Supplementary materials

Calculation of the growth rate for each gene

In the growth rate dataset, each gene has many different growth rates under different conditions. The average growth rate for gene *i* deletion strain (G_i) is calculated by the formula:

$$G_i = \frac{\sum_{n \in N_i} \left| R_{i,n} - 1 \right|}{N_i}$$

where $R_{i,n}$ is the raw growth rate data of gene *i* on medium *n* in the original dataset. N_i is the total number of mediums used in experiments for gene *i*.

Independence of different large-scale datasets

In order to combine the results of the four large-scale datasets, we have to make sure that the four datasets are mutually independent with each other. We plotted the values for each gene in different datasets as scatter plots and did not observe any correlation. Furthermore, we calculated the correlation coefficient between any two datasets (6 pairs in total), none of which is significant (please refer to supplementary table 2).

Different definitions of marginal essentiality

In the main text, we discussed the definition of "marginal essentiality" as the average of the normalized the values in the four datasets. Here, we introduce four new methods to define "marginal essentiality":

1. Marginal essentiality as the maximum value among the four datasets

The marginal essentiality (M_i) for gene *i* is calculated by the formula:

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

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 gene *i* have been tested in the four datasets. All the calculations in figure 2 were repeated using this new definition of "marginal essentiality". Supplementary figure 4 shows that all results remain the same. Specifically, there is a positive relationship in panels A, B, and D, while there is a negative relationship in panel C.

2. Marginal essentiality as the minimum value among the four datasets

The marginal essentiality (M_i) for gene *i* is calculated by the formula:

$$M_i = \min\{F_{i,j} / F_{max,j} \mid j \in J_i\}$$

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 gene i have been tested in the four datasets. All the calculations in figure 2 were repeated using this new definition of "marginal essentiality". Supplementary figure 5 shows that all results remain the same.

3. Marginal essentiality is normalized as the percentile rank in each dataset

In the previous three methods, the data in each dataset are all normalized through dividing by the largest value in the dataset. We could also normalize the data by taking their corresponding percentile ranks in the whole set. Thus, the marginal essentiality (M_i) for gene i is calculated by the formula:

$$M_i = \frac{\sum_{j \in J_i} P_{i,j}}{J_i}$$

where $P_{i,j}$ is the percentile rank for gene i in dataset j. J_i is the number of datasets that gene i have been tested in the four datasets. All the calculations in figure 2 were repeated using this new definition of "marginal essentiality". Supplementary figure 6 shows that all results remain the same.

4. Marginal essentiality is normalized as the z-score in each dataset

We could also normalize the data in each dataset in a "z-score" fashion. Thus, the marginal essentiality (M_i) for gene i is calculated by the formula:

$$M_i = \frac{\sum_{j \in J_i} Z_{i,j}}{J_i}$$

where $Z_{i,j}$ is the z-score for gene i in dataset j. J_i is the number of datasets that gene i have been tested in the four datasets. All the calculations in figure 2 were repeated using this new definition of "marginal essentiality". Supplementary figure 7 shows that all results remain the same.

Network Definitions

1. Topological characteristics

Network parameters allow for a simple yet powerful analysis of a global protein interaction network; every network has specific defining and descriptive characteristics. We chose to look at four characteristics for both the essential and non-essential genes in the network of interacting proteins[1-3] (see figure 1a):

(i) The average degree (K) of a network relates to the average number of connections between a node and all other nodes in the network.

(ii) The clustering coefficient (C) defines the cliquishness of each node, which is an odds ratio of observed over expected interactions between a protein node's first neighbors, which is calculated by the formula:

$$\frac{\sum_{i \in N} \frac{2e_i}{k_i(k_i-1)}}{N}$$

where i: ith node in the network. k_i is the degree of node i. e_i is the number of edges existing between the k_i nearest neighbors of node i. N is the total number of nodes in the network.

(iii) The characteristic path-length (L) describes, on average, the number of intermediate nodes between two nodes of the same class, (in this case essential and non essential), which is calculated by the formula:

$$\frac{2 \times \sum_{i,j \in N} d_{ij}}{N(N-1)}$$

where: d_{ij} is the distance between node i and node j. N is the total number of nodes in the network. $N(N\text{-}1)/2$ is the highest possible number of node pairs. Disconnected node pairs are excluded.

(iv) The diameter (D) of the network describes the greatest distance among the shortest path lengths between any two nodes of the same class. The shortest distance between all pairs of nodes is determined and the largest 'shortest distance' is the diameter. (By definition, "diameter" does not carry too much statistical weight[1,2].)

2. Hub

Usually, hubs are defined qualitatively as the most highly connected nodes. In this paper, we defined hubs based on the distribution of the number of proteins with certain degrees - the degree distribution. Given the continuous distribution of degrees for all nodes, it is difficult to provide an exact cut-off point where we can say that a node with greater than a specific number of degrees is a hub. As discussed in figure 3 legend, we define the cut-off as the point, where the degree distribution begins to straighten out.

3. Directed networks, in degree and out degree

Regulatory networks are directed networks: the edges of the network have a defined direction. For example in a regulator network, regulators regulate their targets, not the other way around. A node in the directed network may have an in-degree and an out-degree (see figure 1c), which are completely independent. For directed networks, it is impossible to determine clustering coefficients[2]. Therefore, we focus on the analysis on the average degree.

Construction of the yeast interaction network

Using the same methodology as previous analyses, we constructed a large interconnecting network of most proteins in the yeast genome, drawing from a large body of yeast protein-protein interactions determined through a variety of high-throughput experiments, most notably two yeast two-hybrid datasets[4,5] and two *in vivo* pull-down datasets[6,7]. However, large-scale interaction datasets are known to be error prone[8,9]. In order to introduce a confidence limit, Jansen et al calculated a likelihood ratio (L) for each pair of proteins within the four datasets[9]. Simply put, the higher the likelihood ratio the more likely the interaction is true. In their paper, $L \geq 300$ was used as an appropriate cutoff for choosing reliable interactions.

Many databases such as MIPS[10], BIND[11], and DIP[12] also record the interactions from small-scale experiments, together with the results of the high-throughput methods, these databases were also included in the makeup of the interaction network. These small-scale interaction datasets are generally believed to be the most-reliable datasets[8,9,13,14].

Therefore, we constructed a comprehensive and reliable yeast interaction networks by taking the union of the three small-scale datasets and the interacting pairs within the four large-scale datasets with $L \geq 300$. The network consists of 23,294 unique interactions among 4,743 proteins.

P values calculated by the cumulative binomial distribution

P values for the difference between percentages of essential genes within different datasets are calculated using the cumulative binomial distribution:

$$\mathrm{P}(c \geq c_o) = \sum_{c=c_o}^{N} [\frac{N!}{N!(N-c)!}] p^c (1-p)^{N-c}$$

N is the total number of genes in certain dataset, c_o is the number of observed essential genes within this dataset, and p is the probability of finding an essential gene within the other dataset (*i.e.* percentage of essential genes). P values for the difference between percentages of hubs within different groups of proteins with different marginal essentiality are calculated in a similar fashion.

Further discussion

1. Comparison between essential and non-essential genes

We ranked the essential and non-essential proteins based on their degrees. Interestingly, we found that the bottom 10% of the essential proteins, on average, have the same number of links (~ 1.0) as that of the non-essential proteins. However, the top 10% of the essential proteins have twice as many links as that of the non-essential proteins. These 'super hubs' contribute the most to the difference between the average degrees of the two groups as discussed in the main text.

2. Hubs in the network

Generally, scale-free networks can lose a large number of nodes and still maintain their connectivity. Alternatively, when even a few hubs are knocked out, the network tends to fall apart. Similarly, essential genes are important for the survival of the cell: knock out a couple of non-essential genes and the cell may continue to live, albeit probably not as healthfully as a normal cell; knock out an essential gene and the cell dies.

How are hubs created? One theory proposes that older nodes in a network are more likely to be hubs[15,16]. As a network grows, these older nodes have had more time to acquire additional links than the younger nodes. We could make the assumption then that essential genes are evolutionarily older than their non-essential counterparts. It was recently established that essential genes are more conserved than their non-essential counterparts[17], as well as proteins with more interaction partners[18].

Age alone may not make a hub. It is also thought that there is some sort of preferential growth of hubs such that the nodes that are rich with interactions get richer. If this is the case, how do newer nodes also become hubs? The theory is that the fitness of a node also plays an important part in its selection to become a hub. Given that interactions with other proteins are important for that protein's maintenance (proteins that do not interact would seem to serve no purpose), it is logical to say that those proteins that do eventually become hubs are those that have, for the most part, made themselves integral to the cell, i.e. essential proteins.

Another theory has it that hubs are more likely to be shared across genomes whereas normal nodes tend to be more species specific. Indeed, it has been postulated that non-essential genes are integral for evolutionary diversification between strains of bacteria[19].

Although hubs have been generally believed to be the most essential part of the networks[18,20-23], our analysis has clearly shown that essential genes are underrepresented among the hubs in the target population of the regulatory network (targets with lots of regulators). Therefore, the meaning of the network, especially that of the edges (e.g. interaction, regulation, etc.), should be carefully considered before making blanket statements about essentiality and connectivity.

Molecular networks are useful tools in understanding cellular biology. While we focused on the broad categories of essentiality and non essentiality in analyzing the protein interaction networks, future more comprehensive interaction and annotation data will allow for additional network analyses involving possibly functional or expression information. For example, one could investigate whether functionally similar genes interact closely with other similarly functional genes. While presently annotation is sometimes transferred from a known interacting gene to its unannotated partner, this may not make sense for genes whose functional cousins do not preferentially interact with each other. Additionally, it would be interesting to determine the cliquishness of interacting partners with similar expression values; given that mRNA expression does not

correlate well with interactions[24], how many interacting sets of genes actually have similar expression patterns.

In order to correctly measure the marginal essentiality of the non-essential genes, more phenotypic analyses need to be performed and the results have to be incorporated. A more systematic method to calculate the marginal essentiality with the consideration of assigning weights to different experiments would be quite useful in this context.

3. Relationship between number of paralogs and essentiality

Essential genes are those that are very important to cell fitness. When an essential gene is deleted, the cell can not survive. Therefore, a gene with many paralogs in the genome can not be essential, even if it is extremely important to the cell fitness. Because, when this gene is deleted, its paralogs can perform its function instead, the cell should be able to survive. We, thus, performed an analysis on the relationship between the number of a gene's paralogs and its essentiality and found that genes without paralogs are indeed much more likely to be essential. However, supplementary figure 15 also shows that genes with less paralogs are not more likely to be essential.

Supplementary figure captions

Supplementary Figure 1. Determination of the cut-off for protein hubs. Given the continuous distribution of degrees for all nodes, it is difficult to provide an exact cut-off point where a node with a specific number of degrees or greater can be called a hub. Here, the cutoff is chosen at the point, where the distribution begins to straighten out (≥10) and the number of the defined hubs (1061) is comparable to the number of essential proteins (977). Therefore, hubs are roughly the top 25% of the proteins with the highest degrees.

Supplementary Figure 2. Correlation between a gene's degree and its likelihood of being essential.

Supplementary Figure 3. Negative correlation between diameter and marginal essentiality. The correlation is not as clear as the other three topological parameters, because diameter is the maximum distance between the nodes and therefore it does not carry too much statistical weight as characteristic path length though they reflects almost the same topological properties, which has been discussed in the text.

Supplementary Figures 4-7. 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 marginal essentiality for a gene is calculated as maximum value, minimum value, percentile rank, and z-score, respectively. Genes are grouped into 5 bins based on their marginal essentialities: In figure 4, bin1, <0.05; bin2 [0.05, 0.1); bin3 [0.1, 0.2); bin4 [0.2, 0.5); bin5, ≥0.5. In figure 5, bin1, <0.09; bin2 [0.09, 0.16); bin3 [0.16, 0.3); bin4 [0.3, 0.6); bin5, ≥0.6. In figure 6, bin1, <0.36; bin2 [0.36, 0.46); bin3 [0.46, 0.56); bin4 [0.56, 0.66); bin5, ≥0.66. In figure 7, bin1, <0.5; bin2 [0.5, 0.8); bin3 [0.8, 1); bin4 [1, 1.4); bin5, ≥1.4.

Supplementary Figure 8. A. Percentage of essential genes increase as the percentile rank of gene's degree increases in the regulator networks (outward networks). B. Percentage of essential genes decreases as the percentile rank of gene's degree increases in the target networks (inward networks). Genes are ranked by their degrees within the corresponding sub-networks. Percentile rank reflects the relative standing of a specific degree value in the networks. The percentile ranks of the genes are binned roughly at a unit of 10%. Because many genes have the same degree (especially in the target networks), the bin of both plot is not uniform.

Supplementary Figures 9-14. Correlation between different large-scale datasets. From the plots, we find no significant correlations between any two datasets (see supplementary table 2).

Supplementary Figure 15. Relationship between the number of a gene's paralogs and its likelihood of being essential. The P value measures the difference between genes without

any paralogs and those with at least one (i.e. the combination of bars “1” and “≥2”). It is calculated by the cumulative binomial distribution.

Supplementary Tables

Supplementary Table 1. Expression level and fluctuation for essential and non-essential genes. The absolute expression level of each gene is measured by Cho et al (Mol Cell. 1998. 2:65-73).

	# of genes	Average expression level	Average expression fluctuation
Essential	1098	577.7	0.256
Non-essential	5024	396.4	0.314
P-value*	–	$<10^{-40}$	$<10^{-39}$

* P-values are calculated by Mann-Whitney U-tests.

Supplementary Table 2. Correlation coefficients between different marginal essentiality datasets

	Small-molecule sensitivity	Growth rate	Sporulation efficiency	Phenotypic microarray
Small-molecule sensitivity	—	0.2664	0.0739	0.1561
Growth rate	0.2664	—	0.1694	0.3075
Sporulation efficiency	0.0739	0.1694	—	0.1520
Phenotypic microarray	0.1561	0.3075	0.1520	—

Reference:

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

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

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

4 Ito, T. et al. (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* 97, 1143-7

5 Uetz, P. et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.[comment]. *Nature* 403, 623-7

6 Ho, Y. et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.[comment]. *Nature* 415, 180-3

7 Gavin, A.C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes.[comment]. *Nature* 415, 141-7

8 von Mering, C. et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403

9 Jansen, R. et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-53

10 Mewes, H.W. et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 30, 31-4

11 Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 31, 248-50

12 Xenarios, I. et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 30, 303-5

13 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

14 Jansen, R., Lan, N., Qian, J. and Gerstein, M. (2002) Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics* 2, 71-81

15 Barabasi, A.L. (2002) *Linked: The New Science of Networks*, Perseus Publishing

16 Rosel, N. (1983) The hub of a wheel: a neighborhood support network. *International Journal of Aging & Human Development* 16, 193-200

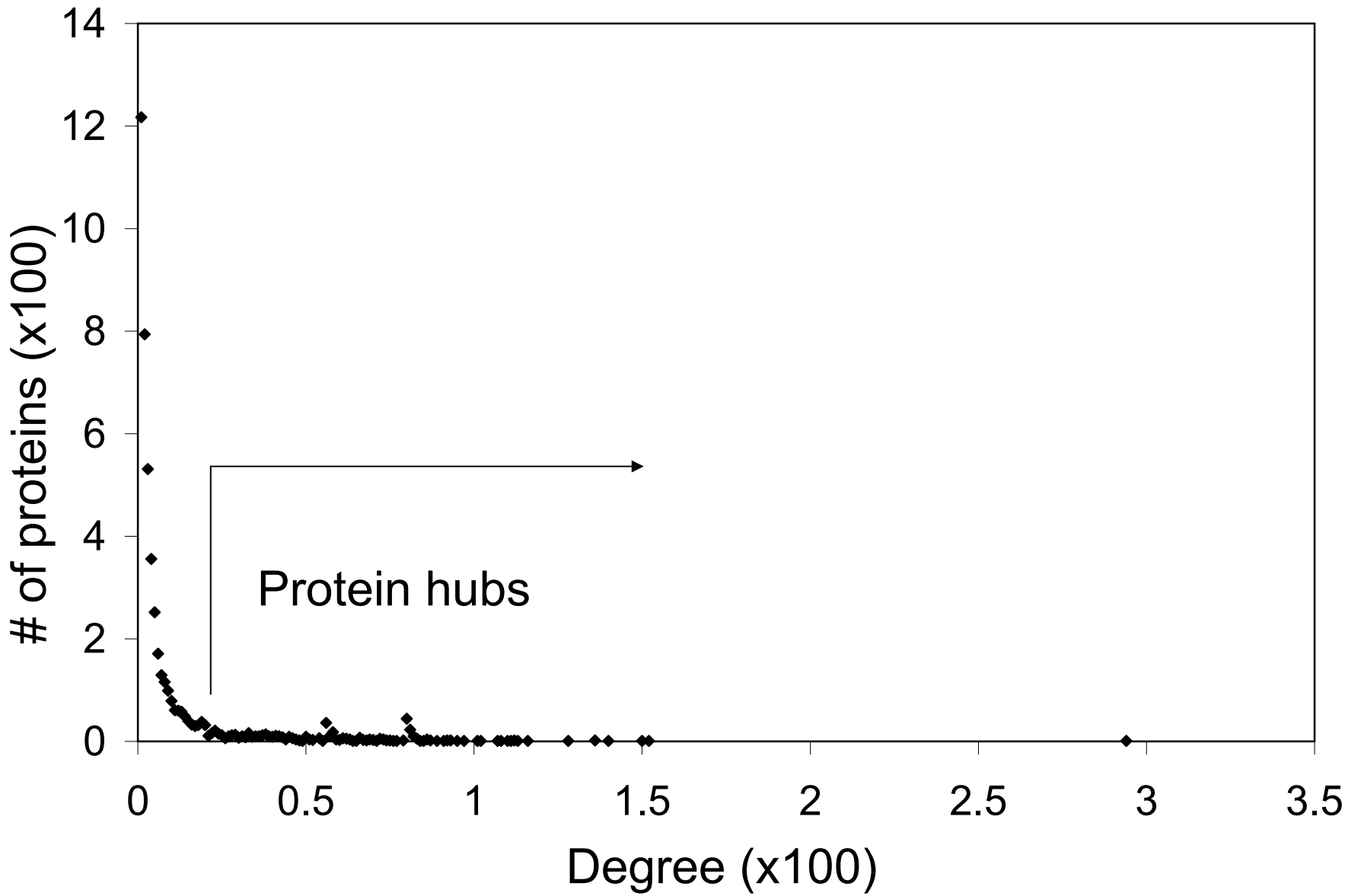
17 Jordan, I.K., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 12, 962-8

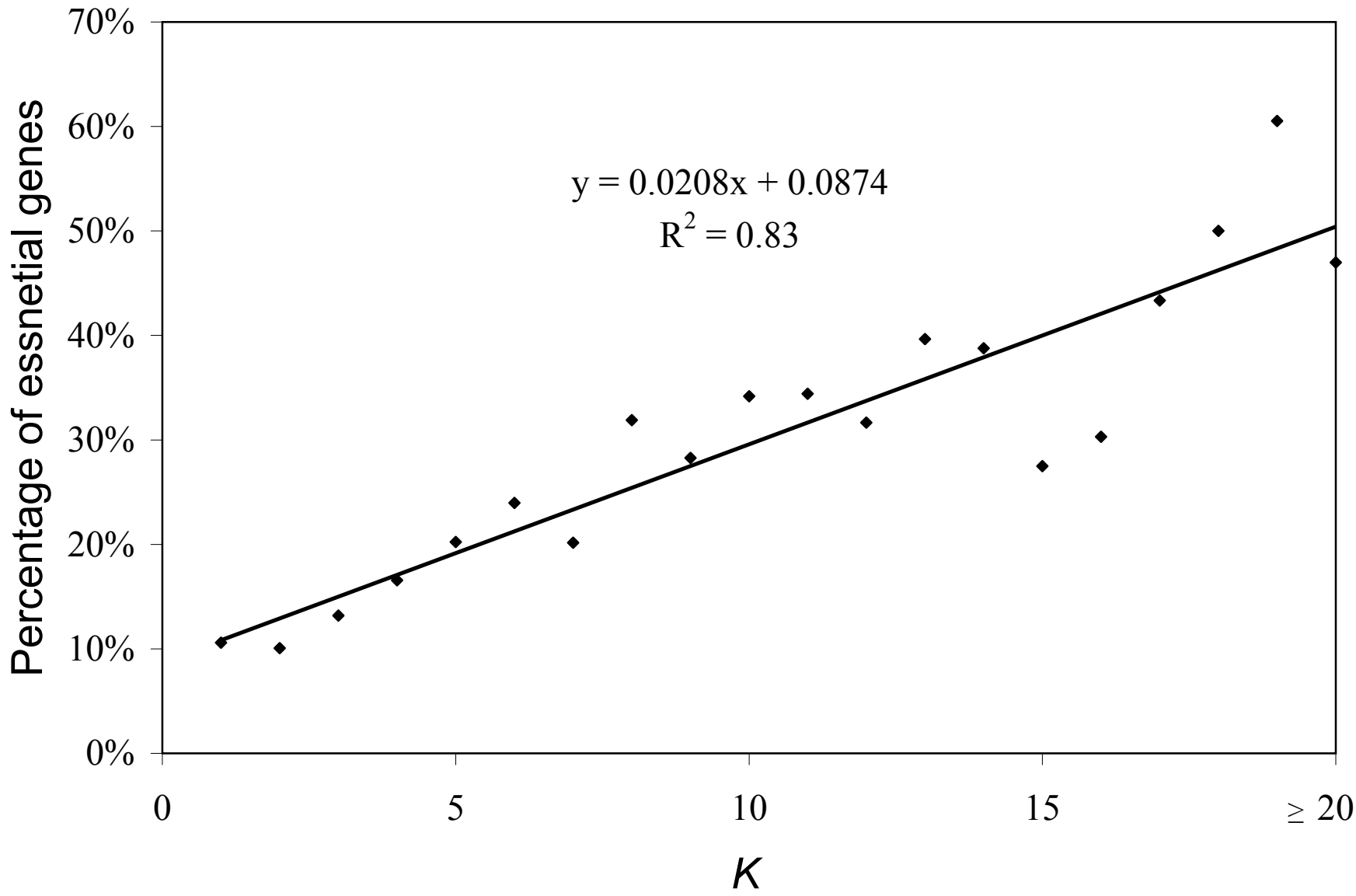
18 Fraser, H.B. et al. (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750-2

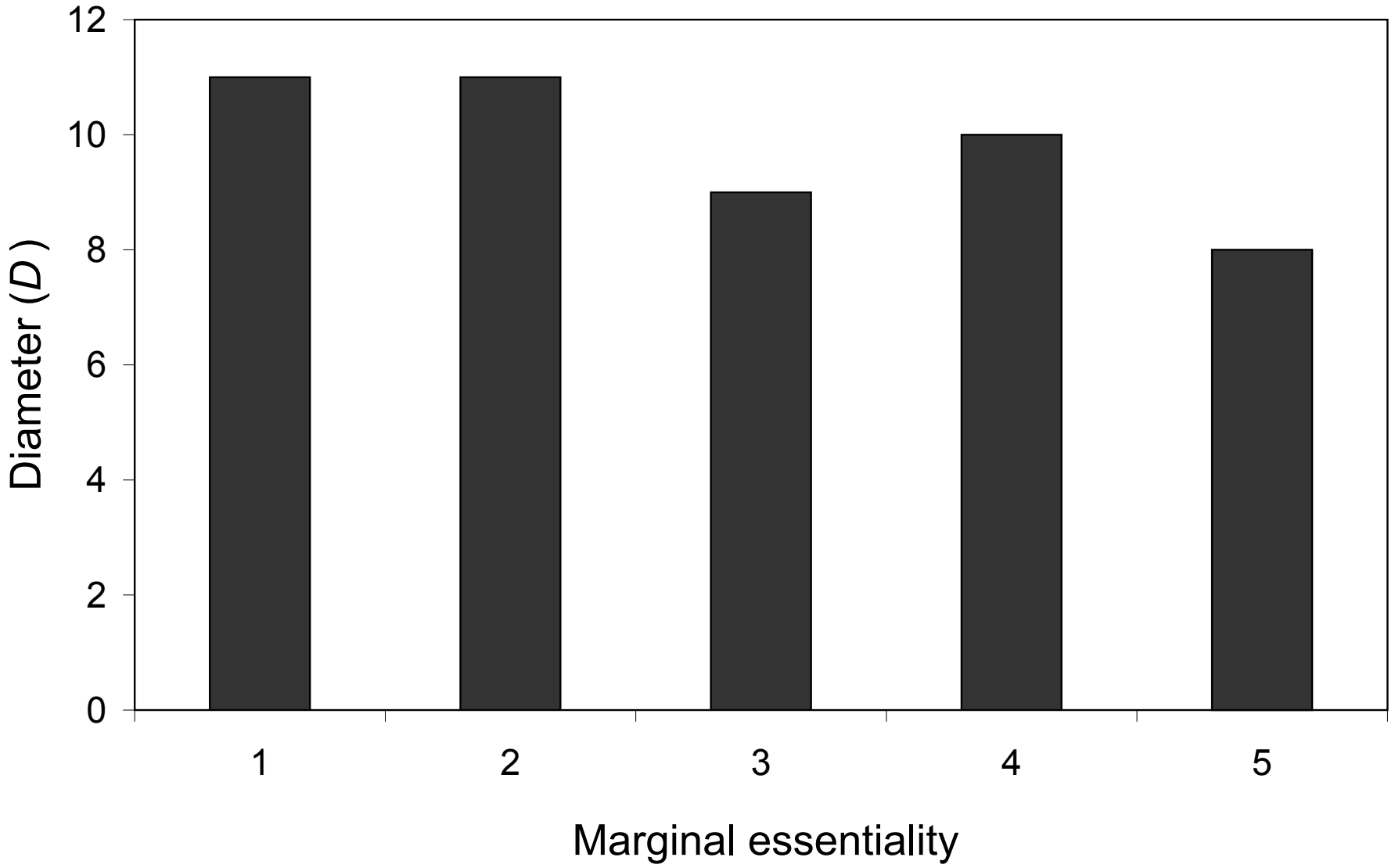
19 Jordan, I.K., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Microevolutionary genomics of bacteria. *Theor Popul Biol* 61, 435-47

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

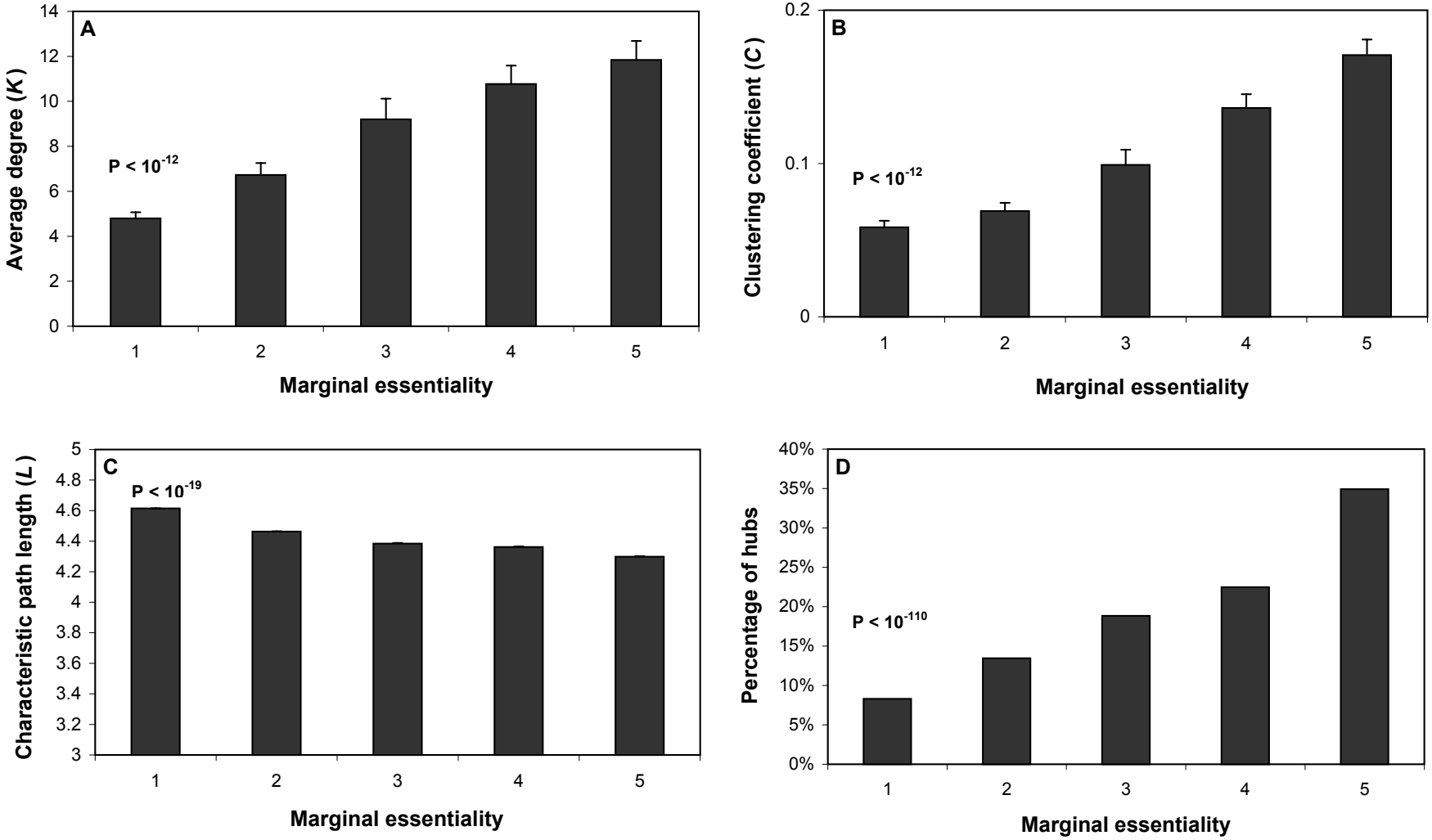
21 Barabasi, A.L. and Albert, R. (1999) Emergence of Scaling in Random Networks. Science 286, 509-512
22 Amaral, L.A., Scala, A., Barthelemy, M. and Stanley, H.E. (2000) Classes of small-world networks. Proc Natl Acad Sci U S A 97, 11149-52
23 Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature 411, 41-2
24 Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res 12, 37-46


14
12
10
8
6
4
2
0
# of proteins (x100)
Protein hubs
0
0.5
1
1.5
2
2.5
3
3.5
Degree (x100)

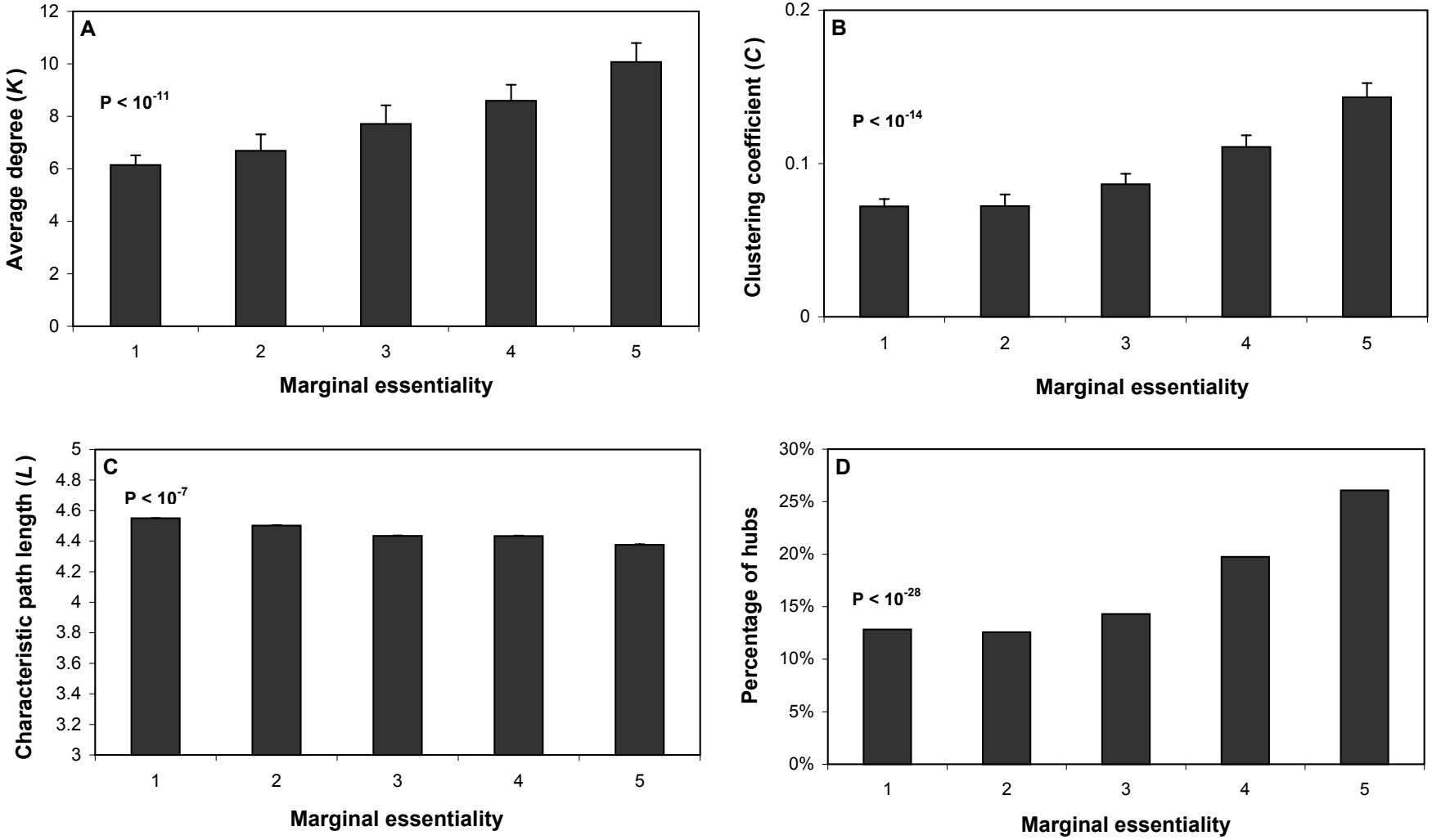

70%
60%
50%
40%
30%
20%
10%
0%
Percentage of essnetial genes
y = 0.0208x + 0.0874
R2 = 0.83
0
5
10
15
≥ 20
K

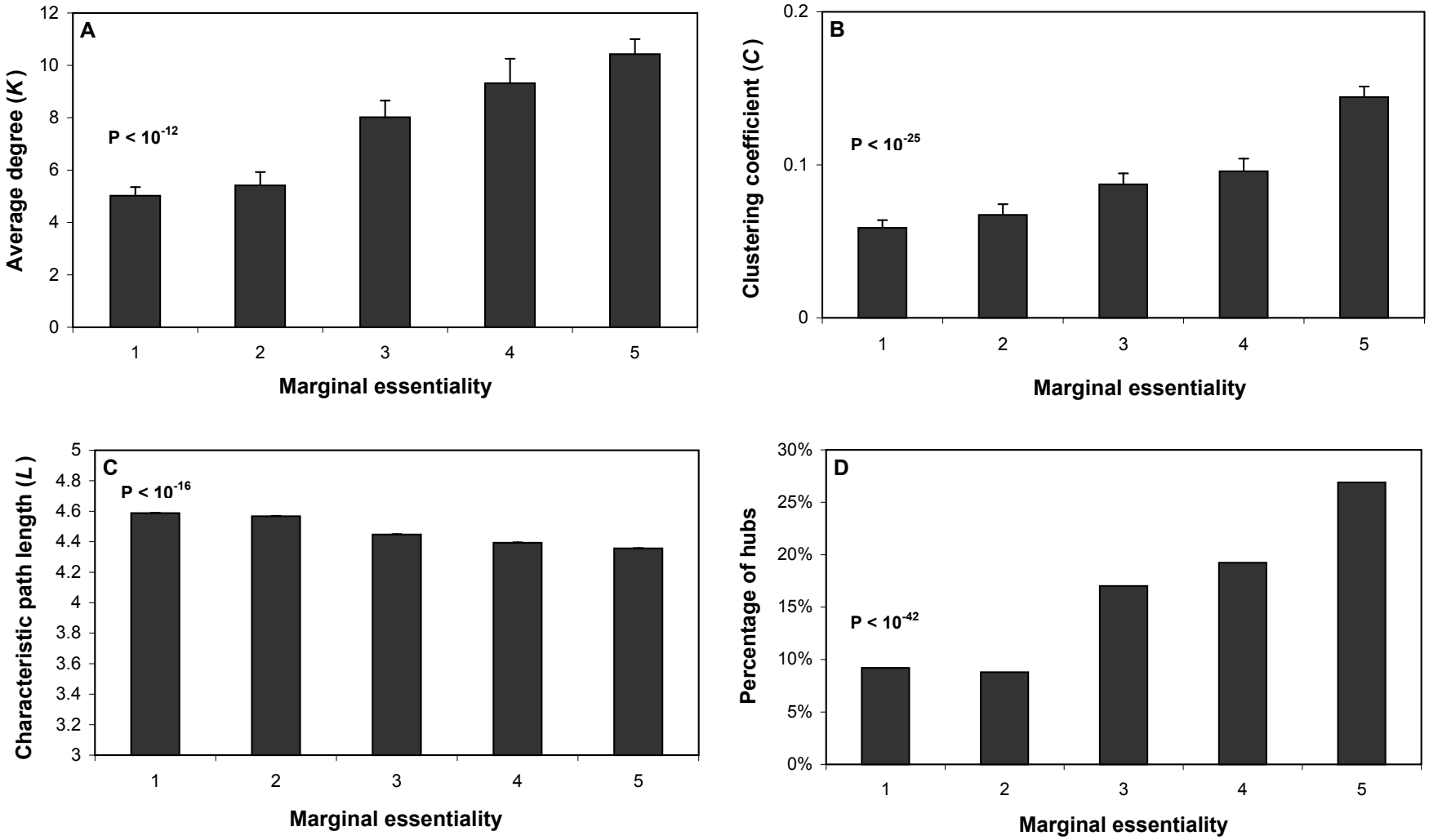

12
10
8
6
4
2
0
Diameter (D)
1
2
3
4
5
Marginal essentiality

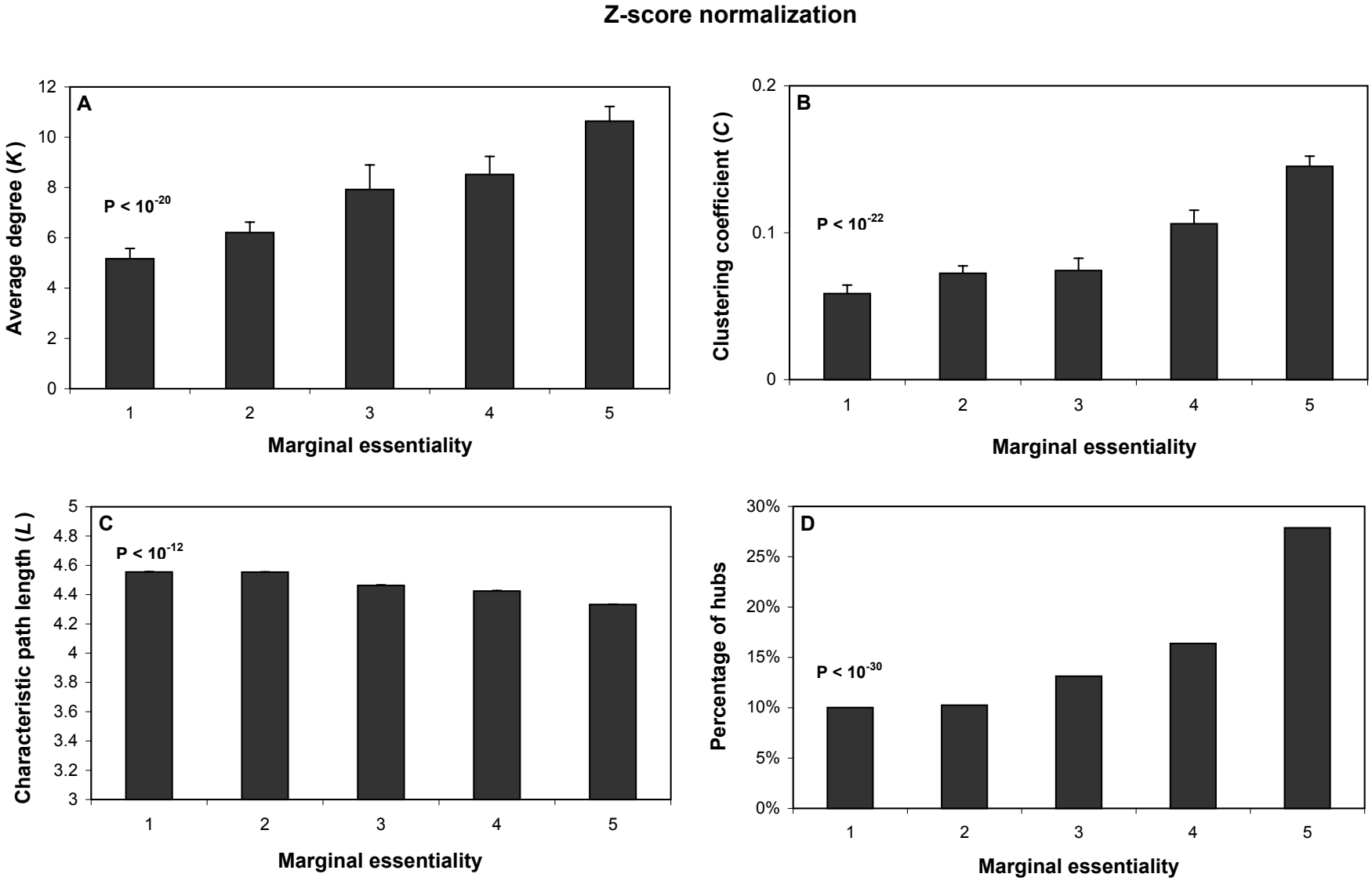
Maximum Values



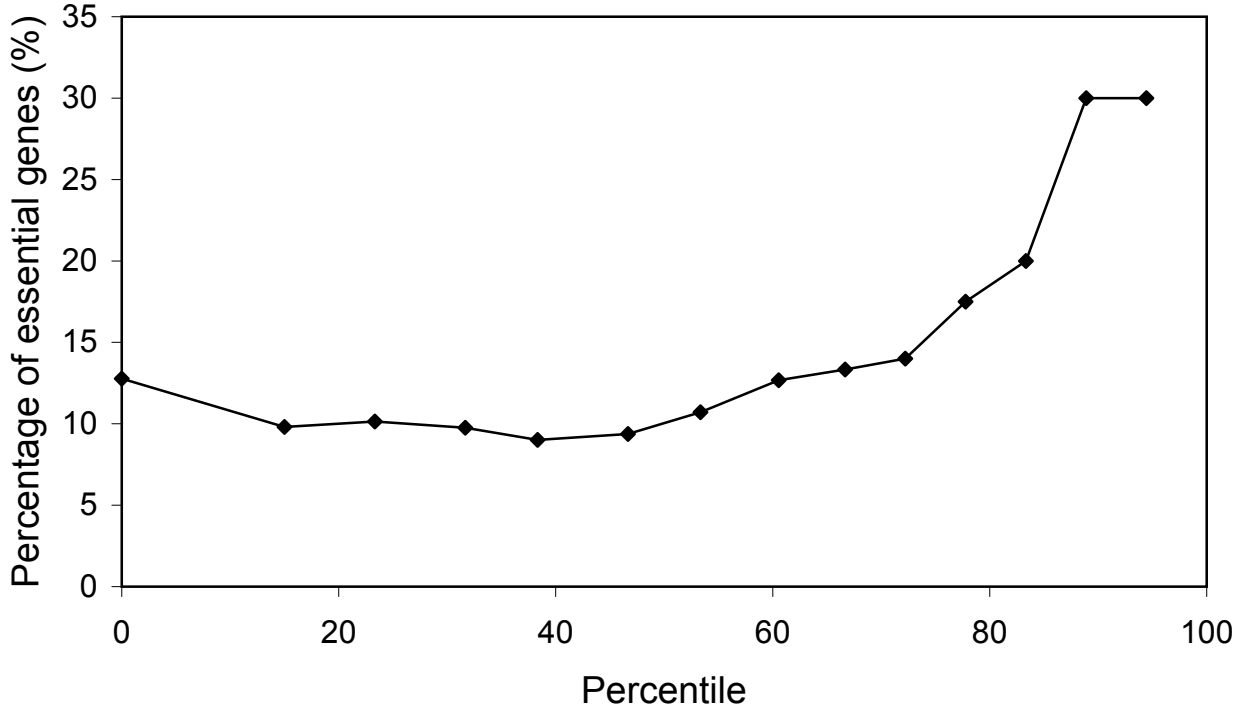
Minimum Values



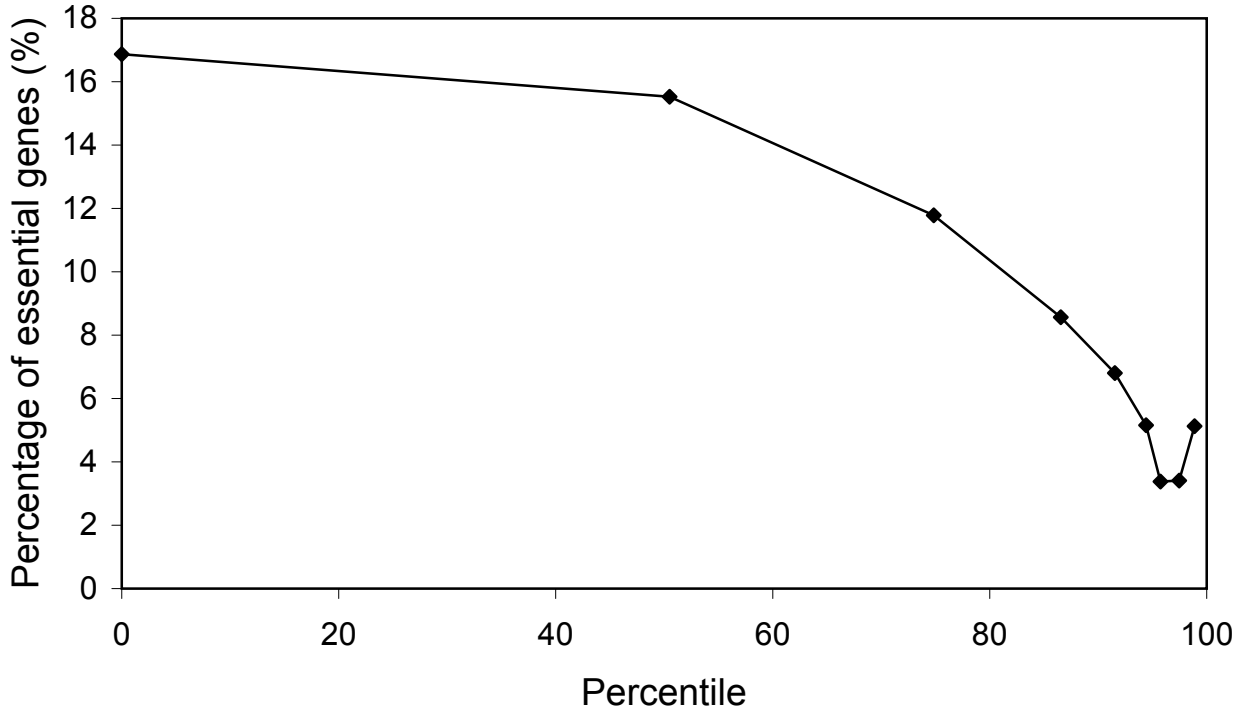

Percentile Rank
A
Average degree (K)
P < 10-12
Marginal essentiality
B
Clustering coefficient (C)
P < 10-25
Marginal essentiality
C
Characteristic path length (L)
P < 10-16
Marginal essentiality
D
Percentage of hubs
P < 10-42
Marginal essentiality

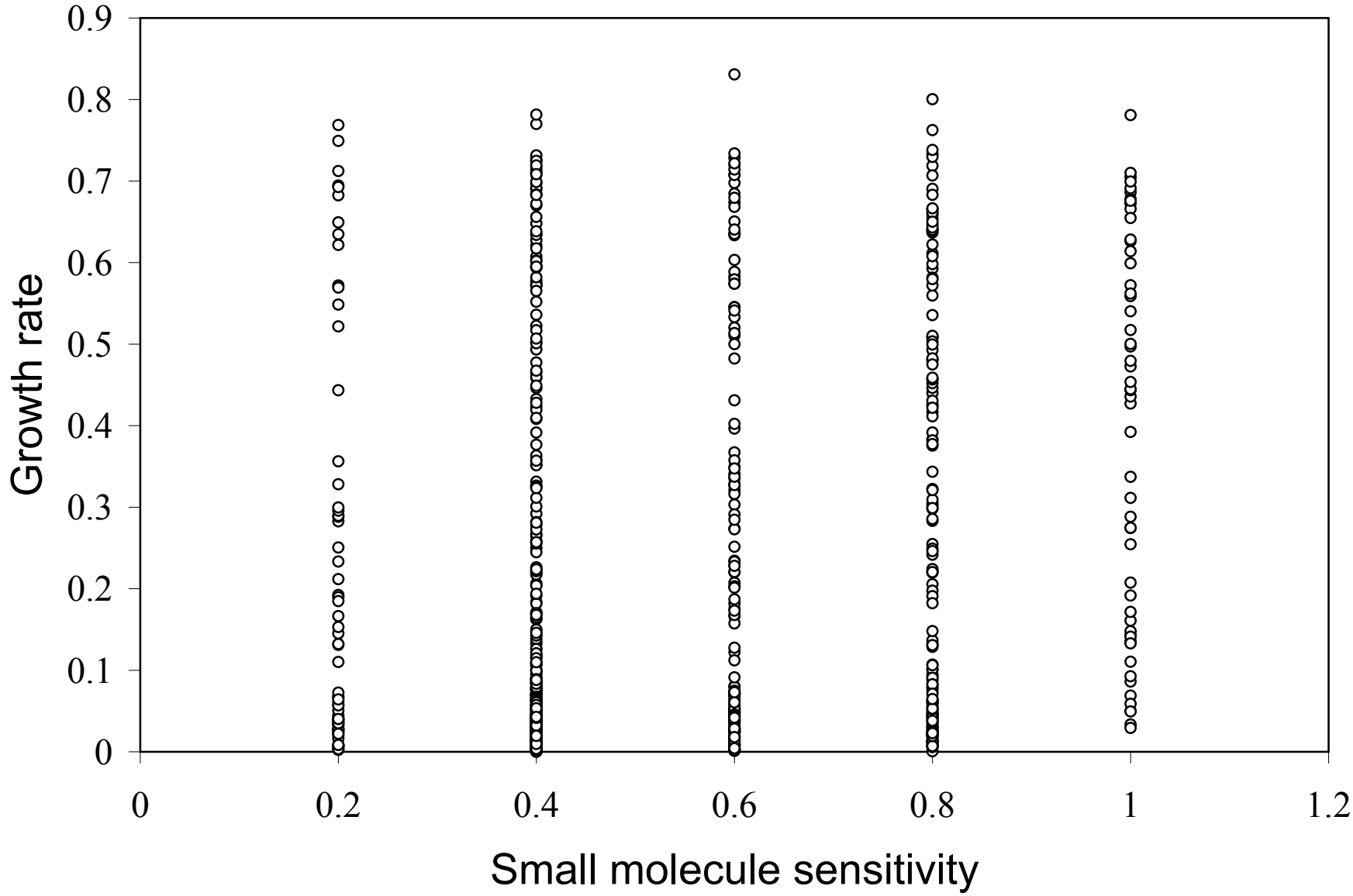

Z-score normalization
A
Average degree (K)
P < 10-20
Marginal essentiality
B
Clustering coefficient (C)
P < 10-22
Marginal essentiality
C
Characteristic path length (L)
P < 10-12
Marginal essentiality
D
Percentage of hubs
P < 10-30
Marginal essentiality

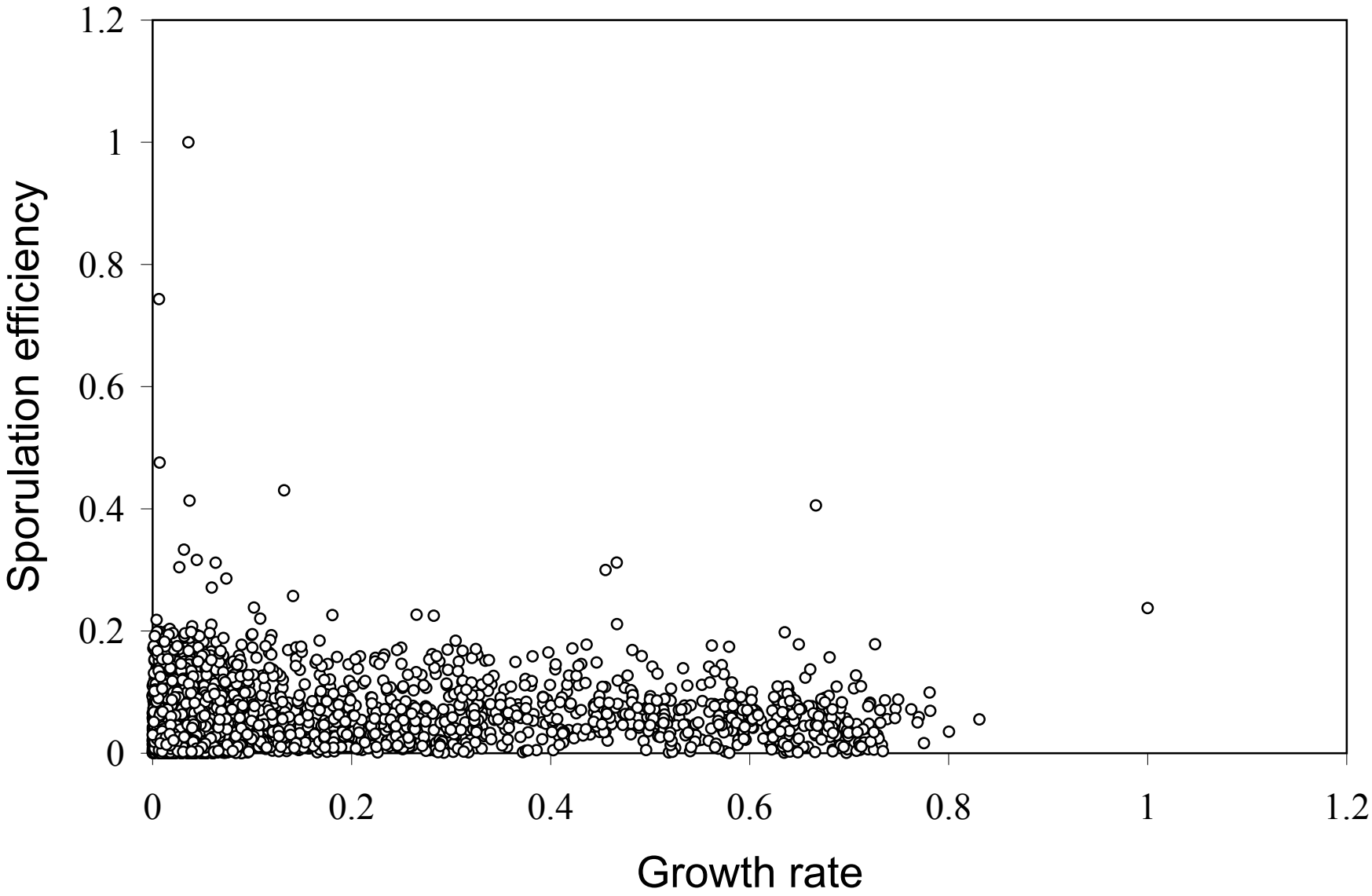
A. Regulator population

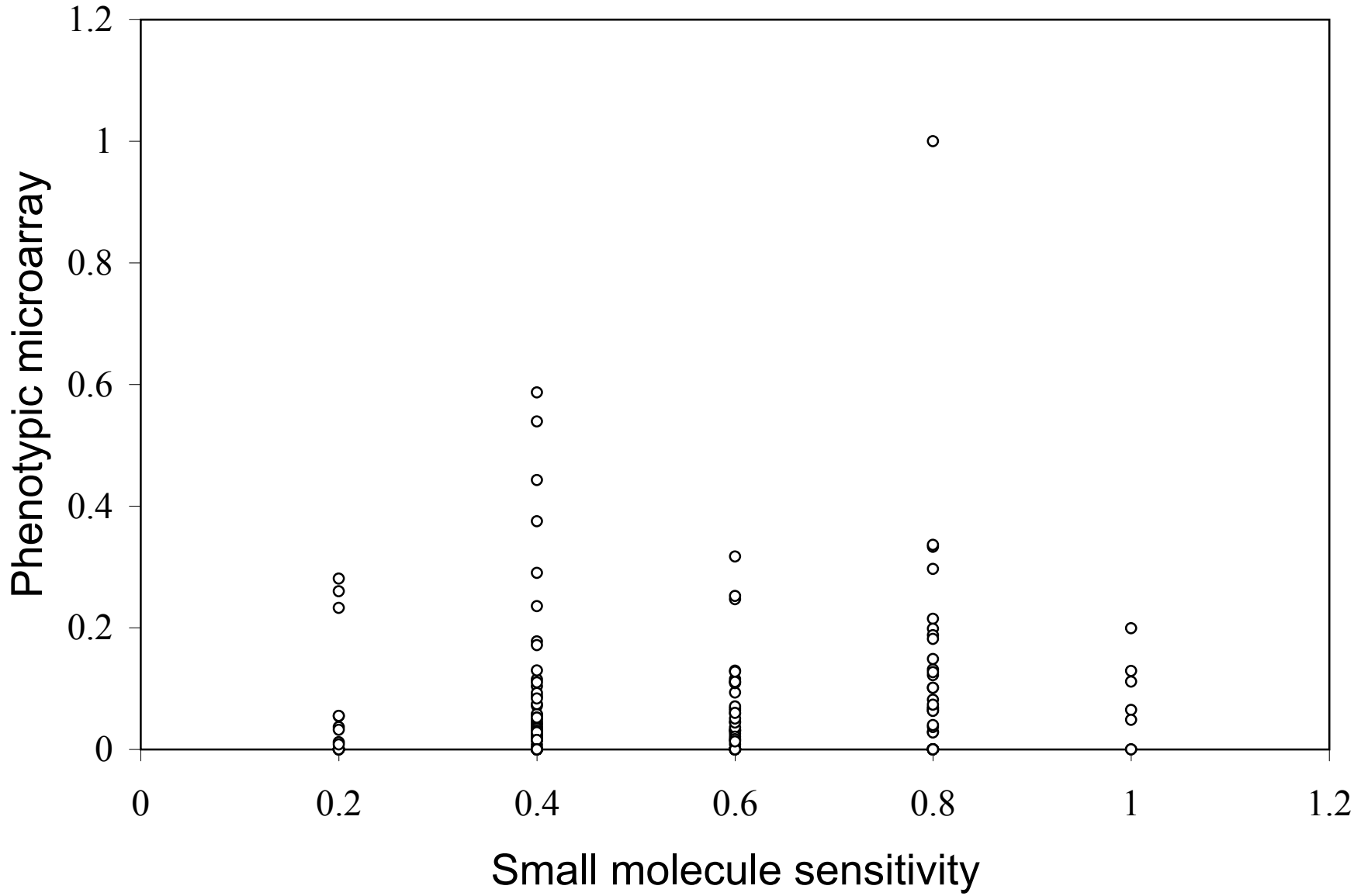


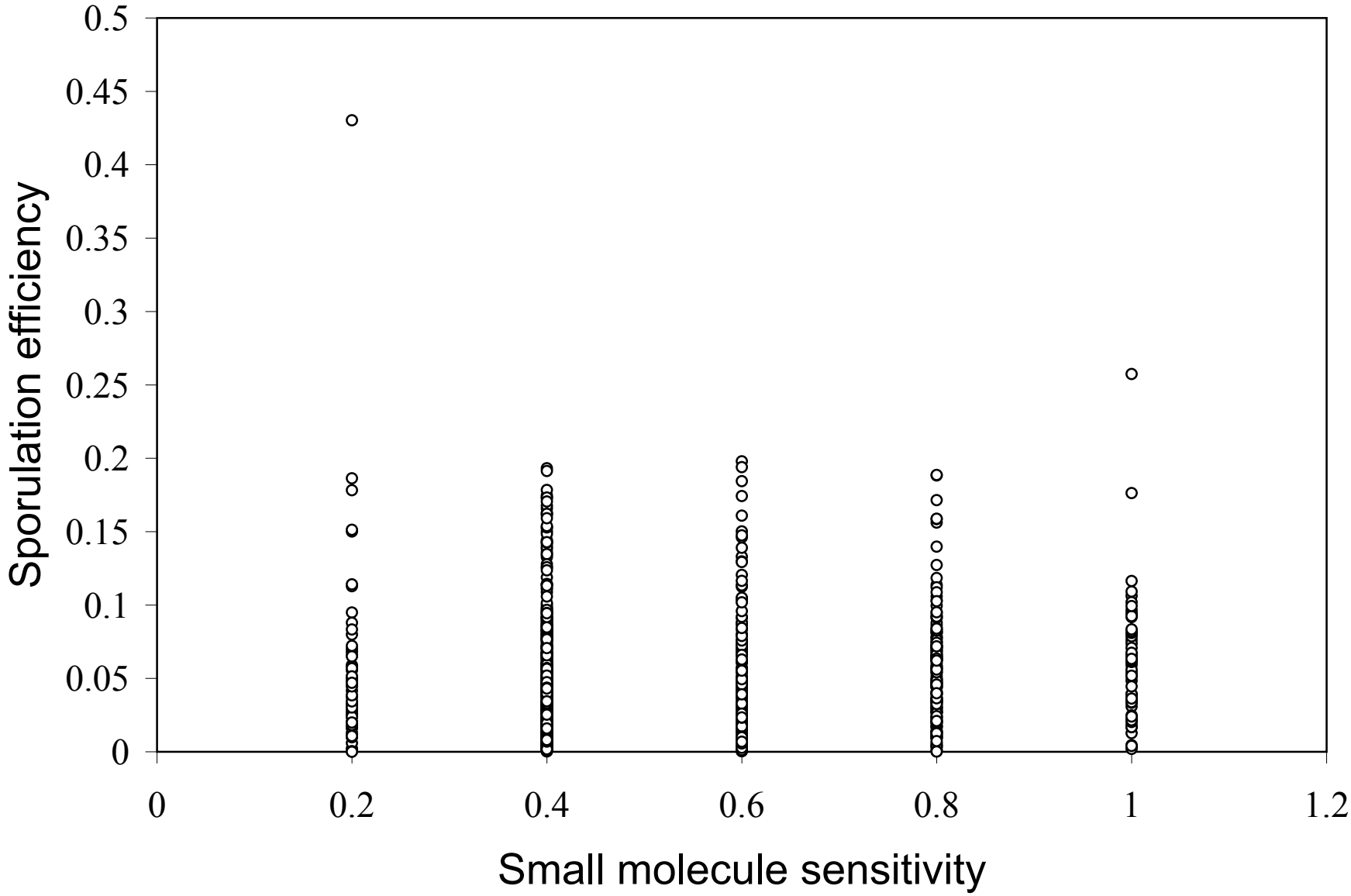
B. Target population

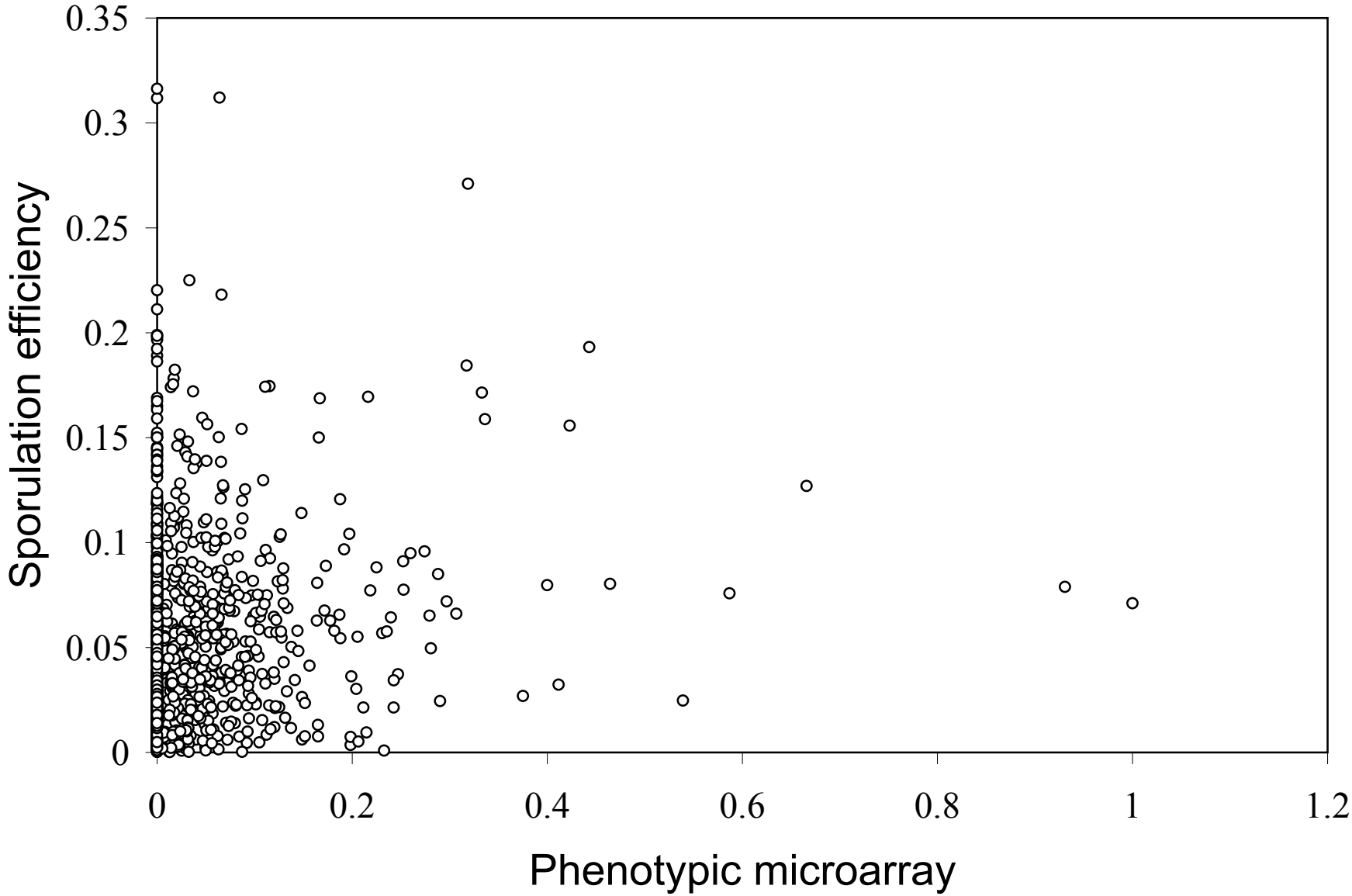


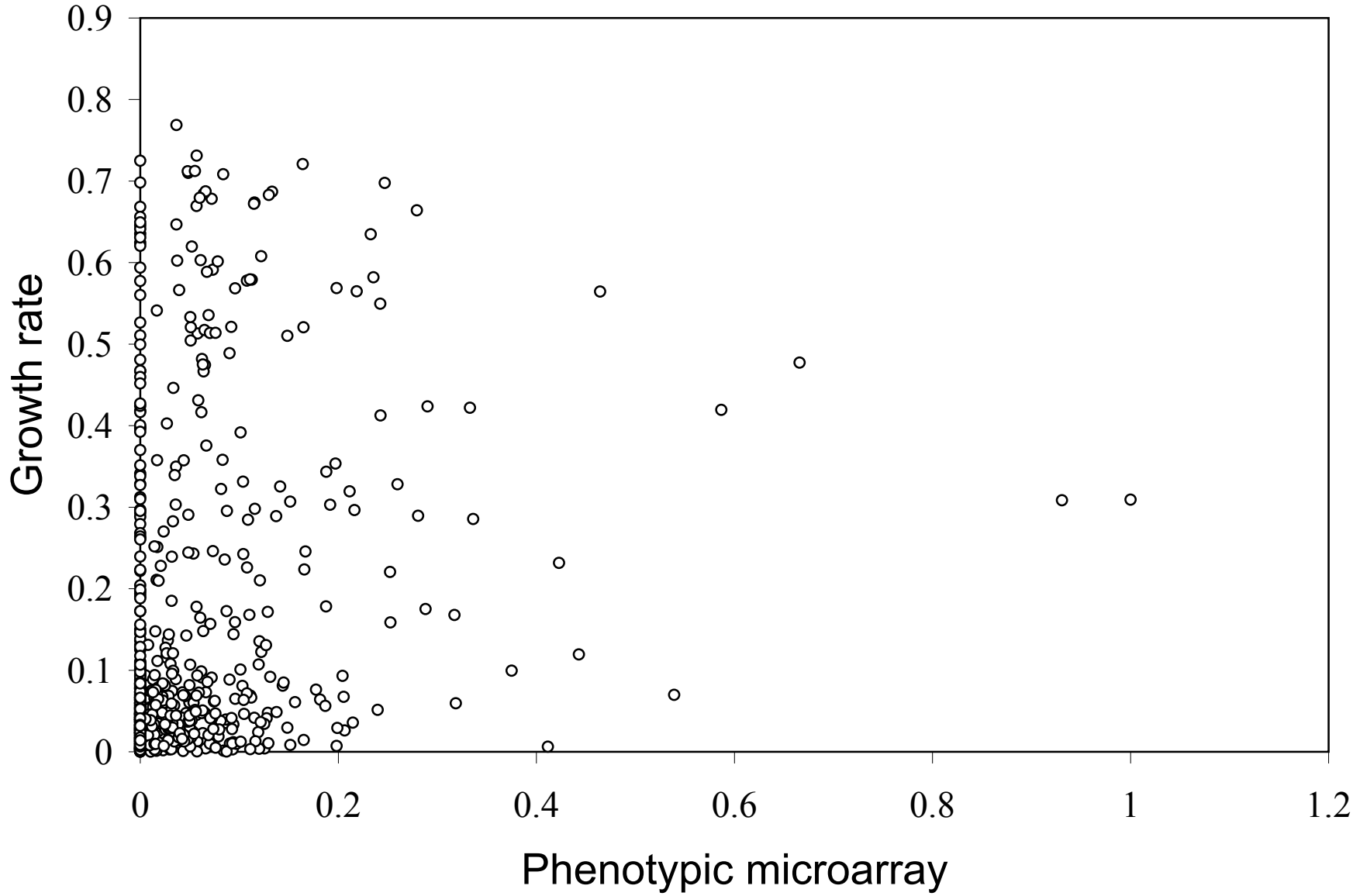

Growth rate
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
1.2
Small molecule sensitivity

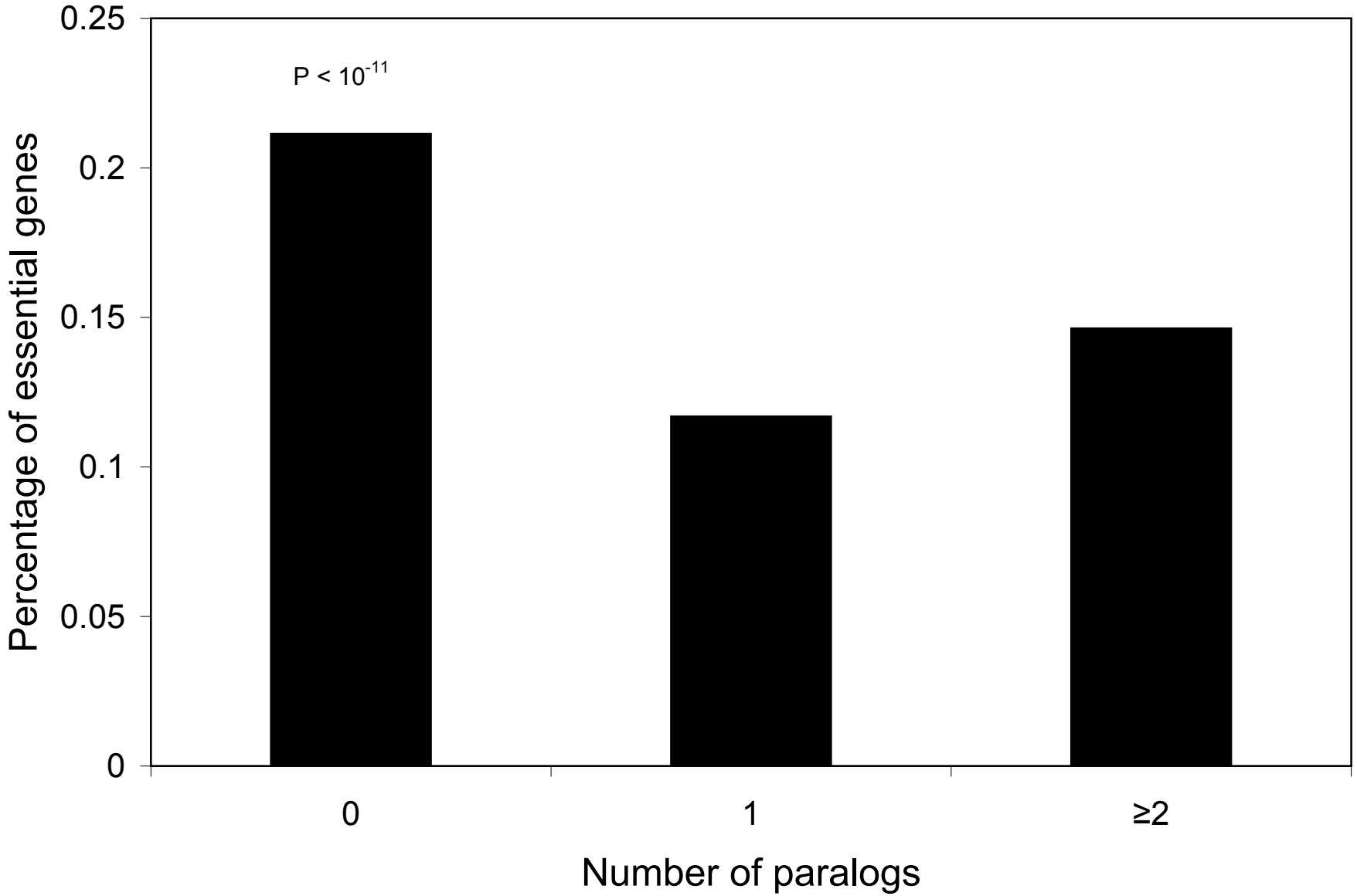

Sporulation efficiency
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
1.2
Growth rate


Phenotypic microarray
Small molecule sensitivity
0
0.2
0.4
0.6
0.8
1
1.2


Sporulation efficiency
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
Small molecule sensitivity


Sporulation efficiency
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
Phenotypic microarray


Growth rate
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
1.2
Phenotypic microarray


0.25
0.2
0.15
0.1
0.05
0
P < 10-11
Percentage of essential genes
0
1
≥2
Number of paralogs