Chapter 15

Discovering Essential Domains in Essential Genes

Yulan Lu, Yao Lu, Jingyuan Deng, Hui Lu, and Long Jason Lu

Abstract

Genes with indispensable functions are identified as essential; however, the traditional gene-level perspective of essentiality has several limitations. We hypothesized that protein domains, the independent structural or functional units of a polypeptide chain, are responsible for gene essentiality. If the essentiality of domains is known, the essential genes could be identified. To find such essential domains, we have developed an EM algorithm-based Essential Domain Prediction (EDP) Model. With simulated datasets, the model provided convergent results given different initial values and offered accurate predictions even with noise. We then applied the EDP model to six microbes and predicted 3,450 domains to be essential in at least one species, ranging 8–24 % in each species.

Key words Essential genes, Domains, Essentiality, Synthetic biology, EM algorithm

1 Introduction

Genes are widely regarded as the basic units of a cell, a complex system made up of a large number of components and reactions. Therefore, a fundamental question in synthetic biology is: what is the minimal gene set that is necessary and sufficient to sustain life [1]? The individual genes that constitute a minimal gene set are called *essential genes*. Experimentally, essential genes are defined as those that when disrupted, confer a lethal phenotype to organisms under defined conditions. Therefore, the essentiality of a gene is the indispensability of this gene's product to the survival of a microorganism. Systematic genome-wide interrogations of essential genes have been conducted by single-gene knockouts [2–5], transposon mutagenesis [6–12], or antisense RNA inhibitions [13, 14]. These experiments have provided a tremendous amount of resources to further our understanding on gene essentiality, one important step closer to unraveling the complex relationship between genotype and phenotype [15].

Long Jason Lu (ed.), *Gene Essentiality: Methods and Protocols*, Methods in Molecular Biology, vol. 1279,
DOI 10.1007/978-1-4939-2398-4_15,

Recent comparative research on the available essential gene datasets has shown surprising results and has challenged many early assumptions in genomics. Firstly, it was discovered that microorganisms share a limited set of essential genes. Independent studies have firmly established that bacterial species share a very limited number of orthologs, regardless of which ortholog detection method is used [16]. This may reflect the physiology of diverse microorganisms expanding the organism list will further decrease the number of orthologs and thus common essential genes. Secondly, and more surprisingly, a recent study showed that when tested experimentally in model bacteria, less than a quarter of the highly conserved genes were essential [17–20]. This suggests that evolutionary conservation of a gene does not necessarily imply that it is essential for microbial survival. Finally, orthologs are often observed to be essential in one organism but not another. For example, the *dapE* gene is essential in *E. coli* but nonessential in *P. aeruginosa* [21]. It is also possible for orthologs to have different functions in different organisms [22], even though it is a fundamental assumption of genomics that most orthologs perform a similar function. This suggests that differences in genetic regulation, genetic redundancy, and divergence in cellular pathways or processes between organisms may all affect gene essentiality; their combined effects result in the discrepancy in essentiality between orthologs.

Here we have reexamined gene essentiality from a novel essential protein domain point of view. With an EM algorithm-based Essential Domain Prediction (EDP) model, we evaluate the contribution of domains in the essentiality of gene. The performance of EDP model are tested on simulated data sets and then used to predict essential domains in six microbes. Our results suggest that this new perspective may offer unique insights into the mechanistic basis of gene essentiality and help to resolve the controversy regarding this phenomenon.

2 Materials

2.1 Essential Gene Data

E. coli K-12 sequence data were downloaded from Comprehensive Microbial Resource (CMR) (http://cmr.jcvi.org/tigrscripts/CMR/GenomePage.cgi?database=ntec01). This database contains 4,289 protein sequences in total [23]. The essential genes of *E. coli K-12* were downloaded from the PEC database [5]. The Kato data set contains 302 essential genes from gene deletion experiments.

P. aeruginosa PAO1 sequence data were downloaded from the Pseudomonas Genome Database (http://www.pseudomonas.com/) (Pseudomonas_aeruginosa_PAO1.faa, revision 2009-07-17). *PA* essential genes were adopted from ref. 24. The Jacobs data set

contains 678 essential genes from transposon mutagenesis experiments in PAO1.

A. baylyi ADP1 sequences were collected from the Magnifying Genomes Database (http://www.genoscope.cns.fr/agc/mage). Out of a total of 3,308 genes, 499 essential genes were acquired from ref. 2.

B. subtilis sequence data were downloaded from Microbial Genome Database (http://mbgd.genome.ad.jp/). This data set contains 4,117 ORFs. The essential gene list was acquired from ref. 25 and consists of three data sets: 150 essential genes determined by Kobayashi's mutant genesis experiments, 42 known essential genes from previous studies, and 79 essential genes by homology mapping to other bacteria, most of which encode proteins involved in ribosome or synthesis.

S. cerevisiae sequences were downloaded from *Saccharomyces* Genome Database (http://downloads.yeastgenome.org/sequence/genomic_sequence/). This resource contains 5,885 ORFs. The essential gene list was obtained from ref. 26. This data set contains 1,049 essential genes from targeted mutagenesis experiments.

N. crassa ORFs were downloaded from *Neurospora crassa* database at Broad Institute (http://www.broadinstitute.org/annotation/genome/neurospora/MultiDownloads.html). Dubious ORFs and pseudo genes were excluded from this list. Essential gene data was kindly provided by K. Borkovich from the systematic genome deletion project in *N. crassa* at UC Riverside. This list contains 1,251 essential genes.

Gene ontology (GO) annotations for domains were downloaded from the Mappings of External Classification Systems to GO (http://www.geneontology.org/GO.indices.shtml).

2.2 The Domain Data Set and Data Filtering

We used InterPro (http://www.ebi.ac.uk/interpro/) [27] and Pfam (http://pfam.sanger.ac.uk/) [28] to derive domain information from protein sequences. Because the InterPro database already includes Pfam-A entries, we combined Pfam-B entries with the InterPro data to construct our domain data set. In total, 9,689 InterPro and 5,098 Pfam-B domains were included in our analysis. Genes that had no domain annotation were excluded, leaving a total of 26,302 genes. Specific numbers of genes and domains for each species are shown in Table 1.

3 Methods

3.1 The Essential Domain Predictor (EDP) Model

Assuming the genome contains n different genes, we defined $G = \{g_1, g_2, \ldots, g_n\}$. For g_i, $i \in (1, n)$, let $g_i = 1$ if the i-th gene is essential, and 0 otherwise. The vector G was obtained from the experiments, and thus was treated as observed data in our model. Suppose gene g_i contains n_i different domains which form the set

Table 1
Details for gene and domain data sets
The number of domains and genes in each data set are shown, as well as the essential count of each type

essential/non-essential		AB	EC	PA	BS	NC	SC
Domain	Interpro	901/2473	597/4071	676/3740	336/3399	1032/3072	855/3423
	PfamB	67/577	56/678	75/1046	33/760	584/2025	343/1090
Gene		487/2432	286/3724	619/4560	190/3434	1165/4504	1010/3891

$D(g_i) = \{D_i^1, D_i^2, \ldots, D_i^{n_i}\}$. Here variable $D_i^j = 1$ if this domain is essential, and 0 otherwise, where (i, j) denotes the j-th domain of the i-th gene. These D_i^j values are unobserved from the experiments, need to be predicted from the model, and are treated as missing values. Each protein may include several distinct domains and each domain may occur in different proteins. Suppose that a genome contains a total of m unique domains denoted $D = \{D_1, D_2, \ldots, D_m\}$, where $D_k = 1$, $k \in (1, m)$ if the k-th unique domain is essential, and 0 otherwise. We also define $S = \{S_1, S_2, \ldots, S_m\}$, where S_k, $k \in (1, m)$ is the set of domains $D_m{}^n s$ that are equal to the kth unique domain D_k. We use $|S_k|$ to denote the size of the set S_k, and we further define δ_k that as the probability that domain D_k is essential.

We also need to describe two kinds of errors that may exist in the prediction process: falsely predicted essential rate (FER) and falsely predicted nonessential rate (FNR). These can be defined as follows:

$$\text{FER} = \Pr(p_i = 1 \mid g_i = 0),\ \text{FNR} = \Pr(p_i = 0 \mid g_i = 1), \quad (1)$$

where $p_i = 1$ if the ith gene is predicted to be essential and 0 otherwise.

Our model also needs two assumptions (*see* **Note 1**):

Assumption I: The essentialities of the domains are independent, which means that the event that one domain is essential is not depend on the essentiality of others.

Assumption II: A gene is defined as essential if and only if at least one of its domains is essential.

The goal of this model is to estimate the parameters set θ to maximize the likelihood of observed essential genes. Because $L(G|\theta)$ is difficult to optimize directly, we augment the observed likelihood $L(G|\theta)$ with missing data D, and the complete data likelihood is thus defined as: $L(G, D|\theta) = L(D|\theta)L(G|D, \theta)$. We further derive the formula as:

$$L(D \mid \theta) = \prod_{k=1}^{m} L_k \quad \text{where}, L_k = \begin{cases} \delta_k, & D_k = 1 \\ 1 - \delta_k, & D_k = 0 \end{cases} \quad (2)$$

$$L(G \mid D,\theta) = \prod_{i=1}^{n} L_i,$$

$$\text{where } L_i = \begin{cases} (1-\text{FNR}) : g_i = 1 \text{ and one } D_i^j \in D(g_i) = 1 \\ \text{FER} : g_i = 1 \text{ and all } D_i^j \in D(g_i) = 0 \\ \text{FNR} : g_i = 0 \text{ and one } D_i^j \in D(g_i) = 1 \\ 1-\text{FER} : g_i = 0 \text{ and all } D_i^j \in D(g_i) = 0 \end{cases} \quad (3)$$

where FER and FNR are defined as before and $\theta = (\delta_k, \text{fer}, \text{fnr})$. Under this framework, we adopted the conventional Expectation-Maximization (EM) algorithm [29] to computer the optimal θ that maximizes $L(G|\theta)$. We derived the EM algorithm as follows:

E-step: during the E step of the t-th iteration, D is updated by the conditional expectation given the estimated θ from last iteration θ_{t-1} and G, that is:

$$\begin{aligned} p_i^j(t) &= E[D_i^j \mid G, \theta^{(t-1)}] = \Pr(D_i^j = 1 \mid G, \theta^{(t-1)}) \\ &= \frac{\Pr(D_i^j = 1 \mid \theta^{(t-1)}) \Pr(g_i \mid D_i^j = 1, \theta^{(t-1)})}{\Pr(g_i \mid \theta^{(t-1)})} \\ &= \frac{\delta_i^j(t-1) \Pr(g_i \mid D_i^j = 1, \theta^{(t-1)})}{\Pr(g_i \mid \theta^{(t-1)})}, i \in (1,n), D_i^j \in D(g_i) \end{aligned} \quad (4)$$

M step: update θ using maximum likelihood estimation (MLE) approach (*see* **Note 2**).

$$\begin{aligned} \theta^{(t)}(\delta_k, fer, fnr) &= \max_{\theta} L(\theta; G, D) \\ \delta_k(t) &= \frac{\sum_{D_m^n \in S_k} p_m^n(t)}{|S_k|}, \forall k \in (1,m) \\ FER(t) &= \frac{\sum_{i=1}^{n} (1-g_i)[1 - \prod_{D_i^j \in D(g_i)} (1 - D_i^j(t))]}{\sum_{i=1}^{n} [1 - \prod_{D_i^j \in D(g_i)} (1 - D_i^j(t))]} \\ FNR(t) &= \frac{\sum_{i=1}^{n} g_i \prod_{D_i^j \in D(g_i)} (1 - D_i^j(t))}{\sum_{i=1}^{n} \prod_{D_i^j \in D(g_i)} (1 - D_i^j(t))} \end{aligned} \quad (5)$$

Each domain group receives a probability score δ_j indicating its likelihood of being essential. In this study those with $\delta_i \geq 0.9$ were classified as essential. The cutoff value was obtained by minimizing the sum of false positive essential gene predictions and false negative predictions.

3.2 Running the EDP Model with Different Initial Values

The results of the EM-algorithm may be different if the initial values of the model are changed. For the EDP Model, three parameters are given by initial values: the probability that one domain is essential $D = \{D_1, D_2, \ldots, D_m\}$, the falsely predicted essential rate

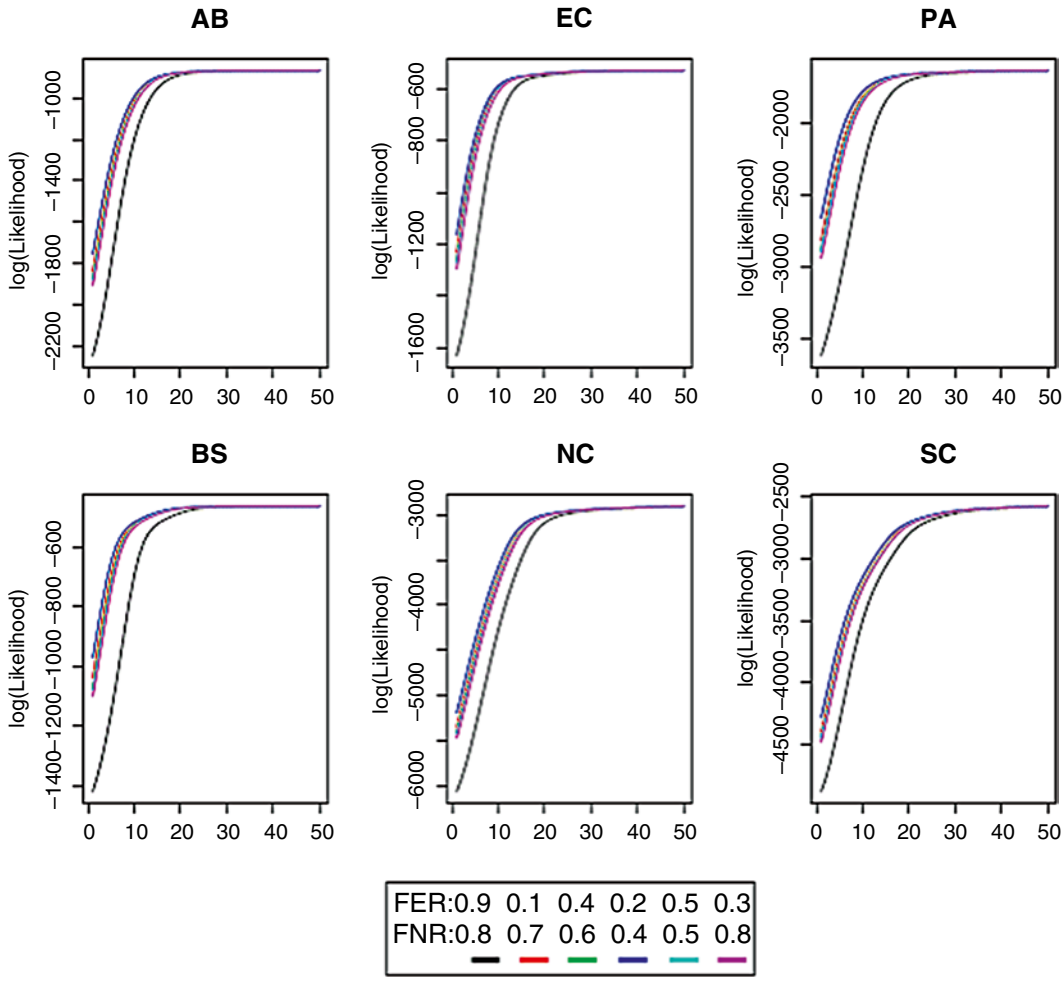

Fig. 1 The convergent process of the EDP Model in real data sets. Shown in the figure are the likelihood result of the EM algorithm performed in six microbes, processes with different initial value are compared. The EDP Model converges quickly for all six microbe data sets. For each species, the calculation provided the same results when given different initial values (each shown in a separate color)

(FER) and falsely predicted nonessential rate (FNR). The initial value of D_m is set equal to the percentage of essential genes in which domain m can be found. We fixed the initial value of D and tested the influence of different FERs and FNRs on the results of the EDP Model. Six pairs of FER and FNR values were tested, and the final result of the EDP Model was convergent as long as the data set remained constant (Fig. 1).

3.3 Testing the Essential Domain Prediction (EDP) Model Using Simulated Data Sets

1. To test the performance of the EDP Model, we generated 20 independent simulated data sets, each of which contains 1,200 genes and 1,000 domains. A previous study has shown that the number of domains for each gene follows a power-law-like distribution [30], and we found the same distribution in the

combined gene-domain annotation for the six microbes. Therefore, we required that the degree of both genes and domains in the simulated data sets follow a power-law distribution. For each data set, we randomly assigned a certain number of essential domains, and then assigned essential genes based on the assumption I. Among the 20 simulated data sets, the number of essential domains ranged from 50 to 83, while the number of essential genes ranged from 120 to 164.

2. We then applied the EDP Model to these simulated data sets, revealing only the essential gene labels and gene-domain association. When given different initial values, the EDP Model produced convergent results for each simulated data set. The results from the 20 simulated data sets were then compared to the original assignment of essential domains (Fig. 2a). All preassigned essential domains were correctly predicted, i.e., no false negative (FN) predictions, while the false positive rate (FPR) is ~0.4. All false positive (FP) assignments were the results of the same scenario: a domain appeared in only one gene and that gene was essential. Additionally, we annotated the essentiality of genes reciprocally based on the predicted essential domains, and no false predictions were made (Fig. 2a).
3. Next, we added noise to the simulated data and repeated the prediction process. The noise data includes ten genes, five were annotated as essential but contained no essential domains, and the remaining five were annotated as nonessential but con-

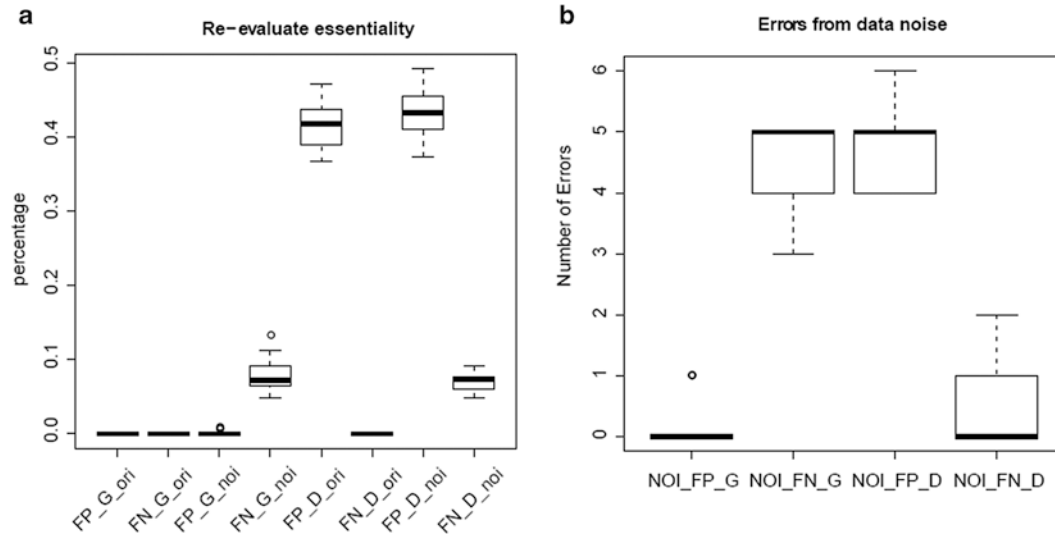

Fig. 2 A comparison between the input and calculated essentiality of domains and genes. (**a**) Shows the False Positive (FP) and False Negative (FN) errors for both essential genes (*G*) and domains (*D*), with or without noise (noi and ori, respectively). (**b**) Shows the number of errors that are the direct result of noise data

tained at least one essential domain. When adding "noise genes" to the data set, the number of domains that a noise gene contained was randomly determined and followed the same degree distribution as the "real genes" in the data set. The errors for repeated predictions are also shown in Fig. 2b. For essential domain prediction, the number of FP errors remained almost the same and even decreased in some cases (due to the influence of noise genes). The FPR increased slightly because of the loss of true positives, which became FN errors. Each of these was caused by the addition of noise genes. As for the re-annotation of genes based on predicted essential domains, 16 simulated data sets had no incorrect essential gene assignments and four of them had only one, each of which were noise genes. The number of FN predictions ranged from 8 to 17, including 3–5 noise genes (Fig. 2b). The test on simulated data shows that the EDP model offers an accurate prediction of essential domains, even with substantial noise (*see* **Note 3**).

3.4 Predicting Microbial Essential Domains Using the EDP Model

1. After testing the predictive capability of the EDP Model on simulated data sets, we applied it to predict essential domains in real data sets, which included six microbes: E. coli (EC), A. baylyi (AB), P. aeruginosa (PA), B. subtilis (BS), S. cerevisiae (SC), and N. crassa (NC). Essential gene annotation for these species were collected and filtered, excluding genes that do not have annotated domains in Interpro or PfamB. After filtering, we obtained 26,302 genes and 14,787 domains (9,689 from Interpro and 5,098 from PfamB) in total. The number for each species is shown in Table 1. Considering that gene essentiality differs across species, we applied the EDP Model to each organism separately.
2. We first tested the influence of initial parameters of the iterative EM algorithm on its ability to converge. There are three parameters in the EDP Model: domain essentiality, falsely predicted essential rate (FER), and falsely predicted nonessential rate (FNR). Specifically, we obtained the initial essentiality of domains by computing the percentage of essential genes associated with each domain. To test the influence of changes in the FER and FNR, we generated six pairs of FERs and FNRs, and compared their convergence process to the final results within each species. As shown in Fig. 1, all processes converged within 50 steps, including the pair with an FER of 0.9 and an FNR of 0.8. This result indicated that the iterative process produces stable results for all data sets despite wide variation in the initial FER and FNR values.
3. Based on the distribution of domain essentiality scores, we set the cutoff for essentiality to 0.9 for the final prediction. Thus, when we generated the set of essential domains, we accounted for 8 ~ 24 % of the total number of domains in different species

(Table 1). We found that the number of essential domains was much larger in eukaryotes than prokaryotes. We identified 1,198 essential domains in SC and 1,616 in NC, while the numbers are 968, 653, 369 and 751 for AB, EC, BS, and PA, respectively. This difference was not caused by the variation in gene numbers; the number of genes in our prokaryotic data set was similar to that of the eukaryotic data set. For example, PA and SC have 5,179 and 4,901 genes in the data set, respectively. We interpreted this phenomenon to be a consequence of the increased complexity of eukaryotic genomes. In order to increase the number of essential functions in a genome without increasing the size (gene count), there must be an increase in functionally essential "core" components, i.e., essential domains.

4 Notes

1. The cross talk between domains may affect the EDP model's accuracy. For example, if a domain D1 is found in five genes and four of them are essential, it might be estimated as essential domain with an initial value of 0.8. However, if there is another domain D2 linked to exactly the same genes with D1, both of them might be estimated as nonessential. This is because when D1 is alone in the annotation matrix with related genes, the probability estimated for the essential gene g_i, $\Pr\left(\mathcal{J}_i \mid \theta^{(t-1)}\right)$, is less than $\Pr\left(\mathcal{J}_i \mid D_i^j = 1, \theta^{(t-1)}\right)$, which provides an underestimation for the essentiality of g_i and leads to the improvement of D1's essentiality score (Eq. 4). However, if there is another domain D2 that is highly correlated with D1, then $\Pr\left(\mathcal{J}_i \mid \theta^{(t-1)}\right)$ may not be less than $\Pr\left(\mathcal{J}_i \mid D_i^j = 1, \theta^{(t-1)}\right)$ (that is also due to the value of FNR and FPR). In this case, the essentiality score for both D1 and D2 may not increase during the iterative process, which would result in both domains being labeled as "nonessential".
2. The estimation of domain essentiality can be improved. For example, in the EDP Model, FER and FNR are estimated based on the entire data set (both essential genes and domains) (Eq. 1). However, FER and FNR influence each domain's essentiality score when reestimating the probability for each domain to be essential (Eq. 4). As a result, applying the EDP Model on two separated gene–domain associations may produce different results when this data is combined and the model is run again.
3. The performance of the EDP Model is dependent upon the quality of the data set. Although we have shown that it is capable of tolerating noise to some degree, the quality of the input data will always affect the accuracy of the predictions.

Acknowledgement

This work was supported by the exchange program fund of doctoral student under the Fudan University Graduate School (to Yulan Lu).

References

1. Mushegian A (1999) The minimal genome concept. Curr Opin Genet Dev 9(6):709–714
2. de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, Durot M, Kreimeyer A, Le Fevre F, Schachter V, Pezo V, Doring V, Scarpelli C, Medigue C, Cohen GN, Marliere P, Salanoubat M, Weissenbach J (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Mol Syst Biol 4:174. doi:10.1038/msb.2008.10
3. Kobayashi M, Tsuda Y, Yoshida T, Takeuchi D, Utsunomiya T, Takahashi H, Suzuki F (2006) Bacterial sepsis and chemokines. Curr Drug Targets 7(1):119–134
4. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006 0008
5. Kato J, Hashimoto M (2007) Construction of consecutive deletions of the Escherichia coli chromosome. Mol Syst Biol 3:132
6. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185(19):5673–5684
7. Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, Ernst S, Will O, Kaul R, Raymond C, Levy R, Chun-Rong L, Guenthner D, Bovee D, Olson MV, Manoil C (2003) Comprehensive transposon mutant library of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 100(24):14339–14344. doi:10.1073/pnas.2036282100
8. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM (2006) An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103(8):2833–2838
9. Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M, Manoil C (2007) A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. Proc Natl Acad Sci U S A 104(3):1009–1014. doi:10.1073/pnas.0606713104
10. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA III, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A 103(2):425–430
11. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286(5447):2165–2169
12. Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, Mekalanos JJ (2002) A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A 99(2):966–971
13. Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, Wall D, Wang L, Brown-Driver V, Froelich JM, C KG, King P, McCarthy M, Malone C, Misiner B, Robbins D, Tan Z, Zhu Zy ZY, Carr G, Mosca DA, Zamudio C, Foulkes JG, Zyskind JW (2002) A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. Mol Microbiol 43(6):1387–1400
14. Ji Y, Zhang B, Van Horn SF, Warren P, Woodnutt G, Burnham MKR, Rosenberg M (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science 293(5538):2266–2269
15. Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, Nislow C, Giaever G, Phillips PC, Fink GR, Gifford DK, Boone C (2010) Genotype to phenotype: a complex problem. Science 328(5977):469
16. Bruccoleri RE, Dougherty TJ, Davison DB (1998) Concordance analysis of microbial

genomes. Nucleic Acids Res 26(19): 4482–4486

17. Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E, Allet E, Fish R, Jamotte T, Curchod ML, Loferer H (1998) A genome-based approach for the identification of essential bacterial genes. Nat Biotechnol 16(9):851–856
18. Freiberg C, Wieland B, Spaltmann F, Ehlert K, Brotz H, Labischinski H (2001) Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. J Mol Microbiol Biotechnol 3(3):483–489
19. Song JH, Ko KS, Lee JY, Baek JY, Oh WS, Yoon HS, Jeong JY, Chun J (2005) Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. Mol Cells 19(3):365–374
20. Zalacain M, Biswas S, Ingraham KA, Ambrad J, Bryant A, Chalker AF, Iordanescu S, Fan J, Fan F, Lunsford RD, O'Dwyer K, Palmer LM, So C, Sylvester D, Volker C, Warren P, McDevitt D, Brown JR, Holmes DJ, Burnham MK (2003) A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. J Mol Microbiol Biotechnol 6(2):109–126
21. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A (2006) Essential genes on metabolic maps. Curr Opin Biotechnol 17(5):448–456
22. Liao BY, Zhang J (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci U S A 105(19):6987–6992
23. Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, Keyamura K, Ote T, Yamakawa T, Yamazaki Y, Mori H, Katayama T, Kato J (2005) Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. Mol Microbiol 55(1):137–149. doi:10.1111/j.1365-2958.2004.04386.x
24. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Yu NY, Hancock RE, Brinkman FS (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. Nucleic Acids Res 39(Database issue):D596–D600. doi:10.1093/nar/gkq869
25. Uchiyama I, Higuchi T, Kawai M (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res 38(Database issue):D361–D365. doi:10.1093/nar/gkp948
26. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418(6896):387–391. doi:10.1038/nature00935
27. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40(Database issue):D306–D312. doi:10.1093/nar/gkr948
28. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40(Database issue):D290–D301. doi:10.1093/nar/gkr1065
29. Hastie T, Tibshirani R, Friedman JJH (2001) The elements of statistical learning, vol 1. Springer, New York
30. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2:18