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

¶ To whom correspondence should be addressed.
Tel: +1 203 432 6105; Fax: +1 360 838 7861;
Email: Mark.Gerstein@yale.edu

haiyuan.yu@yale.edu
dov.greenbaum@yale.edu
haoxin.lu@yale.edu
xiaowei.zhu@yale.edu
mark.gerstein@yale.edu

Abstract

We introduce the notion of 'marginal essentiality' through quantitatively combining 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 (having many interactions) and to have a shorter characteristic path length to others. We extend our network analysis to encompass transcriptional regulatory networks. While transcription factors with many targets tend to be essential, surprisingly, we find that genes regulated by many transcription factors are usually not essential. Further information is available from http://bioinfo.mbb.yale.edu/network/essen.

Introduction

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.

While the definition of essentiality is not novel, Thatcher et al introduced the "marginal benefit" hypothesis[4]. It states that many non-essential genes make significant, but small, contributions to the fitness of the cell, but the effects may not be large enough to be detected by conventional methods. Here, we systematically define "marginal essentiality" (*M*) as a quantitative measure of a non-essential gene's importance to a cell (see figure 2 caption). Our measure incorporates the results from a diverse set of four large-scale knock-out experiments examining different aspects of a protein's impact on cell fitness. These four experiments measure the effect of a particular knock-out 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 knock-out analyses for yeast. (There are a number of other smaller datasets[8-12], which have data only on a small fraction of the genome and are therefore not suitable here.)

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)[13-17]. It has been shown that some protein networks follow power-law distributions[18,19]; i.e., they consist of many interconnecting nodes, a few of which have uncharacteristically high degrees (hubs). Additionally, power-law distributions can also be characterized as scale-free; i.e., 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, as many non-hub (i.e. leaf) genes can be disabled without greatly affecting the viability of the cell[19].

Recently, Jeong et al focused on the relationship between hubs and essential genes, determining that hubs tend to be essential[20]. Fraser et al also observed that an individual protein's effect on cell fitness correlates with its number of interaction partners[21]. Here, we extend the previous work to marginal essentiality and perform a genome-wide analysis of essentiality within a wide variety of protein networks.

Results

Comparison between essential and non-essential proteins within interaction network

We constructed a comprehensive and reliable yeast interaction network containing 23,294 unique interactions among 4,743 proteins[16,17,22]. In a gross comparison we found

that essential proteins, as a whole, have significantly more "links" than non-essential ones, validating earlier findings[20]. Specifically, essential proteins have approximately twice as many links as compared to non-essential ones (figure 1b). We can also see from the power-law plots of the interactions of just essential and non-essential genes (fig. 1c): 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 ones, we determined the fraction of hubs that are essential. Here, we define hubs as the top quartile of proteins with respect to number of interactions[17], giving 1061 proteins as hubs within the yeast network. We found ~43% of hubs in yeast are essential (figure 3a), 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 more closely connected to each other (as determined from the characteristic path length and diameter). Not surprisingly, the values of these topological statistics (except the clustering coefficient) for synthetic lethal genes are between those for the essential and non-essential ones (see figure 1b).

Topological characteristics for marginal essentiality within 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 network statistics (figure 2 and supplementary figure 2). 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 Fraser et al's conclusion[21]. More importantly, the greater a protein's marginal essentiality, 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.

Figure 2d shows that marginal essentially is correlated with a higher likelihood to be one of the 1061 hubs. Because hubs in the protein interaction networks have been shown to be important to cell fitness[20], 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, regulatory networks, i.e. transcription factors and the target genes they regulate[23-27]. These networks differ from protein-protein interaction networks in that they are directed. We looked at regulatory networks from two separate perspectives: the regulator population (e.g., out degree) -where we are looking at a directed network of transcription factors acting on targets, and the target population (e.g., in degree) - where we analyze the sets of target genes that are regulated by any given transcription factors.

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

The results for the regulators and the targets, though 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 would be. Our analysis of the regulator population has, logically, shown that the promiscuous regulators do tend to be essential.

On the other hand, 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). Therefore, the expression of essential genes tends to have less regulation, whereas, non-essential genes often use more regulators to control when and how much the gene products should be expressed. Another possible reason is that essential proteins carry out the most basic and important functions within the cell. They, consequently, always need to be "on". And their expression does not need to be regulated by many factors since this makes the essential gene dependent on the viability of more regulators, which makes the cell less stable.

Relationship between essentiality and function

Having discussed thoroughly that the essentiality of a gene is directly related to its importance to the cell fitness in both interaction and regulatory networks, we now examine the relationship between the number of a gene’s functions and its tendency to be essential, using the MIPS functional classification[28]. 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 paper, we comprehensively defined "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.

Figure Captions

Figure 1. A. Schematic illustration of the diameter of a sub-network. In an undirected network, the diameter of the essential protein network (shown as 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 ones, which is the same for that of the non-essential protein network. In this panel, 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 in this panel. B. 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, measuring the statistical significance of the difference between the values for essential and non-essential proteins. The values are calculated as described in supplementary materials. P-values are calculated using non-parametric Mann-Whitney U-tests. C. Comparison of power-law distributions. The plot is on a log-log scale. The regression equations and correlation coefficients (R) are given close to the corresponding lines in the figure. Open squares: non-essential genes, solid circles: essential genes.

Figure 2. Monotonic relationships between topological parameters and marginal essentiality for non-essential genes. A. positive correlation between average degree and marginal essentiality. B. positive correlation between clustering coefficient and marginal essentiality. C. negative correlation between characteristic path length and marginal essentiality. D. positive correlation between hub percentage and marginal essentiality. The x-axis is the marginal essentiality (M). 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[17]. 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[17]. Genes are grouped into 5 bins based on their marginal essentialities: bin1, <0.05; bin2 [0.05, 0.1); bin3 [0.1, 0.2); bin4 [0.2, 0.3); bin5, ≥0.3. The y-axis is 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 materials.

Figure 3. 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 (see supplementary figure 1), 1061 proteins are considered as protein hubs, within which the percentage of essential proteins is examined. "Whole genome": likelihood of all proteins in the whole genome that have at least one interaction partner to be essential. There are, in total, 4,743 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 rest. 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), while 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 4 functions. The P values in all panels are all calculated by the cumulative binomial distribution.

Reference:

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, 8927-32

2 Ross-Macdonald, P. et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 402, 413-8

3 Winzeler, E.A. et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* 285, 901-6

4 Thatcher, J.W., Shaw, J.M. and Dickinson, W.J. (1998) Marginal fitness contributions of nonessential genes in yeast. *Proc Natl Acad Sci U S A* 95, 253-7

5 Steinmetz, L.M. et al. (2002) Systematic screen for human disease genes in yeast. *Nat Genet* 31, 400-4

6 Deutschbauer, A.M., Williams, R.M., Chu, A.M. and Davis, R.W. (2002) Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomycescerevisiae. *Proc Natl Acad Sci U S A* 99, 15530-5

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, 3345-50

8 Smith, V. et al. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* 274, 2069-74

9 Rieger, K.J. et al. (1999) Chemotyping of yeast mutants using robotics. *Yeast* 15, 973-86

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, 477-83

12 Sakumoto, N. et al. (2002) A series of double disruptants for protein phosphatase genes in Saccharomyces cerevisiae and their phenotypic analysis. *Yeast* 19, 587-99

13 Albert, R., Jeong, H. and Barabasi, A.L. (1999) Diameter of the World-Wide Web. *Nature* 401, 130-131

14 Albert, R. and Barabasi, A.L. (2002) Statistical Mechanics of Complex Networks. *Review of Modern Physics* 74, 47-97

15 Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature* 393, 440-2

16 Yu, H. et al. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* 32, 328-37

17 See supplementary materials.

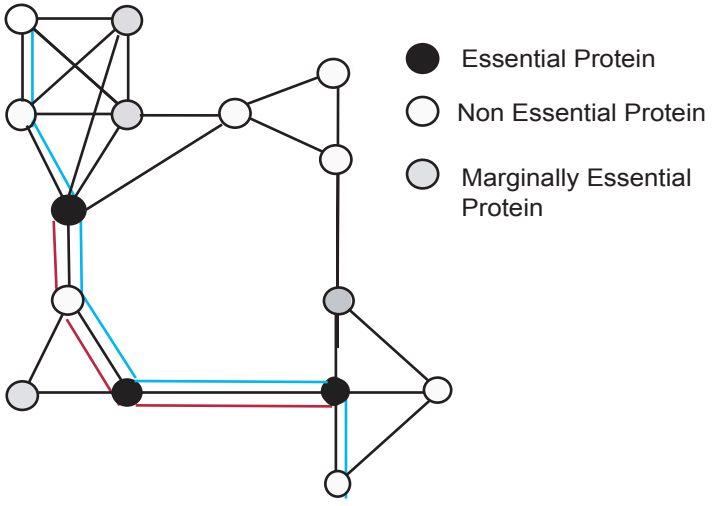
18 Barabasi, A.L. and Albert, R. (1999) Emergence of Scaling in Random Networks. *Science* 286, 509-512

19 Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. *Nature* 406, 378-382

20 Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* 411, 41-2

21 Fraser, H.B. et al. (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750-2
22 Jansen, R. et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-53
23 Yu, H., Luscombe, N.M., Qian, J. and Gerstein, M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 19, 422-7
24 Horak, C.E. et al. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* 16, 3017-3033
25 Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31, 60-3
26 Lee, T.I. et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298, 799-804
27 Wingender, E. et al. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 29, 281-3
28 Mewes, H.W. et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 30, 31-4

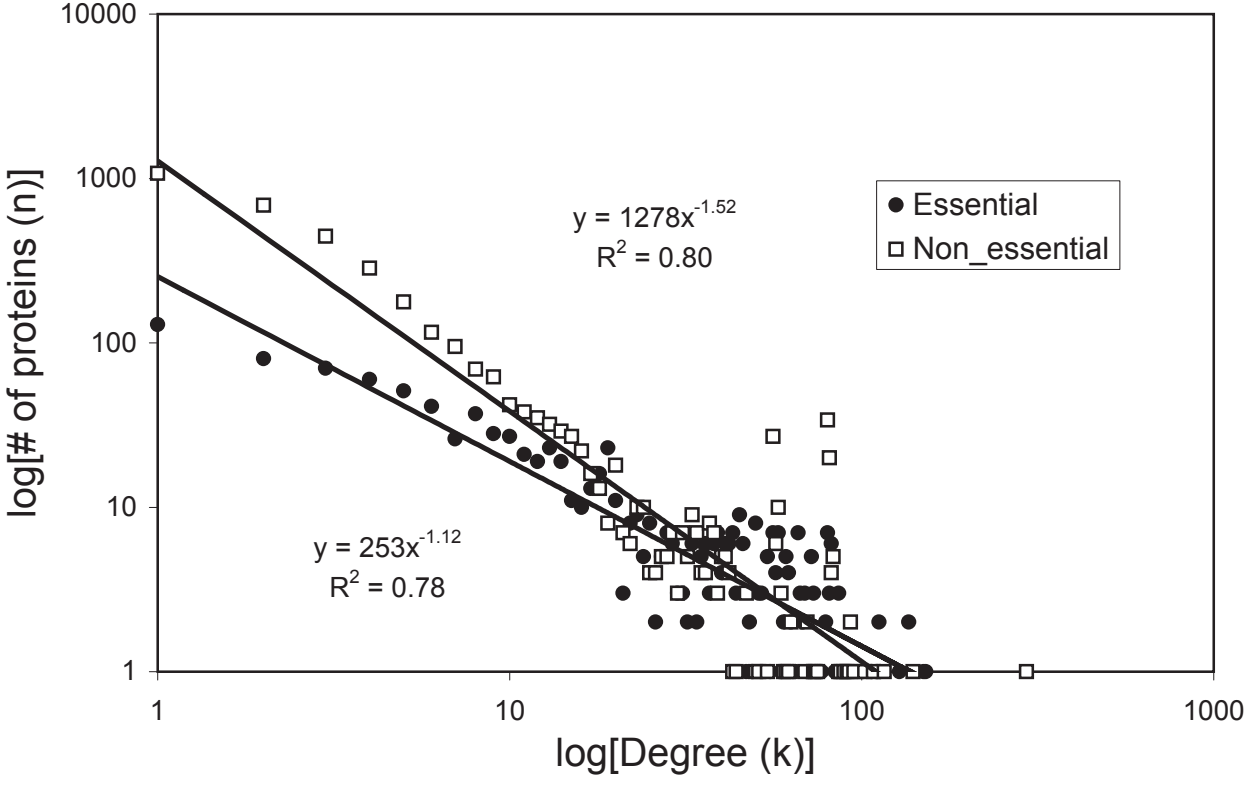
A. Schematic illustration of the network

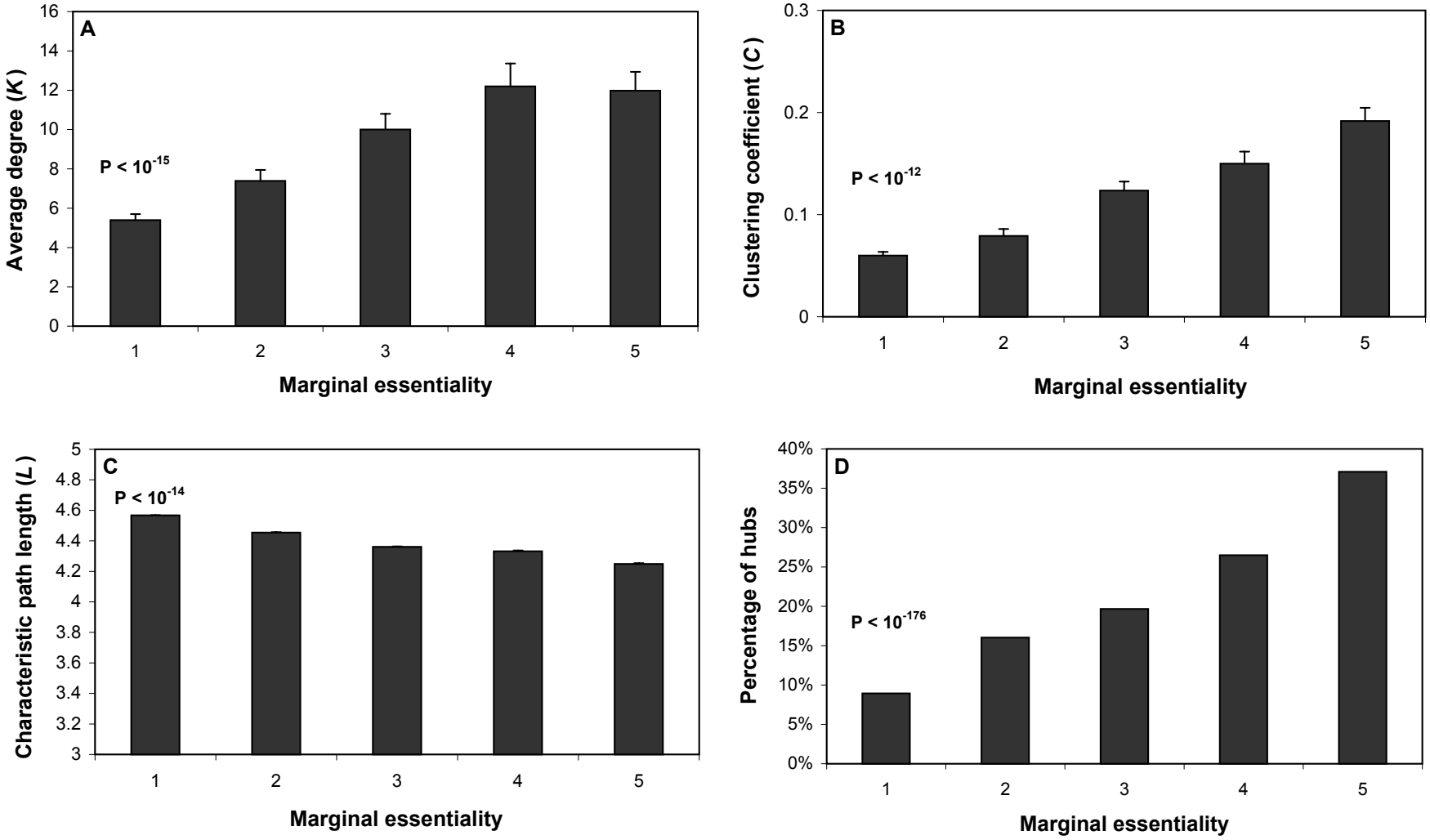


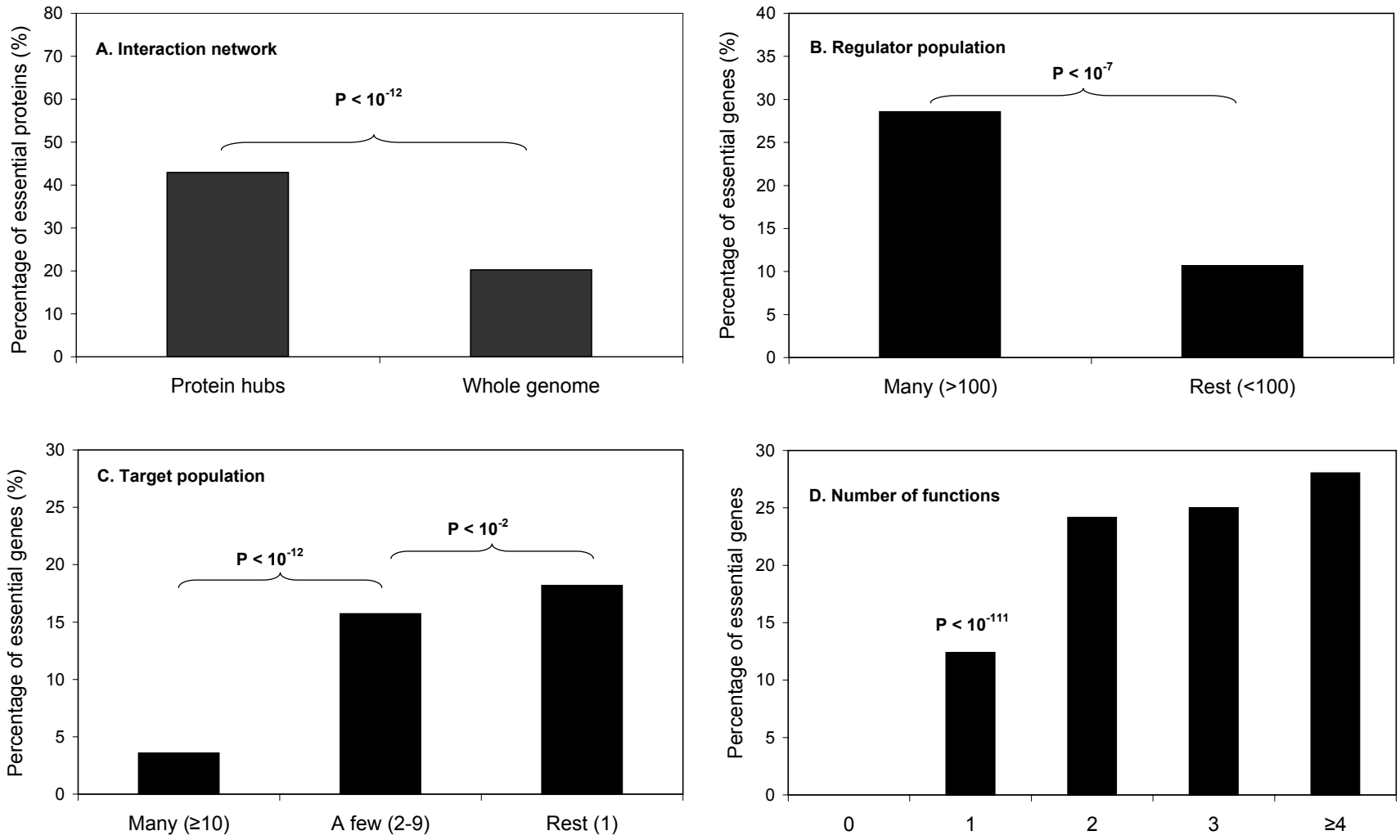
B. Comparison of key topological characteristics

	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	$<10^{-12}$	$<10^{-12}$	$<10^{-12}$	–

C. Comparison of power-law distributions




A
Average degree (K)
P < 10-15
Marginal essentiality
B
Clustering coefficient (C)
P < 10-12
Marginal essentiality
C
Characteristic path length (L)
P < 10-14
Marginal essentiality
D
Percentage of hubs
P < 10-176
Marginal essentiality


A. Interaction network
Percentage of essential proteins (%)
P < 10-12
Protein hubs
Whole genome
B. Regulator population
Percentage of essential genes (%)
P < 10-7
Many (>100)
Rest (<100)
C. Target population
Percentage of essential genes (%)
P < 10-12
P < 10-2
Many (≥10)
A few (2-9)
Rest (1)
D. Number of functions
Percentage of essential genes
P < 10-111
0
1
2
3
≥4