protein dispensability in *Saccharomyces cerevisiae* using machine learning methods for the first time (Chen and Xu, 2005). Then, several bioinformatic methods became available for essential gene prediction. Most of these utilized homology information between the genes whose essentialities need to be identified and those genes whose essentialities already have been experimentally identified (Guo et al., 2010; Lin and Zhang, 2011). Given a threshold, most essential genes can be predicted through basic local alignment search tool analysis of the target genes with the essential gene database (Juhas et al., 2009). Some authors have used this principle to predict potential drug targets (Roemer et al., 2003; Sakharkar et al., 2004; Chong et al., 2006; Doyle et al., 2010). Meanwhile, other researchers used protein-protein interaction networks or genetic interaction networks to iden-

tify essential genes (del Rio et al., 2009; Dolye et al., 2010; Plaimas et al., 2010). All of these techniques require prior experimental information like gene expression levels or protein interaction information, which is only available in a limited number of bacterial species. Using only homology information, some species-specific genes may be classified into the wrong class (Deng et al., 2011b), and it takes a long time to search for the homology. Combining all of the features mentioned above to predict essential genes was also an important strategy for some methods (Gustafson et al., 2006; Roberts et al., 2007; Acenico and Lemke, 2009; Deng et al., 2011a; Lin and Zhang, 2011). If a query essential gene has no homologs in public databases and no prior experimental information, it is impossible to predict its essentiality. A more promising way is to use sequence-derived information to predict essential genes. Sequence composition has already been extensively used to predict other types of genes. The use of sequence compositional information could provide a good alternative in bacterial essential gene prediction because it is economic, fast, and possibly accurate.

Here we attempted to predict essential genes only using parameters derived from the primary sequences. By combining 5 types of sequence compositional features, we obtained an average area under the ROC curve (AUC) that was above 0.9 in self-tests and above 0.7 in 5-fold cross-validation tests in both *Escherichia coli* and *Mycoplasma pulmonis*. A web server was constructed to facilitate the identification of essential genes in complete bacterial genomes or individual sequences. This study will shed light on essential gene prediction in bacteria and other life domains.

MATERIAL AND METHODS

Essential gene dataset and genome sequences

The essential gene datasets were downloaded from the database of essential genes (DEG) (http://tubic.tju.edu.cn/deg/) (version 6.0). DEG is a database of essential genes that collects the essential genes identified by high-throughput experiments in prokaryotes and eukaryotes (Zhang and Lin, 2009). Genome sequences and annotation information were downloaded from the National Center for Biotechnology Information (NCBI) RefSeq database (ftp://ftp.ncbi.nih.gov/genome) in August 2010. Two strains of *Escherichia coli* MG1655 and *Mycoplasma pulmonis* UAB CTIP were analyzed. Particularly, the essential genes of *E. coli* MG1655 were obtained from the profiling of *E. coli* chromosome (PEC) database (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp) because there were 2 very different datasets for the same strain in the DEG. The gene ID from essential gene databases (DEG or PEC) was mapped to a sequence by comparing it with the NCBI RefSeq annotation. Some genes that do not have counterparts in the RefSeq annotation were removed. The information of 16 bacterial strains is listed in Table 1.

Sequence compositional features

All features are generated from primary sequences (gene or protein sequences). These features contain amino acid usage (the frequencies of 20 amino acids in every gene, 20 features), codon usage (the frequencies of codons in every gene, 64 features), nucleotide usage of 3 codon positions (the frequencies of 4 nucleotides at 3 codon positions of every gene, 4 x 3

= 12 features), di-nucleotide usage (the frequencies of 2-tuple nucleotide usage of 2 adjacent codons in 3 codon positions of every gene, 4 x 4 x 3 = 48 features), and CodonW features [features derived from the software CodonW (http://codonw.sourceforge.net/), i.e., G3s, T3s, C3s, A3s, CAI, CBI, Fop, Nc, GC, GC3s, L_sym, L_aa, Gravy, Aromo, please refer to the software homepage for detailed meanings of these abbreviations, 14 features]. The performance of these 5 types of fundamental features was investigated independently and jointly (all 158 features).

Table 1. Detailed information of 16 bacterial strains.

Organism	Accession No.	Annotated gene No.	No. of essential genes	Percentage of essential genes (%)
Acinetobacter baylyi ADP1	NC_005966	3307	498	15.1
Bacillus subtilis 168	NC_000964	4176	221	5.3
Escherichia coli MG1655	NC_000913	4145	296	7.1
Francisella novicida U112	NC_008601	1719	390	22.7
Haemophilus influenzae RD HW20	NC_000907	1657	642	38.7
Helicobacter pylori 26695	NC_000915	1573	322	20.5
Mycobacterium tuberculosis H37Rv	NC_000962	3988	614	15.4
Mycoplasma genitalium G37	NC_000908	475	378	79.6
Mycoplasma pulmonis UAB CTIP	NC_002771	782	309	39.5
Pseudomonas aeruginosa UCBPP PA14	NC_008463	5892	335	6.0
Salmonella enterica serovar Typhi	NC_004631	4314	353	8.2
Salmonella typhimurium LT2	NC_003197	4423	229	5.2
Staphylococcus aureus N315	NC_007795	2891	351	12.1
Staphylococcus aureus NCTC 8325	NC_002745	2583	302	11.7
Streptococcus pneumoniae	NC_003028	2105	111	5.3
Vibrio cholerae N16961*	NC_002505	2741	565	20.6

*Only essential genes located one chromosome I was used for *Vibrio cholerae* N16961.

Support vector machine (SVM)

SVM is one of the most commonly supervised learning methods for classification and regression analysis. SVM can be implemented with the software toolbox LibSVM 3.1, which was written by Lin (Chang and Lin, 2011). It is open sourced and can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. Usually, 4 kinds of kernel functions, i.e., linear, polynomial, sigmoid, and radial basis function (RBF), are available to perform training and predicting. After carefully examining these kernel functions with various parameters, we found that the kernel function RBF achieved the highest accuracy. AUC was used to assess the trained model. Furthermore, sensitivity and specificity were also calculated.

CD-HIT

It is well known that the benchmark dataset with high sequence similarity always contains redundancy, which can overestimate the performance and reduce the generalization ability of a proposed model. CD-HIT is a widely used program for removing the redundant (similar) sequences (Li and Godzik, 2006). It can be freely downloaded from http://weizhong-lab.ucsd.edu/cd-hit/. In this study, we used CD-HIT-EST with the default parameters (i.e., C = 0.95, N = 8) to eliminate redundant sequences.

RESULTS

Prediction of essential genes in *E. coli*

E. coli MG1655 has 302 essential genes and 3843 non-essential genes according to the PEC annotation (all other genes that are annotated in GenBank are considered non-essential). The difference between the number of essential and non-essential genes is so great that it is very hard for any machine learning algorithm to obtain a balanced result (Provost, 2000). There are two common methods to solve the problem. One is over-sampling the minor sample, and the other is under-sampling the major sample. Here, the under-sampling strategy was adopted because a lot of time is required by SVM for over-sampling a large dataset. All essential genes were taken as the positive dataset. Non-essential genes were randomly under-sampled to the same size as the positive dataset. In total, 80 random samplings were performed for *E. coli*. After obtaining the negative and positive dataset, we retrieved the features of each gene as described in the Methods section. Then, the performance of each group of features was estimated by SVM. A 5-fold cross-validation and self-test were used to evaluate the model's performance. For each randomly sampled dataset, the AUCs of the self-test and 5-fold cross-validation were obtained. The average AUCs for 80 replicates were listed in Table 2.

Table 2. Results of self-test and 5-fold cross validation in *Escherichia coli* MG1655.

Features	Self-test			5 x CV
	AUC	Sensitivity	Specificity	AUC
Nucleotide	0.8484	0.7878	0.7387	0.7757
Codon	0.8872	0.8048	0.8098	0.7665
CodonW	0.8524	0.7352	0.8026	0.8068
AA	0.8712	0.8045	0.7838	0.7741
Dinucleotide	0.8914	0.8323	0.7959	0.7689
Combine	0.9595	0.9093	0.8996	0.8168

Prediction of essential genes in *M. pulmonis*

The process of essential gene predication in *M. pulmonis* is similar to that of *E. coli*. According to DEG, *M. pulmonis* has 309 essential genes and 473 non-essential genes. The same under-sampling strategy was used to obtain balanced positive and negative datasets. A total of 20 random samplings were used to obtain the training dataset. The average AUCs of 20 datasets were calculated and listed in Table 3.

Table 3. Results of self-test and 5-fold cross validation in *Mycoplasma pulmonis*.

Features	Self-test			5 x CV
	AUC	Sensitivity	Specificity	AUC
Nucleotide	0.8850	0.8528	0.7722	0.7016
Codon	0.7462	0.8095	0.5311	0.6931
CodonW	0.8515	0.8116	0.7315	0.7186
AA	0.8865	0.8102	0.8102	0.7230
Dinucleotide	0.8693	0.8225	0.7112	0.7125
Combine	0.9532	0.9194	0.8844	0.7371

Prediction of essential genes in all 16 genomes

According to DEG, there are a total of 7215 essential genes and 47,846 non-essential genes (genes that are annotated in GenBank but not in DEG are considered to be non-essential genes here). After using CD-HIT on the essential and non-essential gene dataset with the default parameters (c = 0.95, N = 8), the number of essential genes was reduced to 5985, and the number of non-essential genes was reduced to 35,492. The under-sampling strategy was also used to deal with the unbalanced dataset. We randomly sampled the same number of non-essential genes as essential genes to construct the negative datasets. Because of the computation limitations and to save time, a total of 6 random samples were used. A predictive model was generated, and a 5-fold cross-validation was used to assess the model. The 5-fold cross-validation results were listed in Table 4.

Table 4. Results of self-test and 5-fold cross validation on the combined dataset of 16 bacterial genomes.

Features	Self-test			5 x CV
	AUC	Sensitivity	Specificity	AUC (5 X)
Nucleotide	0.7407	0.6553	0.7021	0.6968
Codon	0.8379	0.7263	0.7976	0.7575
CodonW	0.7549	0.6470	0.7198	0.7175
AA	0.7863	0.6913	0.7367	0.7009
Dinucleotide	0.7862	0.7029	0.7296	0.7023
Combine	0.8345	0.7263	0.7880	0.7583

Web server for predicting essential genes

Based on the model that was built on all 16 bacterial genomes, we constructed a web server for bacteria essential gene prediction. Six different models were combined by averaging the probabilities of the prediction. The server is available at http://cefg.uestc.edu.cn:9999/egp. Only nucleotide sequences of genes are needed to make a prediction. Users are required to either upload a sequence file or copy and paste the query gene sequence into the text box. The input sequence should be in FASTA format. After clicking on the ”Submit” button, the user may check the results on the jumping window. If an e-mail address is provided, an attachment containing the predicted results will be sent to it.

We used the server to predict essential genes in 2 other independent species; another independent species *Shewanella oneidensis*, whose essentiality data were collected from DEG version 10.0 just recently (1), has been used to test the performance of EGP. Running result shows that that the AUC scores were 0.77. The strain *Salmonella enterica* serovar Typhimurium *SL1344* was also used as an independent test set and it attains the AUC score of 0.81. In fact, the former independent testing species belongs to the same class with some of the training species, whereas it does not belong to the same order with any of them. The latter independent testing genome belongs to the same species with some of the training species. So it is thought that the EGP tool may be applicable to each bacterial genome that belongs to the same family with one of the 16 training strains. When predicting the essentiality of an anonymous single gene, the user is needed to only provide its DNA sequence in FASTA format.We suggest that our other web server, Geptop (http://cefg.uestc.edu.cn/geptop), is used when the complete ge-

nome has been sequenced (Wei et al., 2013). Except EGP and Geptop, there are not any other web servers to predict essentiality.

DISCUSSION

In this study, we investigated the performance of 5 types of compositional features for predicting essential genes using SVM. Five-fold cross-validation results exhibited that the AUCs achieved by these features were all greater than 0.7, demonstrating that it is feasible to use primary sequences to predict essential genes. The combination of these features achieved the highest AUC in all the 3 different datasets. These features are not linearly independent. Using principal component analysis, the first 80 components accumulated 99% of the variants from the original 158 features; however, when only the first 80 components were selected as features, the AUC of the 5-fold cross-validation declined significantly.

In 2006, Seringhaus et al. (2006) combined 7 different machine learning methods to predict essential genes in fungal genomes using 14 sequence compositional features. Some of the predicted essential genes were chosen and validated by experiments. That was the first report in which only compositional features were used in essential gene prediction, but the accuracy was not satisfactory. Here, we systematically evaluated 5 different types of basic compositional features. We found that the AUCs in the 5-fold cross-validation were encouraging and exciting. We also considered pseudo amino acid composition, which includes sequence-order information besides composition (Chou, 2001) as features. However, the result was not significantly improved over the method that we described above when we trained and tested them.

Many other features such as genetic interaction, protein-protein interaction, metabolic networks and gene expression patterns can be used in the prediction (del Rio et al., 2009; Hwang et al., 2009; Manimaran et al., 2009; Wang et al., 2012). Our results were comparable with or even better than these features. For example, when adopting 3 types of features derived from sequence, i.e., phyletic retention, strand type, and open reading frame length, Hwang obtained a precision of 82.6% and a recall of 74.3% in the *E. coli* K12 genome based on a 10-fold cross-validation. In our method, the precision is 74.6% and the recall is 66.8% for *E. coli* K12 based on a 10-fold cross validation. When using only topological properties of the protein interaction network, a precision of 58.4% and a recall of 46.4% were obtained for the same strain. Therefore our method is comparative with published methods. On the other hand, these experimental features are unavailable for most bacterial species. In the future, combining these experimental features with the features derived from sequence composition will improve the predictive model (Roberts et al., 2007; Deng et al., 2011a).

Although many features are associated with essential genes, the intrinsic complication of life makes the accurate identification of essential genes in all bacterial species difficult. For example, current experimental techniques can only identify essential genes independently in certain environments. However, when considering 2 or more non-essential genes that can be disrupted separately, disrupting them at the same time may be lethal to the bacteria. This may partly explain the existence of different essential gene datasets that are identified by different research groups, as in the case of *E. coli* MG1655.

This study shed light on essential gene prediction. It employs only sequence-based information without other experimental information. However, just due to few information is required, it is applicable only to some certain phylogenetic lineages. Be cautious to use it when

your input gene belongs to the host that do not be included in the same family with any of the reference species, which have been used in the training set of EGP. A web server for predicting essential genes was constructed for convenience and its practical purposes.

ACKNOWLEDGMENTS

Research supported by the program for New Century Excellent Talents in University (#NCET-11-0059), the Key Technology Research and Development Program of Sichuan Province (Grant #2011FZ0034), the National Natural Science Foundation of China (Grant #31071109 and #60801058), and the China Postdoctoral Science Foundation (Grants #201104687 and #2013M540705). Thanks to Dr. Xianlong Wang for kindly assisting in the use of the FOBOS cluster server for most calculations that were done in this research.

REFERENCES

Acenico ML and Lemke N (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* 10: 290.

Chang C-C and Lin C-J (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 27.

Chen WH, Minguez P, Lercher MJ and Bork P (2012). OGEE: an online gene essentiality database. *Nucleic Acids Res.* 40: D901-D906.

Chen Y and Xu D (2005). Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* 21: 575-581.

Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, et al. (2011). Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* 29: 915-921.

Chong CE, Lim BS, Nathan S and Mohamed R (2006). *In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico Biol.* 6: 341-346.

Chou KC (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43: 246-255.

Christen B, Abeliuk E, Collier JM, Kalogeraki VS, et al. (2011). The essential genome of a bacterium. *Mol. Syst. Biol.* 7: 528.

del Rio G, Koschutzki D and Coello G (2009). How to identify essential genes from molecular networks? *BMC Syst. Biol.* 3: 102.

Deng J, Deng L, Su S, Zhang M, et al. (2011a). Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res.* 39: 795-807.

Deng J, Tan L, Lin X, Lu Y, et al. (2011b). Exploring the optimal strategy to predict essential genes in microbes. *Biomolecules* 2: 1-22.

Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, et al. (2010). Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* 11: 222.

Guo FB, Ning LW, Huang J, Lin H, et al. (2010). Chromosome translocation and its consequence in the genome of *Burkholderia cenocepacia* AU-1054. *Biochem. Biophys. Res. Commun.* 403: 375-379.

Gustafson AM, Snitkin ES, Parker SC, DeLisi C, et al. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 7: 265.

Holman AG, Davis PJ, Foster JM, Carlow CK, et al. (2009). Computational prediction of essential genes in an unculturable endosymbiotic bacterium, *Wolbachia* of *Brugia malayi*. *BMC Microbiol.* 9: 243.

Hwang YC, Lin CC, Chang JY, Mori H, et al. (2009). Predicting essential genes based on network and sequence analysis. *Mol. Biosyst.* 5: 1672-1678.

Jordan IK, Rogozin IB, Wolf YI and Koonin EV (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12: 962-968.

Judson N and Mekalanos JJ (2000). TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* 18: 740-745.

Juhas M, van der Meer JR, Gaillard M, Harding RM, et al. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* 33: 376-393.

Juhas M, Eberl L and Glass JI (2011). Essence of life: essential genes of minimal genomes. *Trends Cell Biol.* 21: 562-568.

Juhas M, Eberl L and Church GM (2012). Essential genes as antimicrobial targets and cornerstones of synthetic biology. *Trends Biotechnol.* 30: 601-607.

Li W and Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

Lin Y and Zhang RR (2011). Putative essential and core-essential genes in *Mycoplasma* genomes. *Sci. Rep.* 1: 53.

Manimaran P, Hegde SR and Mande SC (2009). Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages. *Mol. Biosyst.* 5: 1936-1942.

Plaimas K, Eils R and Konig R (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.* 4: 56.

Provost F (2000). Machine Learning from Imbalanced Data Sets 101. In Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets.

Roberts SB, Mazurie AJ and Buck GA (2007). Integrating genome-scale data for gene essentiality prediction. *Chem. Biodivers.* 4: 2618-2630.

Roemer T, Jiang B, Davison J, Ketela T, et al. (2003). Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. *Mol. Microbiol.* 50: 167-181.

Sakharkar KR, Sakharkar MK and Chow VT (2004). A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico Biol.* 4: 355-360.

Sassetti CM, Boyd DH and Rubin EJ (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U. S. A.* 98: 12712-12717.

Seringhaus M, Paccanaro A, Borneman A, Snyder M, et al. (2006). Predicting essential genes in fungal genomes. *Genome Res.* 16: 1126-1135.

Wang J, Li M, Wang H and Pan Y (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9: 1070-1080.

Wei W, Ning LW, Ye YN and Guo FB (2013). Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One* 8: e72343.

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.

