Genomic analysis of essentiality within protein networks

Haiyuan Yu, Dov Greenbaum, Hao Xin Lu, Xiaowei Zhu and Mark Gerstein

Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA

Corresponding author: Mark Gerstein (Mark.Gerstein@yale.edu).

In this article, we introduce the notion of 'marginal essentiality' through combining quantitatively the results from large-scale phenotypic experiments. We find that this quantity relates to many of the topological characteristics of protein–protein interaction networks. In particular, proteins with a greater degree of marginal essentiality tend to be network hubs (i.e. they have many interactions) and tend to have a shorter characteristic path length than others. We extend our network analysis to encompass transcriptional regulatory networks. Although transcription factors with many targets tend to be essential, surprisingly, we found that genes that are regulated by many transcription factors were usually not essential.

The functional significance of a gene, at its most basic level, is defined by its essentiality. In simple terms, an essential gene is one that, when knocked out, renders the cell unviable. Nevertheless, non-essential genes can be found to be synthetically lethal (i.e. cell death occurs when a pair of non-essential genes is deleted simultaneously). Because essentiality can be determined without knowing the function of a gene (e.g. random transposon mutagenesis [1,2] or gene-deletion [3]), it is a powerful descriptor and starting point for further analysis when no other information is available for a particular gene.

However, the definition of essentiality is not novel; Thatcher *et al.* introduced the 'marginal benefit' hypothesis [4], which states that many non-essential genes make significant but small contributions to the fitness of the cell although the effects might not be sufficiently large to be detected by conventional methods. In this article, we define systematically 'marginal essentiality' (M) as a quantitative measure of the importance of a non-essential gene to a cell. Our measure incorporates the results from a diverse set of four large-scale knockout experiments that examined different aspects of the impact of a protein on cell fitness. These four experiments measure the effect of a particular knockout on: (i) growth rate [5]; (ii) phenotype [2]; (iii) sporulation efficiency [6]; and (iv) sensitivity to small molecules [7]. These datasets are the only available large-scale knockout analyses for yeast. (There are several other smaller datasets [8–12] that have data only on a small fraction of the genome and were therefore not suitable for our analysis.)

Protein networks are characterized by four major topological characteristics: degree [number of links per node (K)], clustering coefficient (C), characteristic path length [average distance between nodes (L)] and diameter [maximum inter-node distance (D); Figure 1a] [13–16]. It has been shown that some protein networks follow power-law distributions [17,18] – that is they consist of many interconnecting nodes, a few of which have uncharacteristically high degrees (hubs). In addition, power-law distributions can be characterized as scale-free – that is the possibility for a node to have a certain number of links does not depend on the total number of nodes within the network (i.e. the scale of the network). Scale-free networks provide stability to the cell because many non-hub (i.e. leaf) genes can be disabled without greatly affecting the viability of the cell [18].

Recently, Jeong *et al.* focused on the relationship between hubs and essential genes and determined that hubs tend to be essential [19]. Fraser *et al.* also observed that the effect of an individual protein on cell fitness correlates with the number of its interaction partners [20]. In this article, we extended the previous work to marginal essentiality and performed a genome-wide analysis of essentiality within a wide variety of protein networks. Further information is available from http://bioinfo.mbb.yale.edu/network/essen.

Comparison between essential and non-essential proteins within an interaction network

We constructed a comprehensive and reliable yeast-interaction network containing 23 294 unique interactions among 4743 proteins [16,21]. In a gross comparison we found that essential proteins, as a whole, have significantly more 'links' than the non-essential proteins, validating earlier findings [19]. Specifically, essential proteins have approximately twice as many links compared with non-essential proteins (Figure 1b). We can also see from the power-law plots of the interactions of essential and non-essential genes (Figure 1c) that the essential genes have a shallower slope, indicating that a proportionately larger fraction of them are hubs.

Given that essential proteins, on average, tend to have more interactions than non-essential proteins, we determined the fraction of hubs that are essential. We define hubs as the top quartile of proteins with respect to the number of interactions (see supplementary material online); therefore, 1061 proteins are defined as hubs within the yeast network. We found ~43% of hubs in yeast are essential (Figure 3a); this is significantly higher than random expectation (20%).

Furthermore, within the interaction network, essential proteins also tend to be more cliquish (as determined from the clustering coefficient) and tend to be more closely connected to each other (as determined from the characteristic path length and diameter). Not surprisingly, the values of these topological statistics (except for the clustering coefficient) for synthetic lethal genes are between those of the essential and non-essential genes (Figure 1b).

Topological characteristics for marginal essentiality within an interaction network

We expanded our analysis to non-essential genes, analyzing the relationship between marginal essentiality and topological characteristics. Overall, we found simple, monotonic trends for all four topological characteristics (Figure 2 and supplementary Figure 2 online). In particular, we found a positive correlation with marginal essentiality for descriptors of local interconnectivity (i.e. degree and clustering coefficient) but an inverse correlation for long-distance interactions (i.e. diameter and characteristic path length). Thus, the more marginally essential a gene is the more likely it is to have a large number of interaction partners – in agreement with the conclusion of Fraser *et al.* [20]. More importantly, the greater the marginal essentiality of a protein, the more likely it will be closely connected to other proteins – as reflected by a short characteristic path length. This implies that the effect of that protein on other proteins is more direct.

Marginal essentially is correlated with a higher likelihood to be one of the 1061 protein hubs (Figure 2d). Because hubs in the protein-interaction networks have been shown to be important for cell fitness [19], this positive correlation further confirms the biological relevance of our marginal-essentiality definition.

Analysis of regulatory networks

Finally, we analyzed protein essentiality within many smaller directed networks of interacting proteins and regulatory networks (i.e. transcription factors and the target genes that they regulate) [22–26]. These networks differ from protein–protein interaction networks in that they are directed. We looked at regulatory networks from two separate perspectives: (i) the regulator population (e.g. out degree) – where we are examined a directed network of transcription factors acting on targets; and (ii) the target population (e.g. in degree) – where we analyzed the sets of target genes that are regulated by any given transcription factor.

Analyzing the regulator population, we found that essential genes contribute to a larger percentage of the more promiscuous transcription factors (Figure 3b). In analyzing the target population, we found that the targets that are associated with the fewest transcription factors have a proportionally higher number of essential genes (Figure 3c).

The results for the regulators and the targets, although seemingly contradictory, are logical. If a regulator is deleted, the expression of all its target genes will be more or less affected. Therefore, the more targets a regulator has, the more important it is. Our analysis of the regulator population has, logically, shown that the promiscuous regulators tend to be essential.

We have found that most essential genes are 'house-keeping' genes [i.e. their expression level is much higher and the fluctuation of their expression is much lower compared with non-essential genes (supplementary Table 1 online)]. Therefore, the expression of essential genes tends to have less regulation, whereas non-essential genes often use more regulators to control the expression of gene products. This might be because essential proteins perform the most basic and important functions within the cell and, consequently, always need to be switched 'on'. Their expression does not need to be regulated by many factors because this makes the essential gene dependent on the viability of more regulators, which makes the cell less stable.

Relationship between essentiality and function

Having concluded that the essentiality of a gene is directly related to its importance to the cell fitness in both interaction and regulatory networks, we examined the relationship between the number of functions of a gene and its tendency to be essential using the functional classification from the Munich information center for protein sequence (MIPS) [27]. Figure 3d shows that genes with more functions are more likely to be essential. More importantly, the likelihood of a gene being essential has a monotonic relationship with the number of its functions.

Conclusion

In this article, we have provided a comprehensive definition of 'marginal essentiality' and analyzed the tendency of the more marginally essential genes to behave as hubs. Surprisingly, we also found that hubs in the target subpopulations within the regulatory networks tend not to be essential genes.

References

1 Akerley, B.J. *et al.* (1998) Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8927–8932

2 Ross-Macdonald, P. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 402, 413–418

3 Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906

4 Thatcher, J.W. *et al.* (1998) Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 95, 253–257

5 Steinmetz, L.M. *et al.* (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.* 31, 400–404

6 Deutschbauer, A.M. *et al.* (2002) Parallel phenotypic analysis of sporulation and postgermination growth in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A.* 99, 15530–15535

7 Zewail, A. *et al.* (2003) Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3345–3350

8 Smith, V. *et al.* (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* 274, 2069–2074

9 Rieger, K.J. *et al.* (1999) Chemotyping of yeast mutants using robotics. *Yeast* 15, 973–986

10 Entian, K.D. *et al.* (1999) Functional analysis of 150 deletion mutants in *Saccharomyces cerevisiae* by a systematic approach. *Mol. Gen. Genet.* 262, 683–702

11 True, H.L. and Lindquist, S.L. (2000) A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* 407, 477–483

12 Sakumoto, N. *et al.* (2002) A series of double disruptants for protein phosphatase genes in *Saccharomyces cerevisiae* and their phenotypic analysis. *Yeast* 19, 587–599

13 Albert, R. *et al.* (1999) Diameter of the World-Wide Web. *Nature* 401, 130–131

14 Albert, R. and Barabasi, A.L. (2002) Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* 74, 47–97

15 Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature* 393, 440–442

16 Yu, H. *et al.* (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* 32, 328–337

17 Barabasi, A.L. and Albert, R. (1999) Emergence of Scaling in Random Networks. *Science* 286, 509–512

18 Albert, R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature* 406, 378–382

19 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

20 Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752

21 Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453

22 Yu, H. *et al.* (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* 19, 422–427

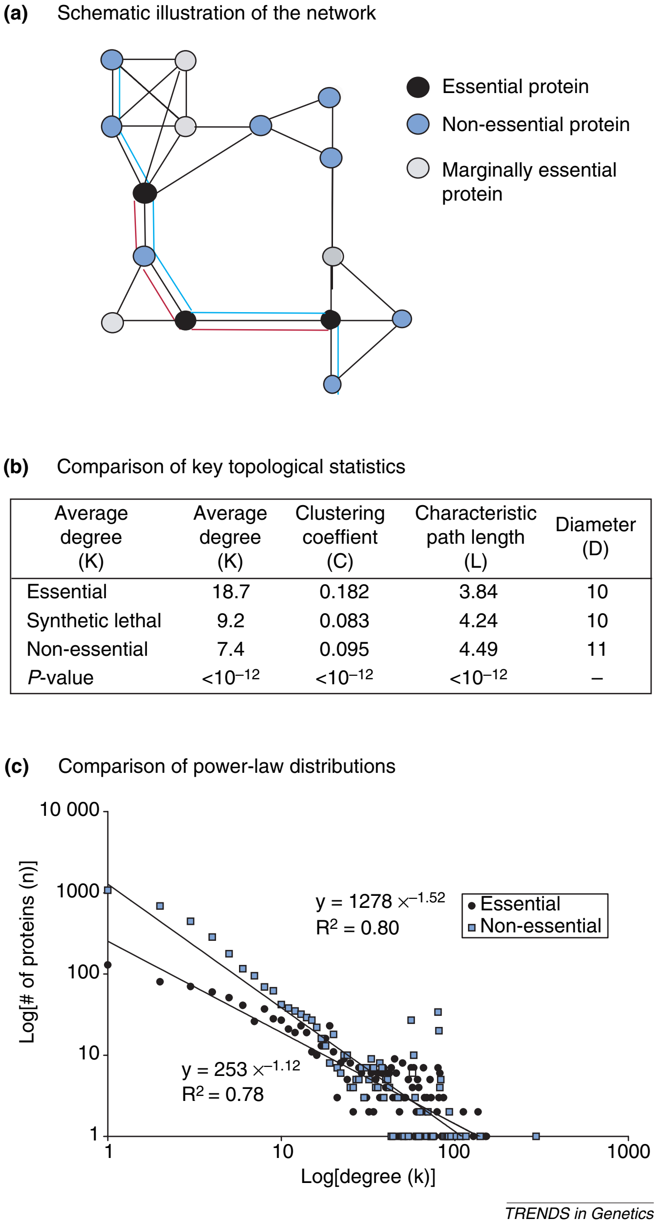
23 Horak, C.E. *et al.* (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae. Genes Dev.* 16, 3017–3033

24 Guelzim, N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 31, 60–63

25 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* 298, 799–804

26 Wingender, E. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* 29, 281–283

27 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34



Average degree (K)	Average degree (K)	Clustering coeffient (C)	Characteristic path length (L)	Diameter (D)
Essential	18.7	0.182	3.84	10
Synthetic lethal	9.2	0.083	4.24	10
Non-essential	7.4	0.095	4.49	11
P-value	$<10^{-12}$	$<10^{-12}$	$<10^{-12}$	–

Figure 1. (a) Schematic illustration of the diameter of a sub-network. In an undirected network, the diameter of the essential protein network (shown by the red line) is the maximum distance between any two essential proteins. The path can go through non-essential proteins but has to start and end at essential proteins; the same conditions apply to the non-essential protein network. Non-essential genes represent those that have no detected effects on cell fitness. The traditional concept of 'non-essential genes' includes both non-essential and marginally essential genes. **(b)** A comparison of key topological characteristics. The values of

different characteristics for essential, synthetic lethal and non-essential proteins are given in the table together with the *P*-values, which measure the statistical significance of the difference between the values for essential and non-essential proteins. The values are calculated as described in the supplementary materials online. *P*-values are calculated using non-parametric Mann-Whitney U-tests. **(c)** A comparison of power-law distributions. The plot is on a log–log scale. The regression equations (y) and correlation coefficients (R) are given close to the corresponding lines in the figure.

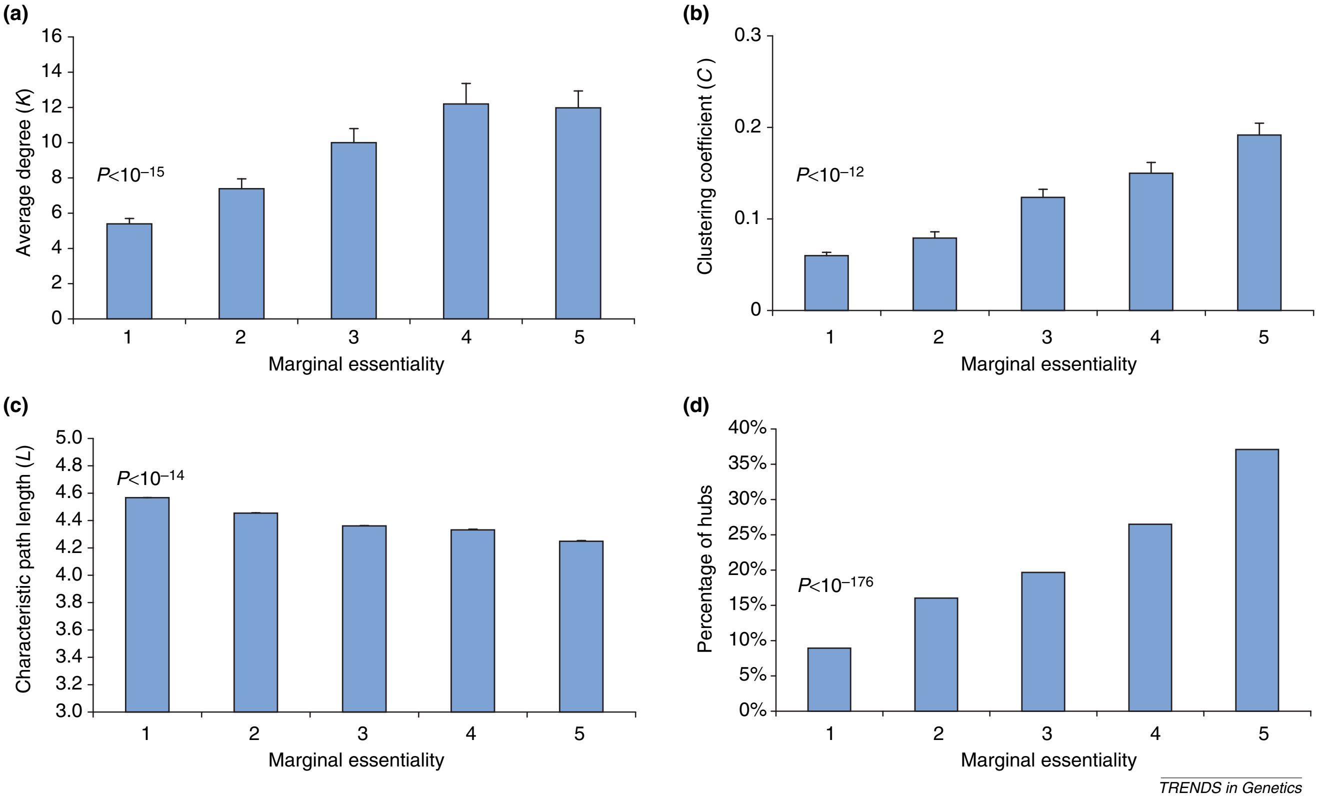


Figure 2. Monotonic relationships between topological parameters and marginal essentiality for non-essential genes. **(a)** A positive correlation exists between the average degree (*K*) and marginal essentiality (*M*). **(b)** A positive correlation exists between the clustering coefficient (*C*) and marginal essentiality. **(c)** A negative correlation exists between the characteristic path length (*D*) and marginal essentiality. **(d)** A positive correlation exists between hub percentage and marginal essentiality. The marginal essentiality for each non-essential gene is calculated by averaging the data from four datasets: (i) growth rate [5]; (ii) phenotype [2]; (iii) sporulation efficiency [6]; and (iv) small-molecule sensitivity [7]. Because the raw data in different datasets are on different scales, all the data points are normalized through dividing by the largest value in each dataset. In particular, the marginal essentiality (M_i) for gene *i* is calculated by:

$$M_i = \frac{\sum_{j \in J_i} F_{i,j} / F_{\max,j}}{J_i}$$

where $F_{i,j}$ is the value for gene *i* in dataset *j*. $F_{max,j}$ is the maximum value in dataset *j*. J_i is the number of datasets that have information on gene *i* in the four datasets. All the data included in the calculations are the raw data from the original datasets, except the growth rate data, which were baseline corrected. Before calculating the marginal essentiality, we verified that the four datasets were mutually independent. Although other methods could also be used to define marginal essentiality, we determined that different definitions have little effect (supplementary material online). Genes are grouped into five bins based on their marginal essentialities: bin one, <0.05; bin two (0.05, 0.1); bin three (0.1, 0.2); bin four (0.2, 0.3); bin five, ≥0.3. The *y*-axis represents the topological characteristics among the genes within the same bin. *P*-values show the statistical significance of the difference between the first and the last bars on each graph. The values of the topological characteristics, the marginal essentialities and the *P*-values are calculated as described in supplementary material online.

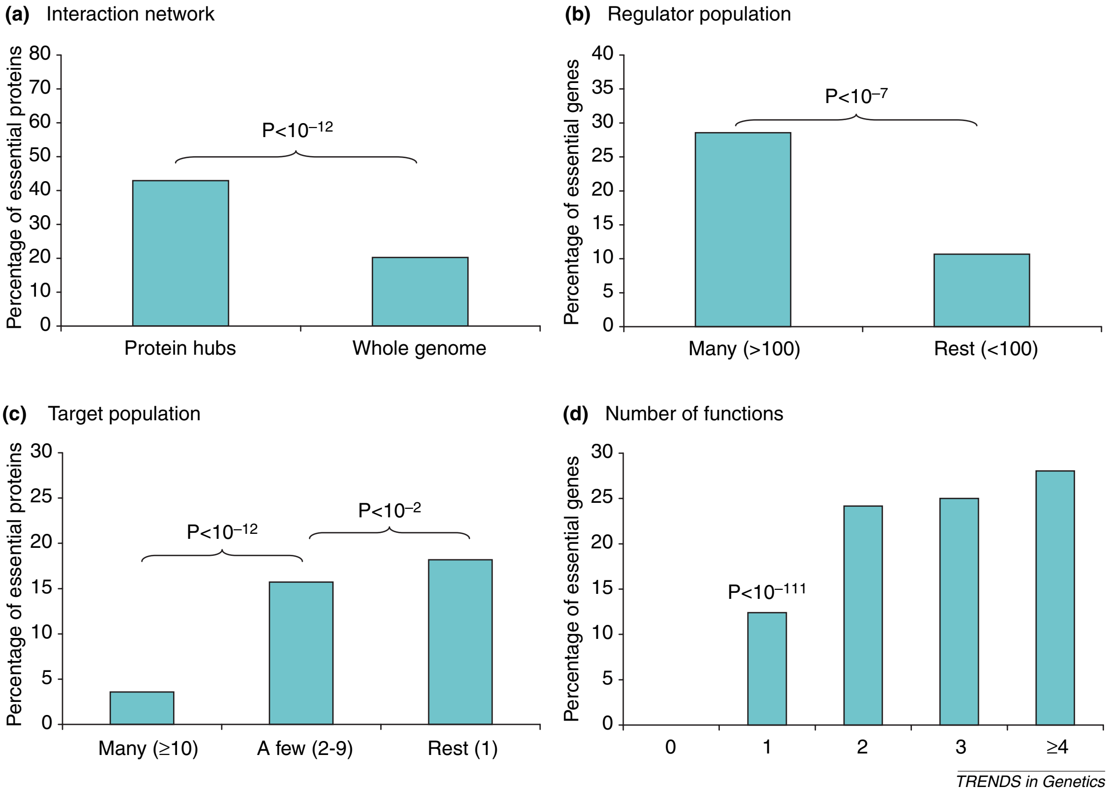


Figure 3. The observed likelihood for different classes of genes being essential. **(a)** Protein hubs in the interaction network tend to be essential. Based on the degree distribution (Figure 1 in the supplementary online), 1061 proteins were considered as protein hubs, within which the percentage of essential proteins was examined. 'Whole genome' refers to the likelihood that all proteins in the whole genome that have at least one interaction partner to be essential. There are, in total, 4743 proteins with at least one interaction partner in the dataset, among which 977 (~20%) are known to be essential. **(b)** Transcription factors (TFs) with many (>100) targets are more likely ($P<10^{-7}$) to be essential than the other proteins. **(c)** Genes with many regulating TFs (≥10) are less likely ($P<10^{-12}$) to be essential than those with only a few TFs (2–9), whereas these genes are less likely ($P<10^{-02}$) to be essential than those with only one TF. **(d)** Genes with more functions are more likely to be essential. The *P*-value measures the difference between genes with only one function and those with more than four functions. The *P*-values in all panels are calculated by the cumulative binomial distribution.