EVOLUTIONARY SYSTEMS BIOLOGY

by

Zhi Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2010

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor Alexy Kondrashov
Assistant Professor Zhaohui Qin
Assistant Professor Patricia Jean Wittkopp

To My Wife and Daughter

ACKNOWLEDGEMENTS

With the completion of this dissertation, I want to give thanks to everyone who has helped me during this memorable 5-year journey.

My first and biggest thanks go to my advisor Jianzhi Zhang. It is my greatest privilege to have worked with such an exceptional advisor that everyone dreams of. I learned from him what a true scientist should be. Without his tremendous devotion, I would not be able to complete and publish any of my chapters. Therefore, I will continue to set him as an example of successful scientists, and try to follow the same way of doing research.

During these years, I have witnessed the succession of this wonderful lab, and met most of the people that have ever studied here. I still remember that I was warmly received by everyone when I joined the lab. I have benefited so much from joining the "afternoon talk" with Xionglei He and Peng Shi, which is also the place where the ideas for my chapters originated. In addition, I have received so many critical and constructive comments from Wenfeng Qian. I want to thank Wendy Grus and Margaret Bakewell for helping me with many kinds of English documents. Soochin Cho is like our big brother, who took care of every aspect of the laboratory life. I have to thank Ben-Yang Liao for giving me lots of encouragement in research. I enjoyed living as neighbors with his family as well. I also obtained quite a lot of help in computation from Zhihua Zhang and more recently from JianRong Yang, as well as his schoolmate Qi He, who helped me to solve mathematical problems. I learned terms and techniques for experimental work from Wenfeng Qian, and Calum Maclean, although I have not yet had a chance to do a real bench experiment. I really enjoyed the time spent with everyone in the lab, and this experience will be precious wealth in my memory.

I also would like to thank my committee members, Alexy Kondrashov, Patricia Wittkopp, and Zhaohui Qin, for their insightful comments on my dissertation, and for assisting me in applying for fellowships and jobs.

This research is partially funded by EEB department fellowship and Rackham One-Term Dissertation Fellowship.

Last but definitely not least, I have to thank my great family. In particular, I give thanks to my wife Li Ying for her unconditional and persistent love. Without her constant support and heartfelt devotion, I would not have completed my graduate study so smoothly. Additionally, I thank my little daughter Joyce Wang, who has brought so much happiness to my graduate life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

By analyzing complex biological networks, I explore the nascent field of systems biology to address some of the most long-lasting and difficult questions in genetics and evolution. First, I study modular structure and test whether the modular organization in cellular function arises from modularity in the underlying molecular interaction networks. I show that although protein interaction networks are highly modular, there is little evidence to suggest that these network modules correspond to functional units or that they are evolutionarily conserved. I demonstrate that network modules can originate simply as a byproduct of gene duplication. Then, I investigate another systems level feature, redundancy, the evolutionary maintenance of which is puzzling. I infer that 37-47% of reactions are functionally redundant in *E. coli* and yeast metabolic networks, but the majority of them are preserved because they are efficiently used under different conditions or their loss causes an immediate fitness reduction. These results challenge the adaptive backup hypothesis and suggest that genetic robustness is likely an evolutionary byproduct. Subsequently, I study the genomic pattern of pleiotropy, another systems attribute of genes. A low level of pleiotropy is observed for the majority of genes in multiple species. A greater per-trait effect size is also observed for genes affecting more traits, which leads to the highest rate of adaptation for organisms of intermediate complexity. These findings suggest that pleiotropy not only allowed but may have also promoted the origin of complexity. Lastly, I apply the systems approach to study protein evolutionary rate. Simulating thousands of nutritional conditions using metabolic networks, I find that there is no condition or combination of conditions for which the gene importance correlates well with the observed gene evolutionary rate. It suggests that the weakness of the empirical correlation between gene importance and evolutionary rate is factual rather than artifactual. Together, my studies using systems

approach deepen our understanding of the genetic systems and provide fresh perspectives on the fundamental characteristics of life.

CHAPTER 1

Introduction

1.1 INTRODUCTION

The reductionist approach has dominated biological research for over 50 years, yielding significant insights into the mechanisms of relatively simple biological phenomena. However, our mechanistic understanding is far less complete on complex biological phenomena, such as homeostasis (Lopez-Maury, Marguerat, and Bahler 2008), circadian rhythm (Takahashi et al. 2008) and cancer (Lazebnik 2010). These biological phenomena generally involve dynamic interactions among multiple molecules. Previous studies on small-scale genetic systems and pathways have revealed biological features that are not observable at the single gene level (Nishizuka 1986; Barinaga 1999). Nevertheless, these case studies are constrained by the availability of large-scale biological data, and thus the application of the conclusions is limited. More importantly, there are far more systemic properties that could not be revealed by small-scale case studies. Because a biological system is not simply an assembly of all of its parts, properties of the system cannot be fully understood without an integrative view of the whole system.

Systems biology is made possible by major advances in two fields. First, functional genomics provides not only functional information for every gene in a genome but also the information of interactions between genes (Uetz et al. 2000; Giot et al. 2003). Second, theoretical advances in several fields, especially network sciences, provide a theoretical framework for describing and testing hypotheses on the relationship between the structure and function of a system (Barabasi and Oltvai 2004). These two fields underwent rapid advancement in the first decade of 21st century, enabling systems

biology to thrive. In this dissertation, I explored the nascent field of systems biology to address some of the most long-lasting and difficult questions in evolutionary biology.

My dissertation is comprised of two main sections. In the first section (Chapters 2-4), I study the genome-wide patterns and evolutionary originations of three commonly observed systems level properties: modularity, redundancy, and pleiotropy.

In Chapter 2, I investigate the modular structure in protein interaction networks. Because cellular functions are organized in a highly modular manner (Hartman, Garvik, and Hartwell 2001; Wagner, Pavlicev, and Cheverud 2007), where each module is composed of a group of tightly linked components and performs a relatively independent task, I ask whether this modularity in cellular function arises from the modularity in the underlying molecular interaction networks. If the network modules correspond to functional units, it is also expected to find evolutionarily conserved modules across species. I examined yeast, fly, and nematode protein networks, but failed to find functionally cohesive and evolutionarily conserved modules. Using computer simulation, I demonstrated that the network modules can originate simply as a byproduct of the process of evolutionary by gene duplication.

Because there is no transmission of signal or mass through the protein interaction network, I decided to study a network that is biologically more meaningful. In Chapter 3, I study the metabolic network because (i) there is actual transmission of atoms through the network, (ii) our knowledge about metabolic networks is quite complete for several model organisms, and (iii) computational tools such as the flux balance analysis (FBA) allow inferences of metabolic functions from metabolic network structures (Price, Reed, and Palsson 2004; Palsson 2006). I use metabolic network analysis to address a key question in evolutionary systems biology, the maintenance of functional redundancy. I inferred that 37-47% of metabolic reactions in *E. coli* and yeast can be individually removed without blocking the production of any biomass component under any nutritional condition. However, the majority of these redundant reactions are preserved, because they have differential maximal efficiencies at different conditions or their loss causes an immediate fitness reduction that can only be regained via mutation, drift, and selection in evolution. The remaining redundancies are attributable to pleiotropic effects

or recent horizontal gene transfers. These results suggest that genetic robustness is likely an evolutionary byproduct.

In Chapter 4, I study another commonly observed attribute of genes, pleiotropy, which refers to the phenomenon of one gene affecting multiple distinct phenotypic traits (Tyler et al. 2009). Although the concept of pleiotropy has far-reaching implications, it remains one of the least characterized biological properties (Williams 1957; Albin 1993; Tyler et al. 2009). I found, based on yeast, nematode, and mouse functional genomic data, that the fraction of traits altered appreciably by the deletion of a gene is minute for most genes, and the gene-trait relationship is highly modular. The size of the phenotypic effect of a gene on a trait is approximately normally distributed with variable standard deviations for different genes, resulting in a greater per-trait effect size for genes affecting more traits. This property alleviates the "cost of complexity" (Orr 2000), leading to the highest rate of adaptation for organisms with intermediate levels of complexity. The findings explain why complex organisms could have evolved and suggest a potential limit of biocomplexity.

In the second section (Chapter 5), I use systems biology tools to address an evolutionary biology question. Slower evolution of functionally more important genes is widely regarded as the foremost principle of molecular evolution and is used by molecular biologists in daily practice (Karp 2008). However, recent genomic analysis of a diverse array of organisms found only weak negative correlations between the evolutionary rate of a gene and its functional importance (Hurst and Smith 1999; Zhang and He 2005). A frequently proposed explanation of the weakness of the correlation is that gene importance is measured under a benign lab condition and thus may differ substantially from the true value in the organism's natural environment (Wolf 2006). However, this hypothesis is difficult to test using traditional methods. In Chapter 5, I am able to show that this difficult question can be approached using systems biology tools. Simulating thousands of nutritional conditions, I test whether there is any condition or combination of conditions for which the importance of a metabolic enzyme gene, measured by FBA, correlates well with the observed rate of its sequence evolution. My result, however, is negative. This and other analyses led to my conclusion that the weakness of the correlation is factual rather than artifactual.

Overall, with the available high throughput functional genomic data, I constructed the complex biological networks and studied the genetic patterns and their evolutionary originations at systems level. This systems biology approach provides me with a fresh perspective on complex genetic and evolutionary phenomena.

1.2 REFERENCES

Albin RL. 1993. Antagonistic pleiotropy, mutation accumulation, and human genetic disease. Genetica 91:279-286.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101-113.

Barinaga M. 1999. Circadian rhythms. CRY's clock role differs in mice, flies. Science 285:506-507.

Giot L, Bader JS, Brouwer C, et al. 2003. A protein interaction map of Drosophila melanogaster. Science 302:1727-1736.

Hartman JL, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol 9:747-750.

Karp G. 2008. Cell and Molecular Biology. Hoboken, New Jersey: John Wiley & Sons, Inc.

Lazebnik Y. 2010. What are the hallmarks of cancer? Nat Rev Cancer 10:232-233.

Lopez-Maury L, Marguerat S, Bahler J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat Rev Genet 9:583-593.

Nishizuka Y. 1986. Studies and perspectives of protein kinase C. Science 233:305-312.

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13-20.

Palsson BO. 2006. Systems Biology: Properties of Reconstructed Networks. Cambridge: Cambridge University Press.

Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Takahashi JS, Hong HK, Ko CH, McDearmon EL. 2008. The genetics of mammalian circadian order and disorder: implications for physiology and disease. Nat Rev Genet 9:764-775.

Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays 31:220-227.

Uetz P, Giot L, Cagney G, et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623-627.

Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nat Rev Genet 8:921-931.

Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. Evolution 11:398-411.

Wolf YI. 2006. Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. Trends Genet 22:354-357.

Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

CHAPTER 2

In Search of the Biological Significance of Modular Structures in Protein Networks

2.1 ABSTRACT

Many complex networks such as computer and social networks exhibit modular structures, where links between nodes are much denser within modules than between modules. It is widely believed that cellular networks are also modular, reflecting the relative independence and coherence of different functional units in a cell. While many authors have claimed that observations from the yeast protein-protein interaction (PPI) network support the above hypothesis, the observed structural modularity may be an artifact because the current PPI data include interactions inferred from protein complexes through approaches that create modules (e.g., assigning pairwise interactions among all proteins in a complex). Here we analyze the yeast PPI network including protein complexes (PIC network) and excluding complexes (PEC network). We find that both PIC and PEC networks show a significantly greater structural modularity than that of randomly rewired networks. Nonetheless, there is little evidence that the structural modules correspond to functional units, particularly in the PEC network. More disturbingly, there is no evolutionary conservation among yeast, fly, and nematode modules at either the whole-module or protein-pair level. Neither is there a correlation between the evolutionary or phylogenetic conservation of a protein and the extent of its participation in various modules. Using computer simulation, we demonstrate that a higher-than-expected modularity can arise during network growth through a simple model of gene duplication, without natural selection for modularity. Taken together, our results suggest the intriguing possibility that the structural modules in the PPI network originated as an evolutionary byproduct without biological significance.

2.2 INTRADUCTION

Many complex networks are naturally divided into communities or modules, where links within modules are much denser than those across modules (Newman 2003) (Figure 2.1). For example, human individuals belonging to the same ethnic groups interact more than those from different ethnic groups (Lin 1999). Studying the modularity of a network not only provides structural information about the network, but may also reveal the underlying mechanisms that determine the network structure. The concept of modularity is not new to biologists. In fact, cellular functions are widely believed to be organized in a highly modular manner, where each module is a discrete object composed of a group of tightly linked components and performs a relatively independent task (Hartwell et al. 1999; Ihmels et al. 2002; Ravasz et al. 2002; Barabasi and Oltvai 2004; Wall, Hlavacek, and Savageau 2004). It is interesting to examine whether this modularity in cellular function arises from modularity in molecular interaction networks such as the transcriptional regulatory network and protein-protein interaction (PPI) network. Many authors have attempted to separate modules in the PPI network based on either the network topology alone or with additional information about gene function and expression (Spirin and Mirny 2003; Tornow and Mewes 2003; Pereira-Leal, Enright, and Ouzounis 2004; Poyatos and Hurst 2004; Chen and Yuan 2006; Farutin et al. 2006; Lu et al. 2006; Valente and Cusick 2006; Zhang, Liu, and Zhou 2006). They generally report high modularity in the PPI network, with evidence for a rough correspondence between PPI modules and functional units. All these analyses, however, suffered from a serious bias in the current PPI data. The PPI data include binary interaction information that is either directly obtained from experiments such as the yeast two-hybrid (Y2H) assay (Uetz et al. 2000; Ito et al. 2001), or indirectly inferred from stable protein complexes (Bader and Hogue 2002). High-throughput protein complex identification is usually mass-spectrometry-based (Gavin et al. 2002; Ho et al. 2002; Gavin et al. 2006; Krogan et al. 2006) (e.g., tandem-affinity purification). These methods involve the discovery of a complex of interacting proteins including a tagged bait protein, but do not provide information about direct pairwise protein-protein interactions (Bader and Hogue 2002; von Mering et al. 2002). Some small-scale biochemical methods, such as co-immunoprecipitation (Sacher et al. 2000) and affinity

precipitation (Fatica et al. 2002), can also identify protein complexes without providing pairwise protein interaction information. Protein complex data obtained by one of these methods are then translated into binary PPIs by either the "matrix" or the "spoke" model (Bader and Hogue 2002) (Figure 2.2). The matrix model assumes that all members of a protein complex interact with each other, whereas the spoke model assumes that all non-bait members of a complex interact with the bait. It is obvious that use of the matrix model creates PPI modules corresponding to protein complexes. The spoke model can also affect modularity because the bait is interpreted by the model as a hub (i.e., a highly connected node), while in reality it may not be a hub. Because the reliability of the two models is unknown, it is possible that the prevailing modularity of PPI networks is an artifact of these models. In this work, we explore the above possibility by analyzing the modularity of two yeast PPI networks. The first is referred to as the PIC network, as it is the *P*PI network *i*ncluding protein *c*omplex data, whereas the second is named the PEC network, as it the *P*PI network *e*xcluding all edges inferred from protein *c*omplexes. Because we are assessing the modularity of the PPI network per se, only the network topology will be used in separating modules. Our analyses show that although both PIC and PEC networks are highly modular, the identified modules lack obvious correspondence to functional units and are not evolutionary conserved. We use computer simulation to show that modularity can arise in a simple model of network growth through gene duplication, without the involvement of selection for modularity. Together, our findings suggest that structural modules in PPI networks may have arisen as an evolutionary byproduct without biological significance.

2.3 RESULTS

2.3.1 Do PPI networks show modular structures?

We downloaded the PPI data for the budding yeast *Saccharomyces cerevisiae* from the Munich Information Center for Protein Sequences (MIPS) (Guldener et al. 2006). The dataset was human-curated and contained mostly binary interactions directly observed in Y2H experiments. In addition, about 10% of the binary interactions in the dataset were inferred using either the spoke or matrix model from protein complexes

identified by high-confidence small-scale experiments. This entire dataset is referred to as the PIC network here. Based on the MIPS annotation, we removed from the PIC network those binary interactions that were inferred from protein complexes, resulting in the PEC network. Because it is only meaningful to separate modules within a connected part of a network, we studied the largest connected subset (i.e., the giant component), of a network. The giant component contains over 90% of all nodes in the yeast PPI network. For simplicity, we refer to the giant component of a network as the network, unless otherwise noted. Table 2.1 lists some important parameters for the PIC and PEC networks studied here.

The extent of modularity for a particular modular separation of a network is often measured by $M = \sum_{s=1}^{N} [\frac{l_s}{L} - (\frac{k_s}{2L})^2]$, where N is the number of modules, L is the total number of edges in the network, l_s is the number of edges within module s, and k_s is the sum of the degrees of the nodes in module s (Newman and Girvan 2004; Guimera and Amaral 2005). The degree of a node is simply the number of edges that the node has. The particular separation that maximizes M is considered the optimal modular separation and the corresponding M is referred to as the modularity of the network (Figure 2.1). In essence, M is the difference between the observed and expected proportions of within-module edges in the network. Here, the expected proportion is computed from a non-modular network where edges are equally likely to be within and between modules.

Several algorithms are available to separate a network into modules and obtain the maximal M. Empirical and simulation studies showed that the method of Guimera and Amaral (Guimera and Amaral 2005) has the best performance because it can give the most accurate module separation and highest M (Danon et al. 2005). We therefore used this method to separate modules in the yeast PIC and PEC networks. To obtain the highest M, we used delicate parameter settings in the simulated annealing algorithm. It took a typical desktop computer ~3 days to separate a yeast PPI network. The PIC network is separated into 26 modules with a modularity of 0.6672, while the PEC network is divided into 22 modules with a modularity of 0.6583 (Table 2.1). The density ratio, defined by the ratio of the number of within-module edges to the number of between-module edges is only slightly lower for PEC than for PIC networks (Table 2.1).

A random network may also have a non-zero modularity by chance or due to certain degree distributions (Guimera, Sales-Pardo, and Amaral 2004). Also, the modularity values of two networks with different sizes or different average degrees cannot be compared directly (Guimera, Sales-Pardo, and Amaral 2004). Thus, to measure the modularity of a network, we compare it with a random network of the same size and same degree distribution, which is generated by the local rewiring algorithm (Maslov, Sneppen, and Zaliznyak 2004). To speed up the computation, we used moderate parameter settings and faster runs (~ 4 hours per network) to estimate modularity. For the yeast PIC network, the modularity for 500 randomly rewired networks has a mean of 0.5466 and a standard deviation of 0.0023, while the real PIC network has a modularity of 0.6555 under this parameter setting (Figure 2.3A). We use z-score, or the number of standard deviations higher than the random expectation to measure the deviation of the modularity of a network from its random expectation. This z-score, referred to as the scaled modularity to differentiate it from z-scores of other properties, is (0.6555-0.5466)/0.0023 = 47 for the PIC network. Under the same parameter setting, the modularity for the real PEC network is 0.6481. The modularity for 500 randomly rewired PEC networks has a mean of 0.5764 and a standard deviation of 0.0027 (Figure 2.3B). In other words, the scaled modularity for the PEC network is (0.6481-0.5764)/0.0027 = 27. Thus, both PIC and PEC networks show significantly greater modularity than randomly rewired networks. As expected, the scaled modularity of PIC is much greater than that of PEC. This difference is largely due to the exclusion of protein complex data in the PEC network. In fact, when we randomly removed 10% of edges from the PIC network, the scaled modularity decreased only slightly (from 47 to 42).

Given the substantive difference in scaled modularity, PIC and PEC networks should also differ in the compositions of their modules. We measured the similarity in module composition between different separations of the same network (or shared nodes in the case of different networks) by the *n*ormalized *m*utual *i*nformation (*NMI*) index (Danon et al. 2005). A higher *NMI* indicates a higher similarity in module composition. The *NMI* between the PIC network and PEC network is 0.35. As a control, we measured the *NMI* between the PIC network and a reduced network generated by random removal

of 10% of the edges in PIC. This control *NMI* has a mean of 0.41 and a standard deviation of 0.018 (from 200 replications). Thus, the *NMI* between PIC and PEC is significant lower than that between PIC and its randomly reduced networks ($P < 0.002$) (Figure 2.3C). Because simulated annealing is a stochastic algorithm, different runs may yield slightly different partitions. We thus separated modules in PIC and PEC networks with different random seeds 50 times and these replications confirmed that the above finding of a lower *NMI* between PIC and PEC than by chance is genuine ($P < 10^{-10}$, Mann-Whitney U test). Together, these analyses demonstrate that the inclusion of interactions inferred from protein complexes in the PPI network has a great impact on network modularity.

2.3.2 Are structural modules functional units?

Because we identified the PPI modules based entirely on the topology of the network, it is important to ask whether such structural modules correspond to functional units. To address this question, we utilized the functional annotation of yeast genes in the CYGD database (Guldener et al. 2005). At the highest level of annotation, each yeast gene is classified into one or several of 17 functional categories (Figure 2.4). If the structural modules correspond to functional units, we should expect a nonrandom among-module distribution of the genes of a given functional category. For example, in the PIC network, there are 361 genes belonging to functional category A (cell type differentiation; see Figure 2.4A). A χ^2 test showed that these genes are not randomly distributed across the 26 PIC modules ($\chi^2 = 317$, df = 25, $P < 10^{-5}$; see the circles in Figure 2.4A). This test was conducted for each functional category and almost all functional categories showed significant nonrandom distributions across PIC modules (even after considering multiple testing). In contrast, the PEC network has fewer functional categories showing significant nonrandom distributions. This trend is particularly evident at the highest level of statistical significance (6 categories in PEC vs. 14 in PIC) (Figure 2.4B).

If structural modules correspond to functional units, we also expect that the majority of genes in a module belong to only one or a few functional categories. In other words, each module should have one or a small number of overrepresented functional

categories. Testing this prediction is not easy because one gene may belong to multiple functional categories. We thus used computer simulations. For example, module 1 of the PIC network comprises of 227 proteins, 92 of which belong to functional category A (Figure 2.4A). We randomly chose 227 genes from the network and counted the number of category A genes. We repeated this procedure 100,000 times to estimate the probability that the number of category A genes in the randomly picked 227 genes is equal to or greater than 92. This probability is indicated with different colors in the small squares of Figure 2.4A. Because 17 functional categories were tested for each module, to control for multiple testing we used 10^{-3} as the cutoff for statistical significance for each category. It can be seen that in 16 (62%) of the 26 PIC modules, at least one functional category is enriched. In comparison, only 7 (32%) of the 22 PEC modules have at least one enriched functional category. The above difference between PIC and PEC modules is statistically significant ($P < 0.05$, χ^2 test).

The two analyses above revealed nonrandom distributions of protein functions across structural modules. To quantitatively measure how well structural modules correspond to functional units, we used a correlation analysis. For a pair of proteins from a PPI network, we ask if they belong to the same module (co-membership) and if they belong to the same functional category (co-functionality). Two proteins are regarded to possess co-functionality as long as they share at least one function. If structural modules correspond well to functional units, protein pairs within the same module should share function whereas protein pairs across modules should not share function. In other words, we should observe a strong positive correlation between co-membership and co-functionality of protein pairs. We enumerated all possible protein pairs and found the correlation to be statistically significant in both PIC ($P < 10^{-300}$) and PEC ($P < 10^{-100}$) networks. However, the level of correlation is extremely low in both PIC ($r^2 = 0.0813\%$) and PEC ($r^2 = 0.00675\%$) networks (Figure 2.4C and 2.4D), indicating that less than 0.1% of the variance in protein-pair co-membership is explainable by co-functionality. We also found that the r value for PEC is significantly lower than that for PIC when we repeated module separations 50 times with different random seeds ($P < 10^{-5}$, Mann-Whitney U test). The observation of a low level of correlation is not due to the presence of many multifunctional proteins, because the low correlation is also observed even when

we consider only monofunctional proteins (r^2 = 0.0384% and $P < 10^{-37}$ for PIC; r^2 = 0.0331% and $P < 10^{-30}$ for PEC). Hence, although there is significant nonrandomness in protein functions across structural modules, the correspondence between structural modules and functional units is extremely weak in both PIC and PEC networks, especially the latter.

We also examined the cellular locations of each protein (Huh et al. 2003) and tested whether members of a structural module tend to be co-localized, as would be expected if structural modules represent functional units. Our results were generally similar to those for functional categories. Although some nonrandom patterns were observed, the correspondence between structural modules and cellular locations is extremely weak in both PIC and PEC networks, especially the latter (Figure C.1).

2.3.3 Are structural modules evolutionarily conserved?

If a structurally defined PPI module represents a functional unit, the composition of the module should be evolutionarily conserved. To test this prediction, we applied the same module separation algorithm to the fruit fly (*Drosophila melanogaster*) PPI network, which was constructed from binary PPIs obtained in high-throughput Y2H experiments (Giot et al. 2003). Because the fly data do not contain any interactions inferred from protein complexes, we expect that the fly PPI network behaves more similarly to the yeast PEC network than to the PIC network. We thus examine the evolutionary conservation of modular structures between the yeast PEC network and the fly network.

We separated the fly network into 27 modules, with a modularity of 0.6851 and a scaled modularity of 29 (Table 2.1). Hence, the scaled modularity of the fly network is comparable to that of the yeast PEC network (27). There are 691 orthologous proteins between the giant component of the yeast PEC network and that of the fly network. We here again use *NMI* to measure the similarity in module compositions between two networks. The *NMI* value between the yeast PEC and fly PPI networks is 0.14. If the modular structures are evolutionary conserved between the two networks, the above *NMI* value should be significantly greater than that between the actual fly network and a randomly separated yeast network. We randomly separated the yeast PEC network into

26 modules by conserving the actual module sizes and then computed *NMI* between the real fly modules and the randomly separated yeast modules. To make this comparison, we repeated this process 10,000 times and obtained the frequency distribution of *NMI* (Figure 2.5A). The observed *NMI* between the real fly and real yeast networks falls in the central part of the distribution, indicating that the yeast and fly modules are no more similar to each other than by chance ($P > 0.6$), and revealing a complete lack of evolutionary conservation in PPI modules between the two species.

Because modular structures are often hierarchically organized (Ravasz et al. 2002), it is possible that a low level of structure is evolutionary conserved despite the lack of conservation at the whole-module level. Pairwise relationships between proteins represent the lowest possible structure in the PPI network. We invented a *c*onservation *i*ndex for *p*airs of proteins (CI_P). Between species X and Y, CI_P is defined as the probability that the Y orthologs of two X proteins belonging to the same module in X also belong to the same module in Y. CI_P is 0.048 between the yeast and fly, which is not significantly different from the expectation derived by comparison of the fly network to a random separation of the yeast network ($P > 0.6$; 10,000 simulations; Figure 2.5B). Thus, even at the lowest structural level, yeast and fly modules are not evolutionarily conserved. Note that CI_P measures the conservation of co-membership in a module between two proteins, regardless of whether these two proteins interact with each other. CI_P does not measure the conservation of PPIs. If two yeast proteins engage in a PPI and their respective fly orthologs also engage in a PPI, these two PPIs are referred to as orthologous PPIs (Matthews et al. 2001). Between the yeast PEC and fly PPI networks, there are 45 orthologous PPIs. In comparison, between the fly network and 1,000 randomly rewired yeast networks (with the degree of each node unchanged), there are only 0.58 orthologous PPIs on average (standard deviation = 0.75). Thus, orthologous PPIs are evolutionary conserved between the two species.

We also examined the evolutionary conservation of structural modules between yeast and the nematode *Caenorhabditis elegans*. Although the PPI data for *C. elegans* are highly incomplete, with only 2387 proteins and 3825 interactions in the giant component, the results we obtained (Figure C.2) are similar to those from the comparison between yeast and fruit fly networks.

2.3.4 Does participation in different modules affect the evolutionary rate of a protein?

If structural modules represent functional units, proteins with links to many modules should be evolutionarily more conserved than those with links largely within a module, because multifunctional or pleiotropic proteins tend to be conserved (He and Zhang 2006a; Salathe, Ackermann, and Bonhoeffer 2006). Guimera and Amaral (Guimera and Amaral 2005) defined the *p*articipation *c*oefficient of a node by $PC = 1 - \sum_{i=1}^{N} (\frac{k_i}{k})^2$, where k is the degree of the node, k_i is the number of links from the node to any nodes in module i and N is the total number of modules. A high PC indicates that a node participates in the functioning of many modules. These authors found that the propensity of an enzyme gene to be lost during evolution is negatively correlated with the PC of the enzyme in the metabolic network (Guimera and Amaral 2005). Such an observation strongly suggests that the modular structure in the metabolic network has biological significance. It is therefore useful to examine PC for the proteins in the PPI network. It has previously been debated whether the degree of a protein in the PPI network influences its evolutionary rate (Fraser et al. 2002; Fraser, Wall, and Hirsh 2003; Jordan, Wolf, and Koonin 2003; Bloom and Adami 2004; Fraser and Hirsh 2004; Batada, Hurst, and Tyers 2006). Because past studies did not exclude PPIs inferred from protein complexes, it is possible that some of previous results were due to artifacts of such inferences. Separate analyses of the PIC and PEC networks may help answer this question.

We first measured the rate of protein evolution by the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) between orthologous genes of yeast species *S. cerevisiae* and *S. bayanus*. We chose this species pair because their divergence level is appropriate for obtaining informative and reliable d_N estimates (Zhang and He 2005). We found that the d_N of a protein is significantly negatively correlated with its total degree in the yeast PIC network ($P < 0.001$; Table 2.2), but not with its degree in the PEC network ($P > 0.4$). Thus, when protein complexes are not considered, there is no significant correlation between d_N and degree. When we separated the links of a node into within-module links and between-module links, we found a significant correlation

between d_N and the within-module degree (i.e., the number of within-module links) in the PIC network. This correlation is again absent in the PEC network, suggesting that the correlation between d_N and within-module degree is largely attributable to protein complexes. In neither the PIC nor the PEC network did we find a significant correlation between d_N and the between-module degree (i.e., the number of links across modules). Similar results were found between d_N and *PC* of a protein (Table 2.2). Furthermore, even when we divided the proteins into different topological roles by their *PCs* and degrees, as was done by Guimera and Amaral for the metabolic network, no significant correlation between these roles and d_N was observed (bottom two rows in Table 2.2).

We also measured the rate of protein evolution by the *p*ropensity of gene *l*oss (*PL*) across 12 fungal species whose draft genome sequences are available. The results obtained for *PL* are qualitatively similar to those for d_N (Table 2.2). Taken together, there is no observable impact of the within-module, between-module, or total PPI degree of a protein on its evolutionary rate when protein complexes are excluded. Furthermore, if structural modules correspond to functional units, a protein with higher participation in various modules should be more pleiotropic (or multifunctional) and thus should be more conserved in evolution (He and Zhang 2006a; Salathe, Ackermann, and Bonhoeffer 2006). However, we found no impact of the extent of participation in various modules on the evolutionary rate of a protein. This negative result is consistent with the idea that structural modules do not correspond to functional units.

The growth rate of a yeast strain with a gene deleted can measure the importance of the gene under the tested condition. Growth rate is known to be negatively correlated with the PPI degree of a gene (Jeong et al. 2001; Hahn and Kern 2005; Batada, Hurst, and Tyers 2006; He and Zhang 2006b). We confirmed this result in both PIC and PEC networks, although the significance is only marginal in the latter (Table 2.2). Interestingly, for both networks, this significance is also found for within-module degrees, but not for between-module degrees. This phenomenon may arise because the between-module degree is often much smaller (mean = 1.04 for PIC and 0.98 for PEC) than the within-module degree (mean = 2.70 for PIC and 2.48 for PEC) and thus contributes less to the total degree of a node. Growth rate also contains the information of gene essentiality, as essential genes have zero growth rates whereas nonessential genes

have positive growth rates. Thus, similar results are obtained when we analyze the genes by gene essentiality rather than growth rate.

2.3.5 Can modularity originate as an evolutionary byproduct?

Because both PIC and PEC networks have significantly higher modularity than that of their randomly rewired networks but the identified modules exhibit little biological significance, it is puzzling as to how the modular structure could have arisen in evolution. Earlier studies suggested that modularity can originate by gene duplication (Sole and Fernandez 2003; Hallinan 2004). However, in these studies modularity is defined by hierarchical clustering or clustering coefficient, which lacks an objective function to identify the best module separation and to compute network modularity. We thus conducted computer simulations to examine whether the network modularity as defined in this paper can arise from evolution by gene duplication. Because duplication-divergence models can generate many network features similar to real PPI networks (Sole and Fernandez 2003; Vazquez et al. 2003) and have clear biological bases (Wagner 2002; Wagner 2003; He and Zhang 2005), we simulated network growth by a duplication-divergence model starting from a pair of connected nodes. Briefly, at each step, a node (A) is randomly picked and duplicated along with all its edges to generate its paralogous node (A'). We refer to two edges, one from A and the other from A', as a pair of edges if they both end at the same third node. To simulate functional divergence after gene duplication, we randomly remove one edge from each pair of edges, until A and A' share 90% of edges. This duplication-divergence process was repeated 300 times to generate a network of 302 nodes. The resulting network has 212 nodes in its giant component (first row in Table C.1). We found the modularity and scaled modularity of this simulated network to be 0.6717 and 29, respectively (Figure 2.6; Table C.1). We conducted 10 simulation replications and all cases show similarly high modularity and scaled modularity that are comparable to those of the yeast and fly PPI networks (Table C.1). In fact, we found that many different combinations of simulation parameters can give rise to modular networks and the specific model of evolution by gene duplication (e.g., the subneofunctionalization model (He and Zhang 2005)) does not appear to matter much to the result of high modularity (data not shown). Although self-interactions can be

biologically important, they are not considered in our simulation because such interactions are disregarded in the module separation algorithm of Guimera and Amaral (Guimera and Amaral 2005).

2.4 DISCUSSION

In this work, we conducted a comprehensive analysis of modular structures in yeast protein interaction networks. Rather than lumping binary PPIs directly observed in experiments with those indirectly inferred from protein complexes, we separately analyzed the PIC network, which includes inferred binary PPIs, and the PEC network, which excludes inferred binary PPIs. This distinction is necessary because inferences of binary interactions from protein complexes introduce errors to the network structure, which hamper accurate measurement of network modularity. Given that protein complexes likely represent true (functional) modules in the network, the unanswered question is whether the network structure is still modular when the PPIs inferred from protein complexes (~10% in our PIC network) are removed. We found that both PIC and PEC networks are significantly more modular than expected by chance, the scaled modularity of the PIC network is substantively greater than that of the PEC network, and the module compositions of the two networks are significantly different. The latter two results are expected, because the current models for inferring binary PPIs from protein complexes tend to increase modularity. Consistent with these results, we found that the fruit fly PPI network, which is entirely based on experimentally determined binary PPIs, has a comparable scaled modularity to that of the yeast PEC network.

In spite of the presence of significant modularity in the yeast PEC network, the identified structural modules do not appear to correspond to functional units. This is reflected in three analyses. First, for some functional categories, their member genes are distributed randomly among structural modules. Second, for most structural modules, there are no enriched functional categories. Third, for protein pairs, the correlation (r^2) between co-membership in a module and co-functionality, although significantly greater than 0, is lower than 0.1%. Our results contradict several previous studies which claimed that PPI modules correspond well to functional units (Spirin and Mirny 2003; Tornow

and Mewes 2003; Pereira-Leal, Enright, and Ouzounis 2004; Poyatos and Hurst 2004; Chen and Yuan 2006; Farutin et al. 2006; Lu et al. 2006; Valente and Cusick 2006; Zhang, Liu, and Zhou 2006). This difference is in part owing to the inclusion of protein complexes in these early studies. Furthermore, some studies utilized more than the PPI network topology in separating modules. For example, Tornow and Mewes considered gene co-expression patterns (Tornow and Mewes 2003). Although such practices may help identify functional modules, they do not objectively evaluate whether the PPI network itself has a biologically meaningful modular structure. Many studies also suffered from the lack of an efficient algorithm to identify the maximum modularity, resulting in suboptimal modular separations with many small modules. For example, Pereira-Leal and colleagues (Pereira-Leal, Enright, and Ouzounis 2004) separated the yeast PPI network into 1046 modules, with an average size of 8 proteins per module. A small module may appear to have a better functional correspondence than a large module, because the chance probability of functional similarity among a few proteins is considerably greater than that among a large number of proteins. Because the module separation algorithm we used here is superior to the earlier algorithms (Danon et al. 2005), under the same definition of modularity our results are expected to be more reliable than those based on inferior algorithms.

Although many authors have claimed that PPI networks are modular with significant functional correspondence, none have examined the evolutionary conservation of PPI modules. By comparing the yeast PEC network and fly PPI network, we found that PPI modules are not more conserved than the chance expectation at the whole-module level. Furthermore, even at the protein-pair level, the PPI modules are not more conserved than by chance. These findings are consistent with our observation of minimal correspondence between yeast PEC modules and functional units. Interestingly, PPIs are found to be conserved between the yeast and fly, suggesting that the lack of conservation of modules cannot be trivially explained by the lack of conservation of individual interactions in the network.

The participation coefficient of a node measures the extent of the distribution of links from the node to all modules. If PPI modules correspond to functional units, proteins with high participation coefficients should have higher degrees of pleiotropy (or

multifunctionality) and be more conserved than those with low participation coefficients, because pleiotropic or multifunctional proteins are known to be evolutionary conserved (He and Zhang 2006a; Salathe, Ackermann, and Bonhoeffer 2006). This correlation was not observed in either the PIC or PEC network when either d_N or propensity for gene loss was used as a measure of a protein's evolutionary rate. Thus, the results again point to the lack of correspondence between PPI modules and functional units.

Taken together, our analyses strongly suggest that the yeast PEC network has a modular structure, which, nevertheless, lacks detectable biological significance. One may argue that the PEC network actually contains biologically important structural modules, but such modules are difficult to identify due to the incompleteness and inaccuracy of current PPI data. While this possibility cannot be entirely ruled out, we note that the PPI data we used here are generally regarded as of relatively high quality (Guldener et al. 2006). Furthermore, according to recent estimates, our PPI data should cover 25% to 50% of all PPIs in the yeast interactome (von Mering et al. 2002; Grigoriev 2003). Several observations, such as the negative correlation between the growth rate of a single-gene deletion yeast strain and the PPI degree of the gene (Table 2.2), suggest that the current PPI data contain biologically meaningful signals. An alternative explanation of the PPI modularity that lacks biological significance is that modularity may be an evolutionary byproduct. Inspired by earlier studies (Sole and Fernandez 2003; Hallinan 2004), we demonstrate by computer simulation that a simple model of gene duplication-divergence can generate networks with a scaled modularity comparable to that observed in the yeast and fly PPI networks. This result suggests that the modularity in the PPI networks may indeed have no biological significance and has not been under selection. Because gene duplication is the primary source of new genes and new gene functions (Zhang 2003), our simulation is biologically relevant. It is possible that evolutionary processes other than gene duplication also contributed to the origin of network modularity. For example, if assortative links (i.e., links between nodes of similar degrees) are disfavored, as has been observed in PPI networks (Maslov and Sneppen 2002; Newman 2002), modularity may arise. PPI networks also have clustering coefficients higher than the chance expectation, meaning that two proteins that both interact with the third one also tend to interact with each other (Barabasi and Oltvai

2004). Natural selection for higher clustering coefficients for some nodes of the network may also raise modularity.

It has been intensely debated as to whether there is a negative correlation between the PPI degree of a protein and the evolutionary rate (d_N) of the protein (Fraser et al. 2002; Fraser, Wall, and Hirsh 2003; Jordan, Wolf, and Koonin 2003; Bloom and Adami 2004; Fraser and Hirsh 2004; Batada, Hurst, and Tyers 2006). We found this correlation to be statistically significant for the PIC network, but not significant for the PEC network. These observations suggest that the significant correlation is simply due to lower evolutionary rates for proteins involved in protein complexes than those not involved in complexes. Our result is consistent with a recent study reporting the lack of a significant correlation when PPIs were curated from literature (Batada, Hurst, and Tyers 2006). Because proteins involved in complexes tend to have exceptionally high degrees as a result of indirect inference of PPIs by the matrix or spoke model, our result is also consistent with the finding that only the most prolific interactors tend to evolve slowly (Jordan, Wolf, and Koonin 2003). Recently, Han and colleagues (Han et al. 2004) classified hubs (i.e., high-degree nodes) in the PPI network into party hubs and date hubs. The former are those proteins whose interaction partners have similar expression profiles across various conditions, whereas the latter are those whose partners have different expression profiles. Party hubs have been interpreted as proteins that function within a biological process (or a functional module), whereas date hubs are thought to link different functional modules. Fraser reported that party hubs are evolutionarily more conserved than date hubs, and suggested that this pattern may reflect a tendency for evolutionary innovations to occur by altering the proteins and interactions between rather than within modules (Fraser 2005). A closer examination of the party hubs and their partners reveals that the majority of them form protein complexes, whereas date hubs and their partners do not form complexes. Thus, Fraser's observation is explainable by a lower evolutionary rate of proteins involved in complexes than those not in complexes, without invoking additional evolutionary forces.

PPI networks have been subject to many structural, functional, and evolutionary analyses in the past few years. Our results show that removing a small fraction (~10%) of PPIs that are inferred from protein complexes can have a substantial effect on the

analysis. This observation raises a warning about many results regarding PPI networks, because they have usually been based on the PIC network that contains many potentially false PPIs inferred for members of protein complexes. As such false interactions are not randomly distributed in the network, their potential detrimental effect is particularly alarming. The PIC data we used do not contain high-throughput protein complex data such as those in (Gavin et al. 2002; Ho et al. 2002). In many PPI databases, such as BIND (Bader, Betel, and Hogue 2003), DIP (Salwinski et al. 2004), and the new literature-curated dataset (Reguly et al. 2006), about half or more of the PPIs are inferred from protein complexes. The recent genome-wide surveys of all protein complexes in the yeast added even more complexes to the PPI data. Inclusion of inferred PPIs from these complexes would affect the network structure even more. We caution that use of such PPI data may produce misleading results.

Systems biology is a nascent field with many hopes as well as much hype (Kitano 2002). It has been of great interest to identify nonrandom topological structures such as motifs and modules in molecular networks (Ihmels et al. 2002; Milo et al. 2002; Guimera and Amaral 2005). Such nonrandom patterns are often interpreted as having functional significance and having been particularly favored by natural selection (Alon 2003; Wuchty, Oltvai, and Barabasi 2003; Guimera and Amaral 2005). While this may be true in many cases, a nonrandom network structure can also originate as a byproduct of other processes without having its own function. Recent studies suggested that motifs in transcriptional regulatory networks do not represent functional units and are not subject to natural selection (Mazurie, Bottani, and Vergassola 2005). Rather, random gene duplication and mutation could give arise to motifs (Dwight Kuo, Banzhaf, and Leier 2006). A recent study even suggested that the high abundance of feed forward loops in regulatory networks could be an evolutionary byproduct (Cordero and Hogeweg 2006). Our results add yet another network structure that is widely believed to be of great biological importance to this growing list of potential evolutionary byproducts. That being said, the modular organization of cellular functions is real, and whether this organization is also an evolutionary byproduct or has been actively selected for remains to be scrutinized.

2.5 MATERIALS AND METHODS

2.5.1 The yeast, fly, and nematode PPI networks

The budding yeast (*Saccharomyces cerevisiae*) PPI data were from the MPact dataset (Guldener et al. 2006) of MIPS (ftp://ftpmips.gsf.de/yeast/PPI/PPI_18052006.tab), which contains human-curated high-throughput and small-scale binary interactions directly observed in experiments, as well as binary interactions inferred from high-confidence protein complex data. Only non-self physical interactions were considered. After excluding PPIs involving mitochondrial genes, we built the PPI network named PIC (*P*PI *i*ncluding protein *c*omplexes). The giant component of the PIC network is composed of 3886 proteins linked by 7260 nonredundant interactions. To build the PEC (*P*PI *e*xcluding protein *c*omplexes) network, we retained only those binary interactions in the PIC network that had direct experimental evidence. The giant component of the PEC network contains 3696 proteins linked by 6403 interactions.

The fruit fly (*Drosophila melanogaster*) PPI data came from (Giot et al. 2003) (http://www.bme.jhu.edu/labs/bader/publications/giot_science_2003/flyconf.txt). A moderate confidence level (0.25) was chosen to generate the fly PPI network with a comparable average degree to the yeast PEC network. In total, the giant component of the fly PPI network contains 6280 proteins linked by 10210 interactions, all generated by Y2H experiments.

The nematode (*Caenorhabditis elegans*) PPI data were from (Li et al. 2004) (http://vidal.dfci.harvard.edu/interactomedb/WI5.txt). Only the PPIs identified by Y2H experiments are used. In total, the nematode PPI network contains 2624 proteins and 3967 interactions, of which 2387 proteins and 3825 interactions are in the giant component.

2.5.2 Functional categories of yeast proteins

We used the yeast functional annotations in the CYGD database (Guldener et al. 2005) (ftp://ftpmips.gsf.de/yeast/catalogues/funcat/funcat-2.0_data_18052006), considering only the highest level of annotation. Functional categories containing < 15 proteins and the category of unknown functions were removed. The cellular localization

data for yeast proteins were from (Huh et al. 2003) (http://yeastgfp.ucsf.edu/allOrfData.txt). Similarly, ambiguous localizations and localizations with < 15 proteins were not used.

2.5.3 Evolutionary conservation of modules and proteins

The list of orthologous genes between the yeast and fly was provided by He and Zhang (He and Zhang 2006b), who used reciprocal best-hits in BLASTP searches to define gene orthology (E-value cutoff = 10^{-10}). The same method was used to identify the yeast and nematode orthologous genes. The d_N values between *S. cerevisiae* and *S. bayanus* orthologous genes were computed by a likelihood method and obtained from Zhang and He (Zhang and He 2005). We used the parsimony principle to infer the propensity of gene loss (i.e., the number of gene loss events) for each of the *S. cerevisiae* genes throughout the known phylogeny of 12 fungi. The protein sequences predicted from the complete genome sequences of the 12 species were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequences (*S. cerevisiae*, *S. bayanus*, *S. paradoxus*, and *S. mikatae*), ftp://ftp.ncbi.nih.gov/genomes/Fungi (*Candida glabrata*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, and *Yarrowia lipolytica*), http://www.broad.mit.edu/annotation/genome/neurospora/Home.html (*Neurospora crassa*), http://www.broad.mit.edu/seq/YeastDuplication (*Kluyveromyces waltii*), and http://www.sanger.ac.uk/Projects/S_pombe/ (*Schizosaccharomyces pombe*). A *S. cerevisiae* gene is considered to be lost in species X if it does not hit any genes in X (Evalue cutoff = 10^{-1}) but has a hit in at least one species that is more distantly related to *S. cerevisiae* than is X related to *S. cerevisiae*. Here X refers to one of the 10 fungi that are neither *S. cerevisiae* nor *S. pombe*, the latter being the most distantly related species to *S. cerevisiae* in our study.

The growth rates of the yeast single-gene deletion strains were originally generated by Stanford Genome Technological Center (Steinmetz et al. 2002), and we here used the dataset curated and provided by Zhang and He (Zhang and He 2005).

2.5.4 Normalized mutual information (*NMI*)

NMI was described in detail in (Danon et al. 2005). Briefly, let us define the matrix N, where each row corresponds to a module in separation X and each column corresponds to a module in separation Y. Each member N_{ij} in the matrix represents the number of nodes in the ith module of X that appear in the jth module of Y. The calculation of *NMI* is given by

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{n_X} N_{i.}\log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{n_Y} N_{.j}\log\left(\frac{N_{.j}}{N}\right)}$$

where n_X and n_Y are the number of modules in module separation X and Y, respectively. The sum over row i of matrix N_{ij} is denoted $N_{i.}$ and the sum over column j is denoted $N_{.j}$. If two module separations are identical, the *NMI* between them reaches the maximum value of 1.

2.5.5 Data and program availability

Datasets used in this work and computer programs made for the analyses can be downloaded from http://www.umich.edu/~zhanglab/download.htm.

2.6 ACKNOWLEDGMENTS

We thank Roger Guimera and Luis Nunes Amaral for providing the module separation program and Margret Bakewell, Wendy Grus, Xionglei He, Ben-Yang Liao, Zhihua Zhang, and three anonymous reviewers for valuable comments. This work was supported in part by research grants from the National Institutes of Health and University of Michigan to JZ.

Figure 2.1 An example of network modular structure. (**A**) A small network with modular structure. (**B**) A randomly rewired network of (**A**). Different colors show different modules separated by Guimera and Amaral's algorithm (Guimera and Amaral 2005). The modularity is 0.5444 for the network in (**A**) and 0.2838 in (**B**), and the scaled modularity is 15 for the network in (**A**) and 0 in (**B**).

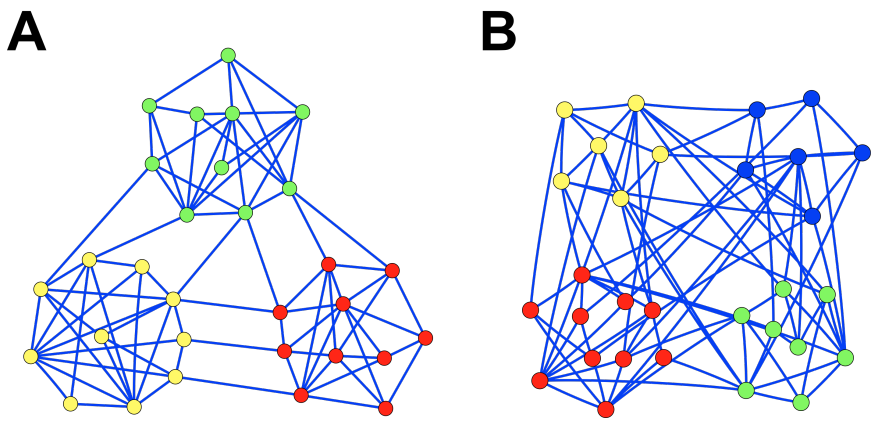


Figure 2.2 PPI network representations of protein complexes. (**A**) A hypothetical protein complex. Binary protein-protein interactions are depicted by direct contacts between proteins. Although five proteins (A, B, C, D and E) are identified through the use of a bait protein (red), only A and D directly bind to the bait. (**B**) The true PPI network topology of the protein complex. (**C**) The PPI network topology of the protein complex inferred by the "matrix" model, where all proteins in a complex are assumed to interact with each other. (**D**) The PPI network topology of the protein complex inferred by the "spoke" model, where all proteins in a complex are assumed to interact with the bait; no other interactions are allowed.

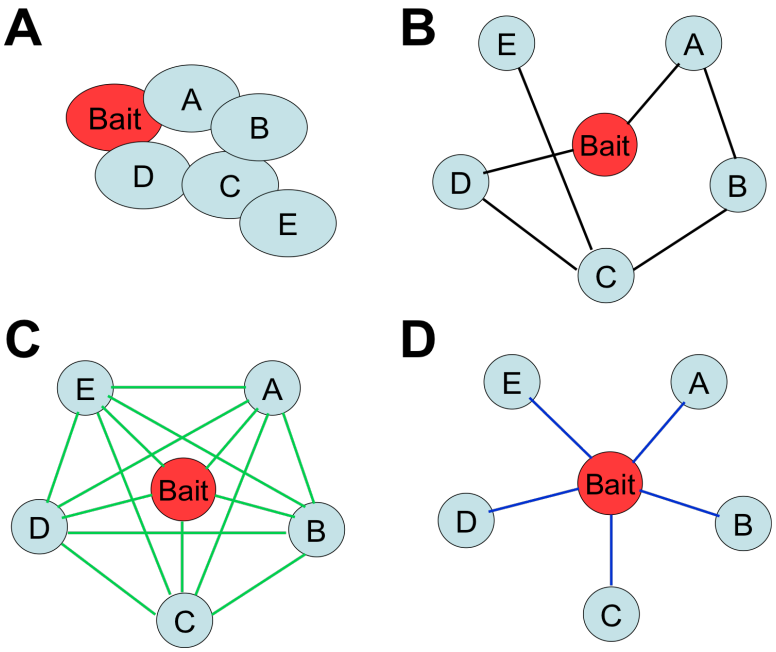


Figure 2.3 Modularity of yeast PIC and PEC networks. The modularity of yeast (**A**) PIC and (**B**) PEC networks compared to that of their randomly rewired networks, and (**C**) the similarity of module compositions between PIC and PEC networks compared to the random expectation. In (**A**) and (**B**), the observed modularity is indicated by the vertical arrow. The bars show the frequency distribution of the modularity from 500 randomly rewired networks. Scaled modularity, or the difference between the modularity of a real network and the expected modularity of a randomly rewired network in terms of the number of standard deviations, is indicated at the top area of the panel. In (**C**), the observed similarity between PIC and PEC networks, measured by *NMI* (normalized mutual information), is indicated by the vertical arrow. The bars show the frequency distribution of the *NMI* between PIC and 200 reduced networks (by random removal of 10% edges from the PIC network). The result shows that the difference between PEC and PIC is not simply because the PEC network is 10% smaller than the PIC network.

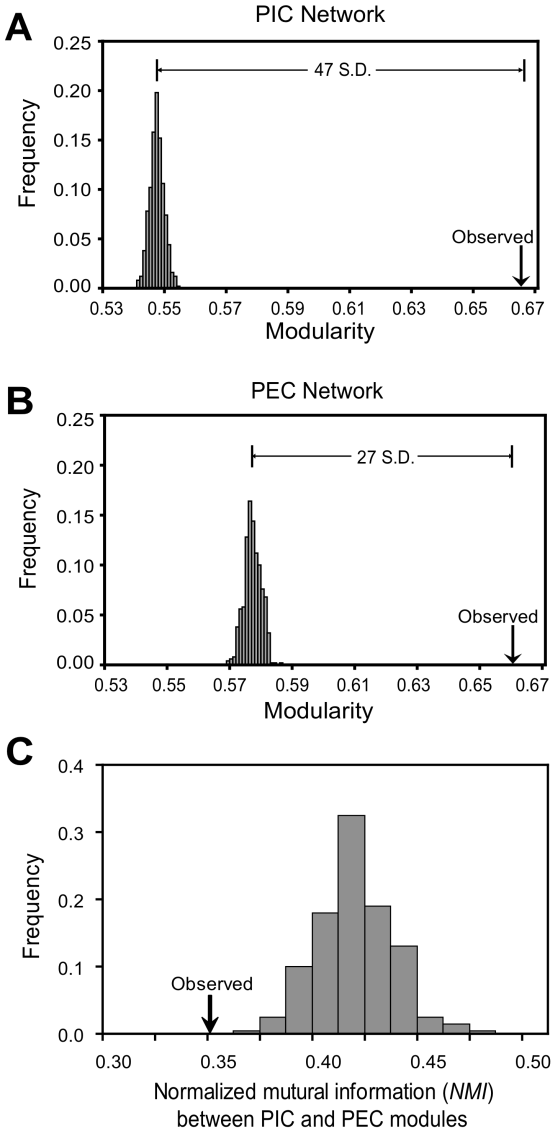


Figure 2.4 Lack of obvious correspondence between structural modules and functional units. In (**A**) and (**B**), each functional category is indicated by a letter (A to Q). In parentheses next to the letter is the percentage of proteins in the network that belong to that functional category. Note that one protein may belong to more than one category. The circles next to the grid show the statistical significance of nonrandom distributions of genes of the same functional categories across modules. Each small square in the grid shows the statistical significance of enrichment of a particular function in a module. For the circles and squares, significance levels are indicated by different colors. Panels (**C**) and (**D**) show the correlation between co-membership in structural modules and co-functionality for all pairs of proteins in the PIC and PEC networks, respectively. The circle size is proportional to the number of protein pairs. The line shows the linear regression and *r* is the correlation coefficient.

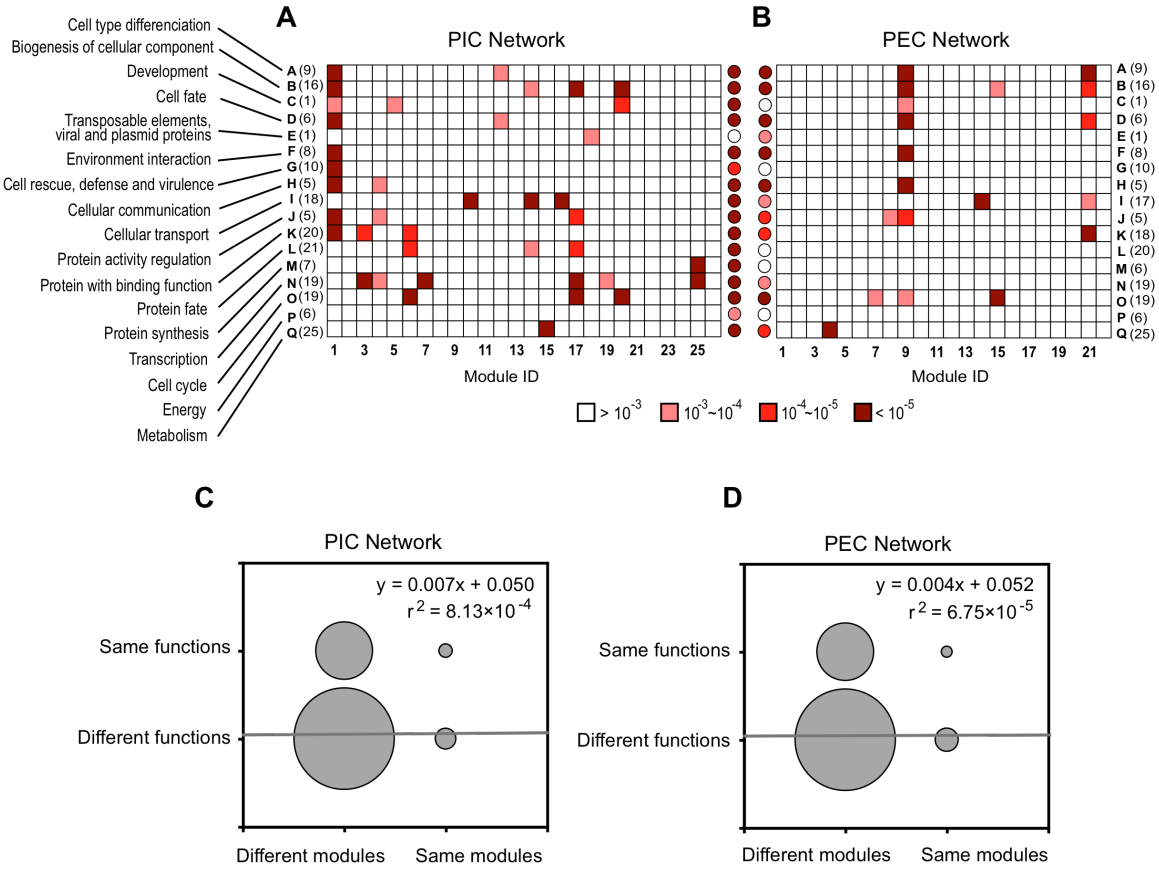


Figure 2.5 Lack of evolutionary conservation between the yeast and fruit fly PPI modules. (**A**) The observed *NMI* (normalized mutual information) between yeast and fruit fly modules is not significantly different from the chance expectation. The bars show the distribution of *NMI* between the yeast and fly modules when the yeast modules are randomly separated. (**B**) The observed CI_P (conservation index for pairs of proteins) between yeast and fruit fly modules is not significantly different from the chance expectation. The bars show the distribution of CI_P between the yeast and fly modules when the yeast modules are randomly separated.

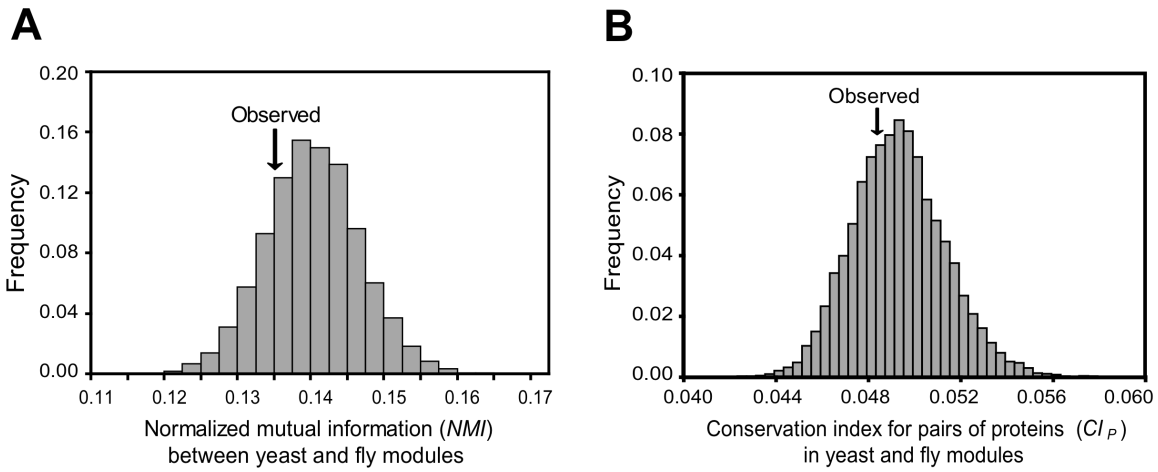


Figure 2.6 A random network generated by gene duplication followed by subfunctionalization shows high modularity. Modularity = 0.6717, and scaled modularity = 29. Different colors represent different modules identified by Guimera and Amaral's algorithm (Guimera and Amaral 2005).

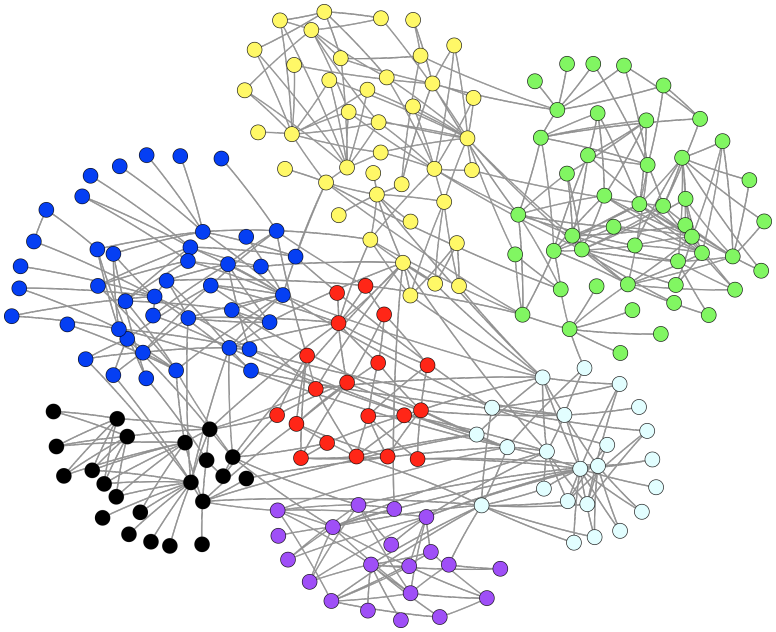


Table 2.1 Summary statistics of the giant component of the protein interaction networks.

	No. of proteins	No. of interactions	Average degree	No. of modules	Average number of proteins per module	Density ratio	Modularity	Scaled modularity
Yeast PIC network	3885	7260	3.74	26	149	2.60	0.6672	47
Yeast PEC network	3695	6403	3.47	22	168	2.52	0.6583	27
Fly PPI network	6279	10094	3.22	27	232	2.99	0.6851	29

Table 2.2 Relationship between the degree of a protein and its importance to growth or evolutionary rate.

	PIC Network			PEC Network		
	d_N	Propensity for gene loss	Growth rate of single-gene deletion strain	d_N	Propensity for gene loss	Growth rate of single-gene deletion strain
Total degree	-0.06* (< 0.001)[δ]	-0.03 (0.055)	-0.05 （0.004）	-0.01 (0.464)	-0.01 (0.507)	-0.04 (0.057)
Within-module degree	-0.07 (< 0.001)	-0.04 (0.014)	-0.07 （< 0.001）	-0.03 (0.104)	-0.02 (0.218)	-0.04 (0.023)
Between-module degree	0.00 (0.866)	-0.01 (0.594)	-0.02 (0.224)	0.01 (0.479)	-0.01 (0.732)	-0.02 (0.315)
Participation coefficient	0.01 (0.545)	0.00 (0.835)	-0.01 (0.524)	0.03 (0.143)	0.00 (0.934)	-0.02 (0.428)

* Spearman rank correlation coefficient

[δ] *P* value of the Spearman rank correlation

2.7 REFERENCES

Alon U. 2003. Biological networks: the tinkerer as an engineer. Science 301:1866-1867.

Bader GD, Betel D, Hogue CW. 2003. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31:248-250.

Bader GD, Hogue CW. 2002. Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 20:991-997.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101-113.

Batada NN, Hurst LD, Tyers M. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol 2:e88.

Bloom JD, Adami C. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC Evol Biol 4:14.

Chen J, Yuan B. 2006. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22:2283-2290.

Cordero OX, Hogeweg P. 2006. Feed forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23:1931-1936.

Danon L, Diaz-Guilera A, Duch J, Arenas A. 2005. Comparing community structure identification. J. Stat. Mech. P09008:1-10.

Dwight Kuo P, Banzhaf W, Leier A. 2006. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. Biosystems 85:177-200.

Farutin V, Robison K, Lightcap E, Dancik V, Ruttenberg A, Letovsky S, Pradines J. 2006. Edge-count probabilities for the identification of local protein communities and their organization. Proteins 62:800-818.

Fatica A, Cronshaw AD, Dlakic M, Tollervey D. 2002. Ssf1p prevents premature processing of an early pre-60S ribosomal particle. Mol Cell 9:341-351.

Fraser HB. 2005. Modularity and evolutionary constraint on proteins. Nat Genet 37:351-352.

Fraser HB, Hirsh AE. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol Biol 4:13.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750-752.

Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11.

Gavin AC, Aloy P, Grandi P, et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631-636.

Gavin AC, Bosche M, Krause R, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141-147.

Giot L, Bader JS, Brouwer C, et al. 2003. A protein interaction map of Drosophila melanogaster. Science 302:1727-1736.

Grigoriev A. 2003. On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res 31:4157-4161.

Guimera R, Amaral LAN. 2005. Functional cartography of complex metabolic networks. Nature 433:895-900.

Guimera R, Sales-Pardo M, Amaral LAN. 2004. Modularity from fluctuations in random graphs and complex networks. Phys Rev E Stat Nonlin Soft Matter Phys 70:025101.
Guldener U, Munsterkotter M, Kastenmuller G, et al. 2005. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res 33:D364-368.
Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V. 2006. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34:D436-441.
Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803-806.
Hallinan J. 2004. Gene duplication and hierarchical modularity in intracellular interaction networks. Biosystems 74:51-62.
Han JD, Bertin N, Hao T, et al. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430:88-93.
Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. Nature 402:C47-52.
He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157-1164.
He X, Zhang J. 2006a. Toward a molecular understanding of pleiotropy. Genetics 173:1885-1891.
He X, Zhang J. 2006b. Why do hubs tend to be essential in protein networks? PLoS Genet 2:e88.
Ho Y, Gruhler A, Heilbut A, et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180-183.
Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. Nature 425:686-691.
Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. Revealing modular organization in the yeast transcriptional network. Nat Genet 31:370-377.
Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98:4569-4574.
Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41-42.
Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1.
Kitano H. 2002. Systems biology: a brief overview. Science 295:1662-1664.
Krogan NJ, Cagney G, Yu H, et al. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637-643.
Li S, Armstrong CM, Bertin N, et al. 2004. A map of the interactome network of the metazoan C. elegans. Science 303:540-543.
Lin N. 1999. Social networks and status attainment. Ann. Rev. Sociol. 25:467-487.

Lu H, Shi B, Wu G, et al. 2006. Integrated analysis of multiple data sources reveals modular structure of biological networks. Biochem Biophys Res Commun 345:302-309.
Maslov S, Sneppen K. 2002. Specificity and stability in topology of protein networks. Science 296:910-913.
Maslov S, Sneppen M, Zaliznyak A. 2004. Detection of topological patterns in complex networks: correlation profile of the internet. Physica A 333:529-540.
Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res 11:2120-2126.
Mazurie A, Bottani S, Vergassola M. 2005. An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6:R35.
Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple building blocks of complex networks. Science 298:824-827.
Newman ME. 2002. Assortative mixing in networks. Phys Rev Lett 89:208701.
Newman ME, Girvan M. 2004. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69:026113.
Newman MEJ. 2003. The structure and function of complex networks. SIAM Rev. 45:167-256.
Pereira-Leal JB, Enright AJ, Ouzounis CA. 2004. Detection of functional modules from protein interaction networks. Proteins 54:49-57.
Poyatos JF, Hurst LD. 2004. How biologically relevant are interaction-based modules in protein networks? Genome Biol 5:R93.
Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. 2002. Hierarchical organization of modularity in metabolic networks. Science 297:1551-1555.
Reguly T, Breitkreutz A, Boucher L, et al. 2006. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol 5:11.
Sacher M, Barrowman J, Schieltz D, Yates JR, 3rd, Ferro-Novick S. 2000. Identification and characterization of five new subunits of TRAPP. Eur J Cell Biol 79:71-80.
Salathe M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol 23:721-722.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449-451.
Sole RV, Fernandez P. 2003. Modularity “for free” in genome architecture? http://arxiv.org/abs/q-bio.GN/0312032.
Spirin V, Mirny LA. 2003. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A 100:12123-12128.
Steinmetz LM, Scharfe C, Deutschbauer AM, et al. 2002. Systematic screen for human disease genes in yeast. Nat Genet 31:400-404.
Tornow S, Mewes HW. 2003. Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res 31:6283-6289.
Uetz P, Giot L, Cagney G, et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623-627.
Valente AX, Cusick ME. 2006. Yeast Protein Interactome topology provides framework for coordinated-functionality. Nucleic Acids Res 34:2812-2819.

Vazquez A, Flammini A, Maritan A, Vespignani A. 2003. Modeling of protein interaction networks. ComPlexUs 1:38-44.
von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399-403.
Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol 19:1760-1768.
Wagner A. 2003. How the global structure of protein interaction networks evolves. Proc Biol Sci 270:457-466.
Wall ME, Hlavacek WS, Savageau MA. 2004. Design of gene circuits: lessons from bacteria. Nat Rev Genet 5:34-42.
Wuchty S, Oltvai ZN, Barabasi AL. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35:176-179.
Zhang C, Liu S, Zhou Y. 2006. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. J Proteome Res 5:801-807.
Zhang J. 2003. Evolution by gene duplication: an update. Trends. Ecol. Evol. 18:292-298.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

CHAPTER 3

Abundant Indispensable Redundancies in Cellular Metabolic Networks

3.1 ABSTRACT

Cellular life is a highly redundant complex system, yet the evolutionary maintenance of the redundancy remains unexplained. Using a systems biology approach, we infer that 37-47% of metabolic reactions in *E. coli* and yeast can be individually removed without blocking the production of any biomass component under any nutritional condition. However, the majority of these redundant reactions are preserved, because they have differential maximal efficiencies at different conditions or their loss causes an immediate fitness reduction that can only be regained via mutation, drift, and selection in evolution. The remaining redundancies are attributable to pleiotropic effects or recent horizontal gene transfers. We find that *E. coli* and yeast exhibit opposite relationships between the functional importance and redundancy level of a reaction, which is inconsistent with the conjecture that redundancies are preserved as an adaptation to backup important parts in the system. Interestingly, the opposite relationships can both be recapitulated by a simple model in which the natural environments of the organisms change frequently. Thus, adaptive backup is neither necessary nor sufficient to explain the high redundancy of cellular metabolic networks. Taken together, our results strongly suggest that redundant reactions are not kept as backups and that the genetic robustness of metabolic networks is an evolutionary byproduct.

3.2 INTRODUCTION

Functional redundancy refers to the situation where one part in a system can completely or partially compensate the loss of another (Hartman, Garvik, and Hartwell

2001; Wagner 2005a). Due to the existence of functionally redundant parts, a system may exhibit no or only mild phenotypic changes upon malfunction of a part. In biological systems, functional redundancy may occur at the component level (Wagner 2005a), exemplified by isoenzymes, which are generated by gene duplication and differ in protein sequence but catalyze the same biochemical reactions in an organism (Gu et al. 2003; Conant and Wagner 2004; DeLuna et al. 2008). Functional redundancy may also occur at the systems level, due to distributed properties of networks (Hartman, Garvik, and Hartwell 2001; Wagner 2005a). For example, glucose-6-phosphate dehydrogenase and D-ribulose-5-phosphate 3-epimerase catalyze distinct reactions and are located in alternative pentose phosphate pathways in yeast; simultaneous removal of the two enzymes is lethal, although individual removal of either enzyme is not (Harrison et al. 2007). While functional redundancy at the component level has been extensively studied in model organisms (Gu et al. 2003; Conant and Wagner 2004; Liang and Li 2007; Liao and Zhang 2007; Dean et al. 2008; DeLuna et al. 2008; Musso et al. 2008), redundancy at the systems level is poorly understood and thus is the focus of the present study.

An important consequence of functional redundancy is robustness against genetic perturbations such as deleterious mutations. Genetic robustness is a characteristic of cellular life, observed in all domains of life and at many levels of biological organizations, from DNA replication, transcription, and translation, to metabolism, cell cycle, and embryonic development (de Visser et al. 2003; Wagner 2005c; Lenski, Barrick, and Ofria 2006). Despite the apparent importance of functional redundancy and genetic robustness to development and health, their evolutionary preservation remains enigmatic (de Visser et al. 2003). This is because mutations that destroy redundancies occur repeatedly and are normally invisible to natural selection, such that redundancies are evolutionarily unstable and are destined to be lost (Clark 1994) except under special circumstances (Nowak et al. 1997). One hypothesis asserts that redundancies are favored by natural selection to ensure optimal performance through backing up important parts of a biological system when the system is attacked by mutations (de Visser et al. 2003), which we refer to as the adaptive backup hypothesis. But, this hypothesis requires a very high rate of deleterious mutation that is not commonly observed in cellular organisms (de Visser et al. 2003; Wagner 2005c).

To identify the evolutionary forces that preserve systems-level redundancies and to understand the origin of genetic robustness, we here take advantage of recent genomic reconstructions of metabolic networks of model organisms and analyze metabolic redundancies using rigorous systems-level flux balance analysis (FBA) (Price, Reed, and Palsson 2004) and its derivatives (Segre, Vitkup, and Church 2002). The metabolic networks are collections of biochemical reactions used to synthesize biomass, which is made up of multiple components such as amino acids and nucleotides. Based on the stoichiometric associations among metabolites, the computational methods provide reliable quantitative predictions of metabolic function and Darwinian fitness under genetic and environmental perturbations (Edwards, Ibarra, and Palsson 2001; Ibarra, Edwards, and Palsson 2002; Segre, Vitkup, and Church 2002; Famili et al. 2003; Papp, Pal, and Hurst 2004), thus allowing a systematic investigation of the amount of functional redundancy as well as the mechanisms of its maintenance in cellular metabolic networks.

3.3 MATERIALS AND METHODS

3.3.1 *E. coli* and yeast metabolic networks

Metabolic network models of *E. coli* (iJR904 GSM/GPR) (Reed et al. 2003) and yeast *S. cerevisiae* (iND 750) (Duarte, Herrgard, and Palsson 2004) were used in this study. The models were downloaded from the BiGG database (http://bigg.ucsd.edu) and parsed by the COBRA toolbox (Becker et al. 2007). The *E. coli* metabolic network contains 931 unique biochemical reactions, associated with 904 known genes. The yeast metabolic network is composed of 1149 reactions, associated with 750 known genes. Some reactions do not have associated genes because the genes whose protein products catalyze these reactions have yet to be identified. The metabolic network models also provide information of stoichiometry, direction of reaction, isoenzyme, and enzymatic protein complex. Classification of reactions by functional category as presented in Figure C.3 follows previous authors (Reed et al. 2003; Duarte, Herrgard, and Palsson 2004).

3.3.2 Flux balance analysis (FBA)

Details of FBA have been described in the literature (Edwards, Covert, and Palsson 2002; Price, Reed, and Palsson 2004). Briefly, FBA can be used to analyze a metabolic network at the steady state under the constraint of stoichiometry. The FBA equation is $S \cdot v = 0$, where S is the stoichiometric matrix and v is the metabolic flux vector. The biomass reaction describes the relative contribution of metabolites to the cellular biomass. The steady state flux distribution is determined by maximizing the rate of biomass production. The formulated linear programming problem is shown below:

Maximize object: $Z = \sum c_i \cdot v_i$

Subject to: $S \cdot v = 0$ and $\alpha \le v \le \beta$.

Here, the vector c is the biomass reaction function, and vectors α and β represent the lower- and upper-bound constraints of metabolic fluxes, respectively. We used the optimization package CLPEX (www.ilog.com) to solve the linear programming problem. To delete a reaction, we constrain the flux of the reaction to zero and obtain the maximal biomass production under the constraint. The relative fitness of the deletion strain to the wild-type strain is the maximal biomass production rate of the deletion strain divided by that of the wild-type (Segre et al. 2005).

3.3.3 Minimization of metabolic adjustment (MOMA)

MOMA has been previously described in detail (Segre, Vitkup, and Church 2002). Briefly, MOMA predicts the maximal biomass production rate upon deletion of a reaction by minimizing the differences in all metabolic fluxes between the deletion strain and the wild-type strain. All the constraints used in FBA are still enforced in MOMA. The formulated quadratic programming problem is

Minimize object: $D = \sum (v_i^{wt} - v_i)^2$

Subject to: $S \cdot v = 0$ and $\alpha \le v \le \beta$.

Here, v^{wt} is the wild-type flux vector calculated by FBA. When there are multiple flux values for a reaction in the wild-type, we randomly choose one of them, as in the original MOMA analysis (Segre, Vitkup, and Church 2002). MOMA results are not sensitive to the use of different wild-type values (Mahadevan and Schilling 2003). The quadratic

programming problem is also solved by CPLEX. As in FBA, deletion of a reaction is realized by constraining the flux of the reaction to zero.

3.3.4 Identification of dead-end reactions

We followed a published protocol (Burgard et al. 2004) to identify dead-end reactions. Dead-end reactions are defined as reactions that must have zero flux under a steady state. These reactions are involved in the generation of metabolites that are neither included in biomass nor transported outside the cell, and may reflect the incompleteness of the metabolic network models. They were identified by maximizing and minimizing in turn each flux under the condition that all nutrients are provided. If both the maximization and minimization result in zero flux, this reaction is considered a dead-end reaction. Because neither active transportation that requires ATP nor ionic transportation is modeled in FBA, these reactions are also not considered in our analysis. After removing all these reactions, the *E. coli* and *S. cerevisiae* metabolic networks used in our analysis contain 737 and 632 reactions, respectively. *S. cerevisiae* has much more dead-end reactions than *E. coli*, probably because the reconstructed metabolic network is less complete for the former than for the latter.

3.3.5 Simulation of nutritional conditions

The simulation of single-usable-carbon-source conditions follows a previous study (Pal, Papp, and Lercher 2005). Briefly, the medium contains one major carbon metabolite as the organic carbon source and the required inorganic metabolites (nitrogen, phosphate, metal ions, etc). In nature, the environments of microorganisms such as *E. coli* and *S. cerevisiae* often change frequently. These organisms usually face nutritionally poor conditions but occasionally encounter rich conditions. To mimic their natural environments, we simulate random nutritional conditions following a recent study (Wang and Zhang 2009). For each condition, we generate a random number g from an exponential distribution with a mean of $m = 0.1$. Here, g is the probability that a carbon-source nutrient is available. The actual presence or absence of each nutrient is then determined stochastically using g. We then add all required inorganic metabolites. Use of other m values (0.05 or 0.5) did not change our results. For each available nutrient, we

fix the uptake rate at a random value between 0 and 20. A condition is considered to be a valid condition only if FBA shows that it supports the growth of wild-type organisms.

3.3.6 Level of redundancy

We define the level of redundancy by the total number of redundant reactions divided by the average number of these reactions that need to be present in zero-redundancy networks. If there are n compensating pathways for a particular function and each of these pathways contains m reactions, the level of redundancy is $(mn)/m = n$. Thus, a redundancy of n is equivalent to the presence of n compensating pathways of equal length for each function.

3.3.7 Identification of reactions catalyzed by pleiotropic enzymes

We used the gene-reaction association annotated in the reconstructed metabolic networks (Reed et al. 2003; Duarte, Herrgard, and Palsson 2004). If an enzyme is annotated to catalyze more than one reaction, the enzyme is considered to be pleiotropic. An otherwise dispensable reaction appears to be indispensable if the enzyme that catalyzes this reaction is required to be present in the network, owing to its pleiotropic function in catalyzing an indispensable reaction.

3.3.8 Recent horizontal gene transfer (HGT)

We used the HGT dataset compiled in a recent study (Lercher and Pal 2008). In this dataset, HGTs were identified by the parsimony method across 31 proteobacterial species with the available phylogenetic tree. Because we concentrated on the metabolic genes of the *E. coli* K-12 strain, only genes within the dataset that are horizontally transferred into *E. coli* K-12 or its ancestors are used for the calculation. Genes that are horizontally transferred into the common ancestor of *E. coli* K-12 and *E. coli* CFT073 (within 4 steps away from *E. coli* K-12 in the phylogeny of Figure 1 in (Lercher and Pal 2008)) or a more recent ancestor of *E. coli* K-12 are considered as recent HGTs.

3.3.9 Comparison between experimentally determined and computationally predicted fitness values of single-gene-deletion yeast strains

The growth rates of *S. cerevisiae* single-gene-deletion strains in the YPD medium were previously measured (Steinmetz et al. 2002) and were obtained from http://www-deletion.stanford.edu/YDPM/YDPM_index.html. In the original dataset, the relative growth rate of every gene-deletion strain is normalized such that the average growth rate of all viable deletion strains is 1. In order to obtain the fitness of the deletion strains relative to the wild-type, we scaled the relative growth rate to another dataset (Sliwa and Korona 2005) which accurately measured the fitness of 12 gene-deletion strains by competing them individually with the wild-type strain. Specifically, we averaged the growth rate of the 12 gene deletion strains using the data from (Steinmetz et al. 2002) (f_g=1.026) and averaged their fitness relative to the wild-type using the data from (Sliwa and Korona 2005) (f_m=1.010). Then, for every deletion strain in the large dataset (Steinmetz et al. 2002), its fitness is calculated by multiplying the growth rate by f_m/f_g. In our analysis, f_m/f_g is 0.984. Our results are virtually unaffected even when we use f_m/f_g=1.

The parameters used in FBA and MOMA to mimic the YPD medium that is used in the experimental determination of the growth rates of yeast gene deletion strains follow previous authors (Forster et al. 2003). Comparison between the FBA predicted and experimentally determined fitness values shows that only in 8% (39/486) of cases, essential genes are misidentified as nonessential by FBA. However, a reaction is considered redundant only when it is nonessential in all 10^5 examined conditions. The probability of misclassifying a non-redundant reaction as redundant is the probability that FBA misclassifies it as nonessential in every condition where it is essential, which should be low. Thus, it is improbable for a non-redundant reaction to be misclassified as redundant. We regarded a redundant reaction as indispensable if its deletion strain has an FBA or MOMA predicted fitness of f<0.99. In fewer than 2% (8/486) of cases, FBA predicted fitness is <0.99 while the experimentally determined fitness is >0.99. The corresponding error rate for MOMA is slighter higher (24/486=5%). A dispensable redundant reaction is misclassified by FBA or MOMA as indispensable when the true fitness of the deletion strain is >0.99 in all 10^5 conditions but the predicted fitness is <0.99 in at least one condition, which is expected to be low.

3.3.10 Preservation of rarely used genes

A rarely used gene can be preserved in the genome during evolution as long as the null mutation rate is sufficiently low, compared to the product of the probability that the gene is used at any given time and the fitness contribution of the gene when it is used. This is demonstrated below, first in haploid organisms and then in diploids. Let us use *A* to collectively denote all functional alleles of the gene under study and *a* to collectively denote all null alleles of the gene, and use *p* and *q* to denote the frequency of *A* and *a* alleles, respectively. Let the null mutation rate, or the rate of mutation from *A* to *a*, be *u* per gene per generation. We assume that the mutation rate from *a* to *A* is zero because it is extremely unlikely for a null allele to mutate back to a functional allele. Random mutations increase the frequency of *a*, while occasional natural selection reduces it. Let us first consider the possibility of a mutation-selection balance. At the balance, new *a* alleles generated by mutations are completely removed by selection. In haploids, let us assume that the relative fitness of *A* and *a* individuals be 1 and 1-*s*, respectively, and that selection occurs once every *n* generations. For simplicity, let us assume that in every cycle of *n* generations, selection occurs at the end of the *n*th generation in the form of a viability difference. Thus, when the balance is reached, in *n* generations, the allele frequency of *a* increases from q_0 to q_n by mutation, and then decreases to q_0 by natural selection. The mutational process is described by the difference equation

$$q_n = q_{n-1} + (1 - q_{n-1})u \,. \tag{3.1}$$

Solving Equation 3.1, we obtained

$$\begin{aligned} q_n &= (q_0 - 1)(1 - u)^n + 1 \\ &\approx (q_0 - 1)(1 - un) + 1 \\ &= q_0 + (1 - q_0)un \end{aligned} \,. \tag{3.2}$$

In the case of haploid organisms such as *E. coli*, the selection process is described by

$$q_n' = \frac{q_n(1-s)}{(1-q_n) + q_n(1-s)} = \frac{q_n(1-s)}{1 - q_n s} \quad , \tag{3.3}$$

where q_n' is the frequency of *a* after selection. At the mutation-selection balance, we have

$$q_n' = q_0 \,. \tag{3.4}$$

Using Equations 3.2, 3.3, and 3.4, we can obtain

$$q_n = un / s . \tag{3.5}$$

For diploid organisms, the fitness of *AA*, *Aa*, and *aa* individuals are assumed to be 1, 1, and 1-*s*, respectively, because enzyme genes are largely haplosufficient (Kondrashov and Koonin 2004; Deutschbauer et al. 2005). Then, Eqation 3.3 can be rewritten as

$$\begin{aligned} q_n' &= \frac{2p_n q_n + 2q_n^2(1-s)}{2[p_n^2 + 2p_n q_n + q_n^2(1-s)]} \\ &= \frac{q_n(1-q_n) + q_n^2(1-s)}{1-q_n^2 s} \\ &= \frac{q_n(1-q_n s)}{1-q_n^2 s} \end{aligned} . \tag{3.3'}$$

Using Euqations 3.2, 3.3', and 3.4, we obtain

$$q_n = \sqrt{un / s} . \tag{3.5'}$$

Thus, for both haploids and diploids, when $un/s < 1$, null alleles cannot be fixed through the mutation-selection process. In other words, functional alleles can be preserved in the population. Note that the above mutation-selection equilibrium is a stable equilibrium, because if q is by chance slightly larger than its equilibrium value, the effect of selection in removing null alleles (qs for haploids and q^2s for diploids) becomes larger and the mutation rate per generation in generating null alleles ($(1-q)u$) becomes lower. Consequently, q will return to its equilibrium value. The same argument can be made if q is by chance slightly smaller than its equilibrium value. Thus, random genetic drift cannot push q much away from its equilibrium value. This is particularly so, given the large population size of *E. coli* and *S. cerevisiae*.

Although $un/s < 1$ can ensure that functional alleles at a locus will not be lost in evolution, in practice, one may consider a more stringent criterion of $q_n < 0.5$ so that a randomly sampled allele of the gene from the population is more likely to be functional than null. Thus, we consider that the gene can be retained by selection if $n < 0.5s/u$ for haploids or $n < 0.25s/u$ for diploids. The mean mutation rate u for *E. coli* metabolic enzyme genes is 7.7×10^{-8} per gene per generation (see the next section). If we use $s = 0.01$, n has to be smaller than 6.5×10^{4}. If we use $s = 0.1$, n has to be smaller than 6.5×10^{5}. The mean u for *S. cerevisiae* metabolic enzyme genes is 4.0×10^{-8} per gene per

generation (see the next section). When $s = 0.01$, n has to be smaller than 6.3×10^4. When $s = 0.1$, n has to be smaller than 6.3×10^5. Because the vast majority of metabolic enzyme genes have u values not exceeding twice the mean u (Figure C.4), the above results are also largely correct for virtually every individual gene. In our study, although a redundant reaction is considered indispensable when it is used in at least one of the 10^5 examined nutritional conditions, the vast majority of redundant reactions were found to be used in the first 10^4 conditions examined (Figure 3.1). In addition, although our main analysis used a fitness differential of $s = 0.01$ to define indispensability, additional analysis showed that the results are virtually unchanged even when we require $s = 0.1$. Taken together, these considerations and the population genetic analysis in this section demonstrate that the criteria we used in determining indispensability of redundant reactions are appropriate.

Note that the above population genetic formulation ignores occasional back mutations from *a* to *A*. When the frequency of *a* increases, back mutations may become more common. Because back mutations effectively reduce u, the above results are conservative. In other words, the power of selection in preserving rarely used reactions should be slightly higher than calculated above.

3.3.11 Null mutation rate in *E. coli* and *S. cerevisiae*

If we know the point mutation rate and indel mutation rate, we can estimate the null mutation rate for any given gene (Zhang and Webb 2003). This is based on the idea that nonsense mutations and frame-shifting mutations usually generate null alleles. Note that some nonsense and frame-shifting mutations may not generate null alleles if they only disrupt a short C-terminal region of the encoded protein, while some missense mutations can generate null alleles. Our estimates are thus approximate, but they are not expected to differ from the true values by more than one fold, because the above three types of mutations are rare and their opposite influences on our estimates tend to cancel out. We consider an indel as frame-shifting if its size is not multiples of three nucleotides. The point mutation rates used here are 5.4×10^{-10} and 2.2×10^{-10} per nucleotide site per generation for *E. coli* and *S. cerevisiae*, respectively (Drake et al. 1998). We assume that the indel mutation rate is 10% of the point mutation rate and that

83% of indel mutations are frame-shifting, based on previous comparative genomic analysis (Podlaha and Zhang 2003; Zhang and Webb 2003). We then use the coding sequences of the genes associated with *E. coli* and *S. cerevisiae* metabolic reactions studied in this work to estimate the null mutation rates per gene per generation, by a modified version of the program PSEUDOGENE (Zhang and Webb 2003). The frequency distributions of u for *E. coli* and *S. cerevisiae* metabolic enzyme genes are shown in Figure C.4. Given the null mutation rate u, it is easy to see that it takes on the order of $1/u$ (or 10^7) generations for a functional allele to be replaced with null alleles.

3.4 RESULTS

3.4.1 Abundance of redundant metabolic reactions

Here we study the bacterium *Escherichia coli* (Reed et al. 2003) and yeast *Saccharomyces cerevisiae* (Duarte, Herrgard, and Palsson 2004) because their reconstructed metabolic networks are of high quality and have been empirically verified and because they represent prokaryotes and eukaryotes, respectively. The metabolic networks of *E. coli* and *S. cerevisiae* contain 737 and 632 biochemical reactions, respectively, after the removal of dead-end reactions (see Materials and Methods). For three reasons, we focus on biochemical reactions rather than genes encoding the enzymes that catalyze these reactions. First, we are interested in the functional redundancy at the systems level of a metabolic network, which is composed of reactions. Second, there is no one-to-one relationship between genes and reactions. Third, annotations of enzyme genes are incomplete, making it impossible to conduct a gene-based analysis that is as comprehensive and accurate as a reaction-based analysis.

Assuming a steady state in metabolism, flux balance analysis (FBA) maximizes the rate of biomass production under the stoichiometric matrix of all metabolic reactions and a set of flux constraints (see Materials and Methods). The FBA-optimized rate of biomass production can be regarded as the Darwinian fitness of the cell under the condition specified. If removing a reaction blocks the production of one or more biomass components, biomass production becomes zero or undefined due to imbalanced compositional stoichiometry of the biomass. In order to estimate the number (m) of

metabolically redundant reactions, we need to identify the reactions whose single removal does not block the production of any biomass component under any nutritional condition. Because it is infeasible to enumerate all possible conditions, we investigate how the estimate of m changes when the number (c) of examined conditions increases. In *E. coli*, m reduces from 737 to 320 after we examine all single-usable-carbon-source conditions (see Materials and Methods) (Figure 3.1A). We then create random nutritional conditions in which wide-type organisms can grow (see Materials and Methods). As expected, m decreases at a much reduced pace as c increases (Figure 3.1A). When c is doubled from 5×10^4 to 10^5, m decreases by only 6 (or 2%). Thus, 10^5 conditions appear sufficient for providing a reasonably accurate estimate of m. Using this method, we identified 276 (37% of the network) and 295 (47% of the network) redundant reactions from *E. coli* and *S. cerevisiae*, respectively (Figure 3.1 and Figure 3.2).

Non-redundant metabolic reactions can be divided into two groups: always-essential and sometimes-essential. Deletion of an always-essential reaction blocks biomass production under all conditions, whereas deletion of a sometimes-essential reaction blocks biomass production under some but not all conditions. Always-essential reactions can be identified unambiguously, because the metabolic network models allow us to know all nutrients that can be used by the cells under the metabolic models. If a reaction is essential when all these usable nutrients are available, it must be essential when one or more of these nutrients are absent and hence must be an always-essential reaction. The rest of the non-redundant reactions are then classified as sometimes-essential reactions. Using this strategy, we identified 95 (13%) always-essential and 366 (50%) sometimes-essential reactions in *E. coli* (Figure 3.2A), and 24 (4%) always-essential and 313 (49%) sometimes-essential reactions in *S. cerevisiae* (Figure 3.2B). Not unexpectedly, sometimes-essential reactions considerably outnumber always-essential reactions (Almaas, Oltvai, and Barabasi 2005). We also observe that (i) different functional subgraphs of the metabolic network contain different fractions of always-essential, sometimes-essential, and redundant reactions and (ii) *E. coli* and *S. cerevisiae* show different distributions of the three types of reactions among subgraphs (Figure C.3).

3.4.2 Zero-redundancy metabolic networks

While redundant reactions can be individually removed from a metabolic network without blocking biomass production, they may not be simultaneously removed. To estimate the number of redundant reactions that can be simultaneously removed, we build a functional metabolic network with zero redundancy under all conditions. To achieve this goal, we randomly pick a redundant reaction and examine if its deletion still permits biomass production in all conditions examined. If so, this reaction is permanently deleted from the network; otherwise, it is restored. We then pick another redundant reaction from the remaining network and repeat the procedure until no more reactions can be deleted from the network. We generate 250 such networks by using variable random orders in deleting reactions, and find that the zero-redundancy networks have on average 534 (72% of the original network) and 418 (64%) reactions in *E. coli* and *S. cerevisiae*, respectively (Figure C.5). Because on average 203 of the 276 redundant reactions can be simultaneously deleted in *E. coli*, the level of redundancy is 276/(276-203) = 3.8. Simply put, this level of redundancy among the redundant reactions of *E. coli* is equivalent to the presence of 3.8 compensating pathways of equal length for each function (see Materials and Methods). The corresponding number is 3.7 in *S. cerevisiae*. Due to high demands of computational time and memory, our main analysis of zero-redundancy networks examines only 10^3 nutritional conditions. Nevertheless, our subsequent analysis with 10^4 conditions shows that the result is largely unchanged (Figure C.5). In sum, the zero-redundancy network analysis further demonstrates the high redundancy of the *E. coli* and *S. cerevisiae* metabolic networks, because as many as 28%-36% of reactions can be simultaneously removed from the metabolic networks without blocking the biomass production under any condition.

3.4.3 Preservation of redundant reactions: Efficient reactions

How can redundant reactions be preserved in a metabolic network during evolution? One possibility is that these functionally redundant reactions have differential metabolic efficiencies under different conditions, allowing the cell to use different reactions to achieve maximal growth in many different environments. Under this hypothesis, deleting a redundant reaction at a given condition may reduce (but not block)

biomass production when the deleted reaction is more efficient than other reactions of the same function at this condition. Population genetic theories predict that mutations causing a fitness reduction of $>1/N$ will be subject to substantial purifying selection, where N is the effective population size, on the order of 10^9 for *E. coli* (Ochman and Wilson 1987) and 10^7 for *S. cerevisiae* (Wagner 2005b). Thus, natural selection can keep a redundant reaction in the network if its deletion renders even a tiny decrease in biomass production. Furthermore, because the null mutation rate in *E. coli* and *S. cerevisiae* is on the order of 10^{-7} to 10^{-8} per gene per generation, a gene can be selectively kept in a population even if it is used only once every 10^4 to 10^5 generations (see Materials and Methods). Given these theoretical considerations and potential errors associated with FBA-predicted fitness, we regard a redundant reaction to be indispensable if its removal reduces biomass production by more than 1% in one or more of the 10^5 conditions examined. Such indispensable redundant reactions are referred to as efficient reactions, as they are more efficient than other reactions of the same functions under at least one condition. Our analysis identifies 64 and 89 efficient reactions in *E. coli* and *S. cerevisiae*, respectively, accounting for 23-30% of all redundant reactions (Figure 3.1 and Figure 3.2). The remaining 70-77% of redundant reactions are as efficient as or less efficient than other reactions of the same functions under all conditions and are referred to as non-efficient reactions (Figure 3.1 and Figure 3.2).

3.4.4 Preservation of redundant reactions: Non-efficient, active reactions

In the above analysis, we assumed that when a redundant reaction is deleted, its compensating reaction is immediately activated to its optimal flux to produce the maximal biomass predicted by FBA. This assumption requires that the cell has a regulatory emergency plan for every possible reaction deletion, which seems unrealistic. In general, the growth performance of a perturbed metabolic network is suboptimal and the FBA-predicted maximal growth can only be achieved through evolution by mutation, drift, and selection (Ibarra, Edwards, and Palsson 2002; Fong et al. 2005). In other words, when a reaction is deleted from a cell, the cell may be outcompeted by wild-type cells and has no chance to evolve to its FBA-predicted maximal fitness. To consider this possibility, we employ the method of minimization of metabolic adjustment (MOMA), a

derivative of FBA that has also been empirically verified (Segre, Vitkup, and Church 2002). Under all the assumptions and constraints used by FBA, MOMA calculates the rate of biomass production after the deletion of a reaction by minimizing flux changes (see Materials and Methods). Because MOMA minimizes flux changes while FBA does not, the biomass production predicted by MOMA is always lower than or equal to that predicted by FBA. A non-efficient reaction is considered to be indispensable if its removal reduces the MOMA-predicted biomass production by more than 1% in one or more of the 10^5 examined conditions. Such reactions are referred to as active reactions because they must have non-zero fluxes; otherwise their removal will not cause biomass reductions. We identify 158 and 166 active reactions in *E. coli* and *S. cerevisiae*, respectively, accounting for more than half of all redundant reactions or 75-80% of non-efficient redundant reactions (Figure 3.2). The rest of non-efficient reactions are referred to as non-active reactions because their removal does not affect MOMA-predicted biomass appreciably. Unlike the numbers of redundant and non-efficient reactions, the number of non-active reactions may have been overestimated, as the number continues to drop even when c reaches 10^5 conditions (Figure 3.1). This means that the above numbers of active reactions are conservative estimates.

Although we showed how a non-efficient redundant reaction can be indispensable and kept in the network by natural selection, it is puzzling as why such reactions were incorporated into the metabolic network in the first place, as non-efficient reactions are never more efficient than other reactions of the same functions. We suggest that non-efficient reactions were incorporated by neutral processes. They became active reactions if they were equally efficient as their redundant reactions under some conditions. When multiple equally efficient redundant reactions exist, (regulatory or structural) degenerate mutations may be fixed so that the total activity of the enzymes catalyzing the redundant reactions is optimized while the activity of each enzyme becomes insufficient for the maximal growth should the other redundant enzymes be removed.

3.4.5 Preservation of redundant reactions: Non-efficient, non-active reactions

Our analysis identified 54 (7% of the total network) and 40 (7%) non-active redundant reactions in *E. coli* and *S. cerevisiae*, respectively (Figure 3.2). Among them, 38 *E. coli* and 20 *S. cerevisiae* reactions are less efficient than other reactions of the same functions and have zero fluxes under all conditions. The rest may be as efficient as their redundant reactions and have non-zero fluxes, but their removal does not reduce MOMA-predicted biomass production by more than 1%.

How are the non-active reactions maintained in the metabolic network? Some enzymes can catalyze multiple reactions, a phenomenon known as pleiotropy (He and Zhang 2006). In *E. coli*, 266 reactions (36% of the total network), including 27 non-active reactions, are catalyzed by pleiotropic enzymes. In *S. cerevisiae*, 171 reactions (27% of the total network), including 13 non-active reactions, are catalyzed by pleiotropic enzymes. A non-active reaction can be stably retained in the network if the enzyme that catalyzes it also catalyzes one or more indispensable reactions. Indeed, we find that every non-active reaction catalyzed by pleiotropic enzymes can be retained by this "guilt-by-association" mechanism. In both *E. coli* and *S. cerevisiae*, there are only 27 redundant reactions whose retentions are unexplained (Figure 3.2). Further examinations show that they are unexplained by FBA and MOMA simply because of the incompleteness of the reconstructed metabolic networks, limitations of the metabolic models (e.g., lack of connection to regulatory and signal transduction networks), and existence of environments difficult to simulate (e.g. temperature changes). For instance, *E. coli* gene *otsB* encodes trehalose-6-phosphate phosphatase, which is required for cell viability at 4°C (Kandror, DeLeon, and Goldberg 2002) and thus may be maintained by selection if *E. coli* sometimes experiences this low temperature in nature. We also observed 6 *E. coli* non-active reactions that are catalyzed by enzymes encoded by genes that were recently horizontally transferred into *E. coli* (see Materials and Methods). Horizontal gene transfers occur so frequently among prokaryotes (Gogarten, Doolittle, and Lawrence 2002) that the presence of some redundant genes may be attributable to this mechanism rather than preservation under purifying selection. Indeed, analyzing an *E. coli* horizontal-gene-transfer dataset (Lercher and Pal 2008), we find that the fraction

of recently horizontally acquired genes is significantly greater among non-active reactions (43%) than among other reactions (19%) ($P < 0.05$, Fisher's exact test; see Materials and Methods). We did not analyze recent horizontal gene transfers into *S. cerevisiae*, because such information is not readily available and because horizontal gene transfers are thought to be much less frequent in eukaryotes than in prokaryotes. After considering all these additional mechanisms, there are only 14 (8 with associated genes) *E. coli* and 13 (4 with associated genes) *S. cerevisiae* redundant reactions whose preservation in the metabolic networks remain unexplained (Figure 3.2).

3.4.6 A direct test of the adaptive backup hypothesis

Our analysis showed that the vast majority of the functionally redundant reactions in *E. coli* and *S. cerevisiae* are selectively maintained because they cause fitness reductions when singly removed from the cell. This explanation is different from the adaptive backup hypothesis, in which only simultaneous removal of compensating redundant reactions is deleterious. Hence, the adaptive backup hypothesis is not needed to explain the maintenance of metabolic redundancy. The adaptive backup hypothesis also has a key prediction of higher redundancy for more important functions, because the fitness gain from backing up more important functions is greater than that from backing up less important ones (e.g., (Kafri et al. 2008) on redundant duplicate genes). To test this prediction, we measure the importance of reactions using zero-redundancy networks, because they are free from the confounding influence of redundant reactions. We calculate the average biomass reduction upon removal of a reaction from a zero-redundancy network across 10^3 conditions and repeat this calculation in 125 random zero-redundancy networks to obtain the mean. For *E. coli*, contrary to the prediction of the backup hypothesis, reactions that are redundant in the original metabolic network tend to perform less important jobs than reactions that are non-redundant ($P = 0.056$, Mann-Whitney U test; Figure 3.3). But for *S. cerevisiae*, the observation appears to be consistent with the backup prediction ($P = 3.5\times10^{-5}$, Mann-Whitney U test; Figure 3.3). These opposite patterns in *E. coli* and *S. cerevisiae* show that the adaptive backup hypothesis is either inadequate or wrong. Our subsequent computer simulation shows that the observations in both species are explainable without invoking adaptive backup.

Let us assume that a population rotates among many different environments so frequently that a metabolic reaction needed for a particular environment does not have chance to be lost from the population before the population switches back to this environment after moving through other environments. This scenario is quite likely because even without selective constraint it takes on average 10^7 generations for a functional allele of a gene to be replaced by a nonfunctional allele in *E. coli* and *S. cerevisiae* (see Materials and Methods). We ask whether this scenario would result in a metabolic network whose redundancy mimics that observed in *E. coli* and *S. cerevisiae*. We first generate a random nutritional condition (see Materials and Methods). A zero-redundancy metabolic network for this condition is then generated by removing redundant reactions from the original network, as described earlier. We repeat this process 10^3 times, each under a different condition. We then merge the 10^3 resultant zero-redundancy networks to form the final simulated metabolic network. We measure the relative importance of redundant and non-redundant reactions of this simulated network as was done for the real network. Interestingly, for both *E. coli* and *S. cerevisiae*, the results are similar between the simulated networks and their respective real networks (Figure 3.3). Because we did not invoke adaptive backup in the simulation, our result strongly suggests that the observation of higher redundancy for more important functions in *S. cerevisiae* is a byproduct of its evolutionary history. This simulation was repeated 10 more times and the above finding always holds (Table C.2).

3.5 DISCUSSION

In this work, we used FBA and MOMA to study the level of redundancy as well as the mechanisms of its preservation in metabolic networks. It is important to emphasize that in our definition, a reaction is functionally redundant only when its removal does not block the biomass production in any condition. This definition is fundamentally different from that used in all earlier studies of metabolic redundancy. In these studies, a reaction is considered redundant when it is dispensable in only one or a few conditions. It is clear that many redundant reactions such defined are actually sometimes-essential reactions by our definition and thus are not truly redundant in all

conditions, as has been demonstrated in numerous studies (Papp, Pal, and Hurst 2004; Blank, Kuepfer, and Sauer 2005; Dudley et al. 2005; Harrison et al. 2007). It is the number and preservation of those truly functionally redundant reactions that are of particular interest and are the subject of our study. However, to our surprise, the percentage of reactions that are redundant under our strict definition substantially exceeds earlier FBA-based and experiment-based estimates that were based on the looser definitions of redundancy (Papp, Pal, and Hurst 2004; Blank, Kuepfer, and Sauer 2005). A careful comparison reveals that these earlier studies only considered reactions with non-zero fluxes under rich or minimal media, which tend to be non-redundant (Table C.3), while our study considers all reactions in a metabolic network. Our finding of large numbers of redundant reactions in *E. coli* and *S. cerevisiae* confirms the prediction of high metabolic redundancy from extreme pathway analysis of subnetworks, which cannot deal with an entire cellular metabolic network due to computational difficulties (Papin et al. 2002; Price, Papin, and Palsson 2002).

Because FBA and MOMA predictions are not without errors, it is important to evaluate how such errors affect our results. Previous studies demonstrated that FBA and MOMA make good qualitative predictions of gene essentiality (Segre, Vitkup, and Church 2002; Papp, Pal, and Hurst 2004). Here we plot the experimentally determined fitness values of single-gene-deletion *S. cerevisiae* strains in rich media and their corresponding values predicted by FBA and MOMA (Figure 3.4). Only in 8% (39/486) of cases did we observe misidentification of essential genes as nonessential by FBA (yellow bars in Figure 3.4A). The accuracy of FBA should be similarly high in other conditions, because the models and assumptions used in FBA are not specific to the rich media. Because a reaction is regarded as redundant only when it is nonessential in all 10^5 examined conditions, it is improbable for a non-redundant reaction to be misclassified as redundant (see Materials and Methods). In other words, the number of redundant reactions is unlikely to have been grossly overestimated in our study. We regarded a redundant reaction as indispensable if its deletion strain has an FBA or MOMA predicted fitness of $f<0.99$. In fewer than 2% (8/486; red bars in Figure 3.4A) of cases, the FBA predicted fitness is <0.99 while the experimentally determined fitness is >0.99. The corresponding error rate for MOMA is slighter higher (24/486=5%; red bars in Figure

3.4B). These observations suggest that the estimation of the number of indispensable redundant reactions is relatively accurate (see Materials and Methods). To be conservative, we also used a more stringent cutoff of $f<0.9$ in defining indispensable reactions, for which the error rate is expected to be 0.6% (3/486) for FBA and 1.2% (6/486) for MOMA, respectively. We found that the number of unexplained redundant reactions is not much increased (Table C.4), suggesting that our conclusion remains valid even when the few errors made by FBA and MOMA are considered. To explore a wider range of f, we further carried out an analysis with a cutoff of $f<0.999$, which is closer to the fitness boundary between deleterious and neutral mutations ($1/N$), and found that the results are consistent with those under other cutoffs (Table C.4). An even larger f is theoretically preferred but it may generate less accurate results due to the limited precisions of FBA and MOMA. Use of a larger f should result in more redundant reactions to be indispensable and thus provide stronger support to our conclusion. Therefore, our conclusion of indispensability of redundant reactions is robust to the choice of f.

In this study, we examined 10^5 different nutritional conditions to identify redundant reactions and to determine the mechanisms of evolutionary preservation of the redundant reactions. Although we could have examined more combinations of nutrients (e.g., additional nitrogen and phosphate sources), our results showed that even with the limited number and type of nutritional conditions considered, the preservation of virtually every redundant reaction can be explained without invoking backup. Furthermore, some redundant enzymes may perform non-catalytic functions that are not considered in FBA (He and Zhang 2006). Thus, our results are conservative. It is worth mentioning that the above conclusion is strongly supported by a recent experimental study by Hillenmeyer and colleagues, who showed that most yeast genes have fitness effects in at least one of many conditions examined (Hillenmeyer et al. 2008). However, Hillenmeyer *et al.* did not study functional redundancy and their results may be misinterpreted as a complete lack of functional redundancy in yeast. More importantly, the majority of the conditions they used are artificial drug treatments that are likely to be substantially different from the natural environments of yeast (with the exception of some clinical strains of yeast).

The conditions we computationally examined are combinations of existing nutritional metabolites and thus are more realistic.

The adaptive backup hypothesis makes the key prediction of higher redundancy of more important function. Using zero-redundancy networks, we showed that *E. coli* and *S. cerevisiae* exhibit different relationships between the importance of a reaction and its redundancy level. Thus, the adaptive backup hypothesis is inadequate in explaining the actual observations. By contrast, our simulation showed that a simple scenario where a population frequently alternates among many environments can explain both the *E. coli* and *S. cerevisiae* results and thus provides a better explanation of the observations. One caveat of the simulation and other analysis in the present study is that it is unknown how close our simulated conditions match and how well they represent the wide range of natural nutritional environments in the evolution of *E. coli* and *S. cerevisiae*. Nevertheless, identification of conditions where some redundant reactions directly contribute to the organismal fitness suggests that these reactions *can* be maintained without invoking the adaptive backup hypothesis. Furthermore, these identified nutritional conditions may provide information about the natural environments where the organisms live or have recently lived, which are potentially useful for the study of organismal evolution as well as environmental changes.

In summary, our systems analysis of *E .coli* and *S. cerevisiae* metabolic networks revealed the presence of 37-47% redundant reactions. The vast majority of these redundancies are stably preserved in the network owing to their direct contribution to fitness or pleiotropic effect of some enzymes. In the case of *E. coli*, a few redundant reactions were recently acquired via horizontal gene transfers and thus may not be stably maintained in the genome. It is likely that even the small fraction of redundant reactions that are unexplained by our analysis can be explained when more information about them become available. Furthermore, we invalided a key prediction of the adaptive backup hypothesis about the relationship between the functional importance and redundancy level of a reaction. Taken together, adaptive backup is neither necessary nor sufficient to explain the high redundancy of cellular metabolic networks. Thus, the genetic robustness of metabolism is likely an evolutionary byproduct. In this context, genetic robustness does not constrain evolvability (Lenski, Barrick, and Ofria 2006), but rather enhances it,

because genetic robustness reflects the ability of an organism to survive in different environments. Note that although our analysis is limited to the metabolic network, the obtained biological principles and insights may be applicable to redundancies in other biological systems, because all biological systems can be treated as complex networks.

3.6 ACKNOWLEDGEMENTS

We thank Meg Bakewell, Wendy Grus, Xionglei He, Ben-Yang Liao, and Wenfeng Qian for valuable comments. This work was supported by research grants from National Institutes of Health and University of Michigan Center for Computational Medicine and Biology to J.Z.

Figure 3.1 Estimates of the numbers of various redundant reactions in *E. coli* and *S. cerevisiae* stabilize as the number of examined nutritional conditions increases. The first 158 conditions examined in (**A**) *E. coli* and first 60 conditions examined in (**B**) *S. cerevisiae* are single-usable-carbon-source conditions, whereas the remaining conditions are randomly generated following a specific sampling scheme. Note that the number of non-active reactions might be overestimated, because the estimate continues to decline as the number of examined conditions increases. This leads to a conservative estimate of the number of active reactions.

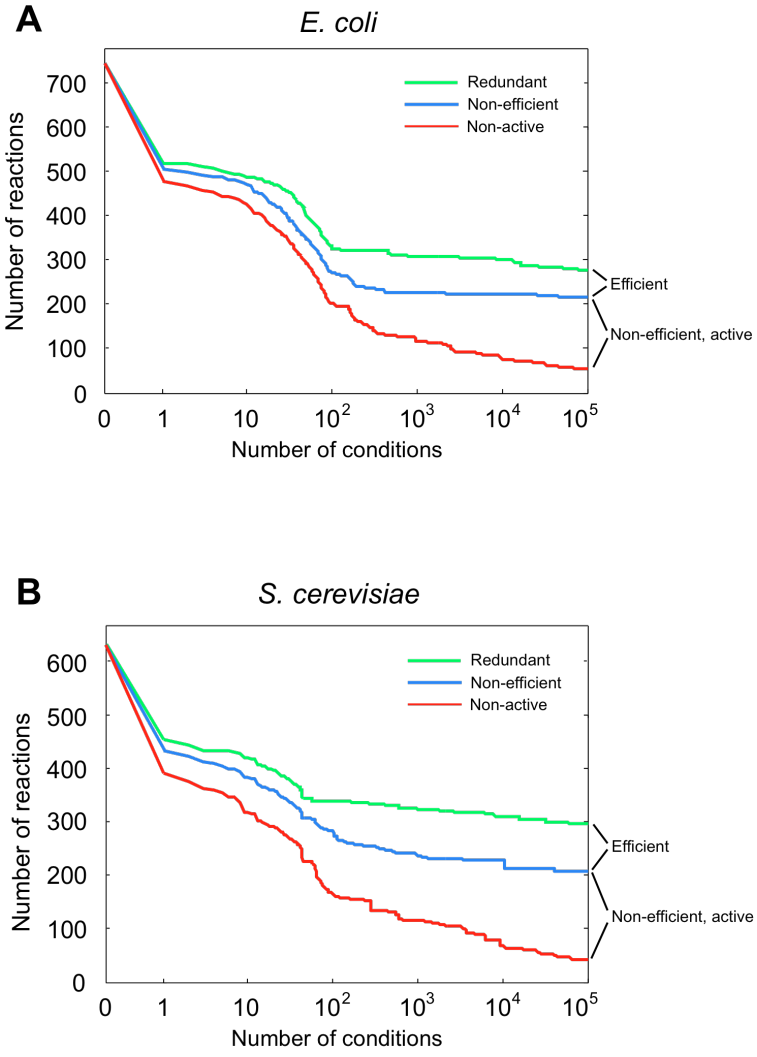


Figure 3.2 Numbers and fractions of redundant and non-redundant reactions in *E. coli* and *S. cerevisiae* metabolic networks. (**A**) *E. coli* and (**B**) *S. cerevisiae*. The total number of reactions after the removal of dead-end reactions is given in the parentheses after the species name. Because the enzyme genes associated with some reactions have yet to be identified, the number of genes known to be associated with the unexplained redundant reactions is given in brackets. For each species, the middle and right circles show various explanations for the existence of redundant reactions. Explanations in the middle circle are considered before those in the right circle; within each circle, explanations depicted with darker colors are considered before those depicted with lighter colors. For each redundant reaction, only the first applicable explanation considered is counted.

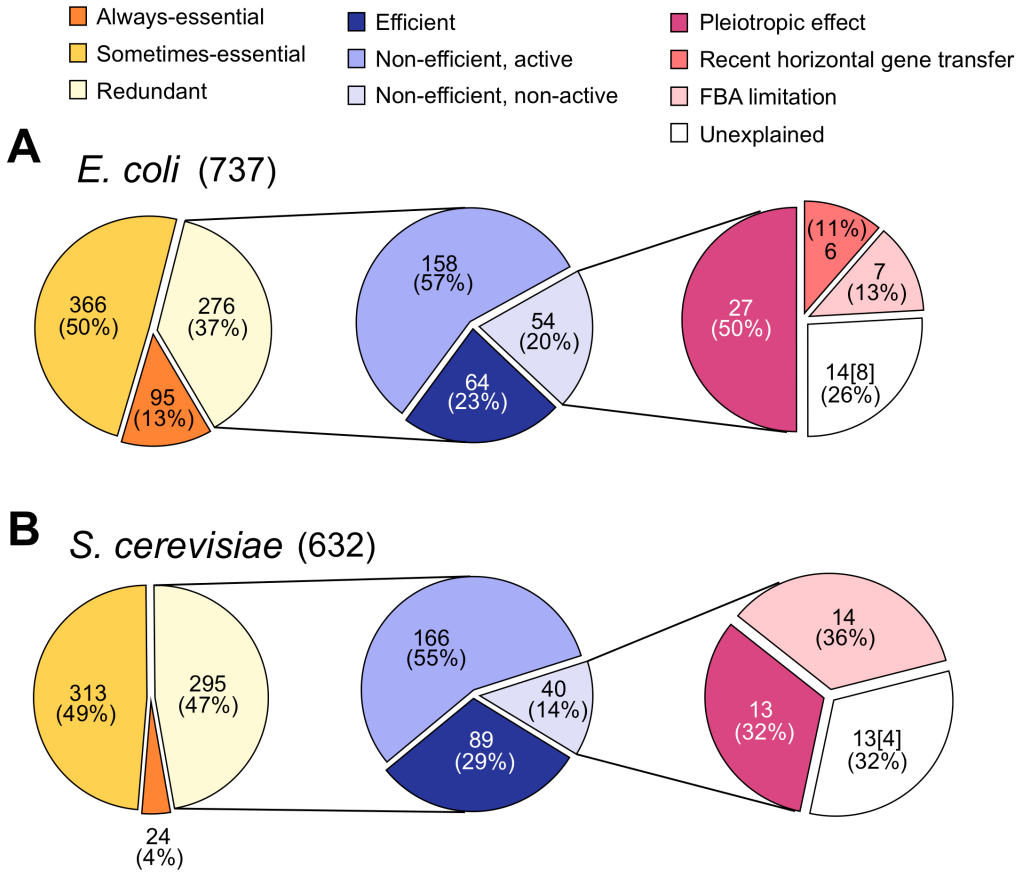


Figure 3.3 Relationships between the importance and redundancy of metabolic reactions. Error bars show one standard error. *P*-values are from Mann-Whitney U test. The redundancy of a reaction is determined from the complete network, whereas the importance of a reaction is determined from zero-redundancy networks. Redundant reactions perform less important functions than non-redundant functions in *E. coli*, whereas the opposite is true in *S. cerevisiae*. The same patterns are recapitulated in simulated metabolic networks that are formed by merging 10^3 zero-redundancy networks that each functions in a different condition.

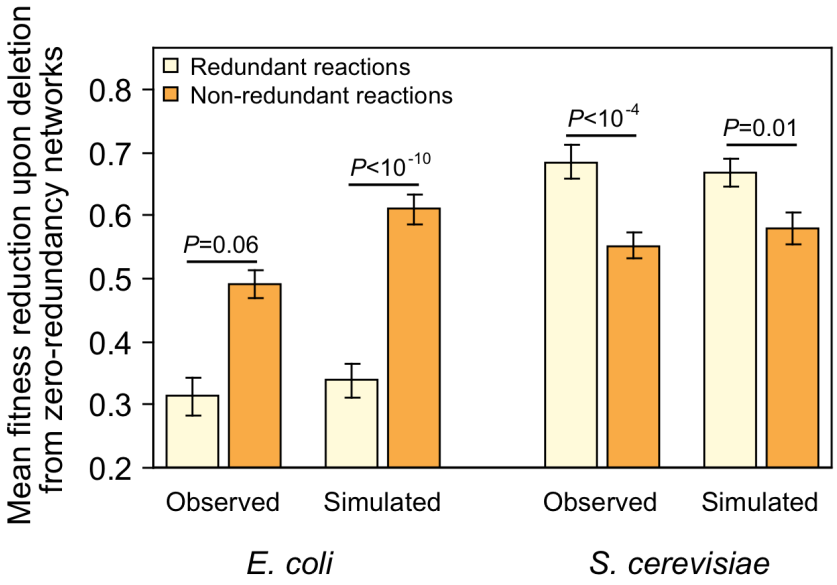
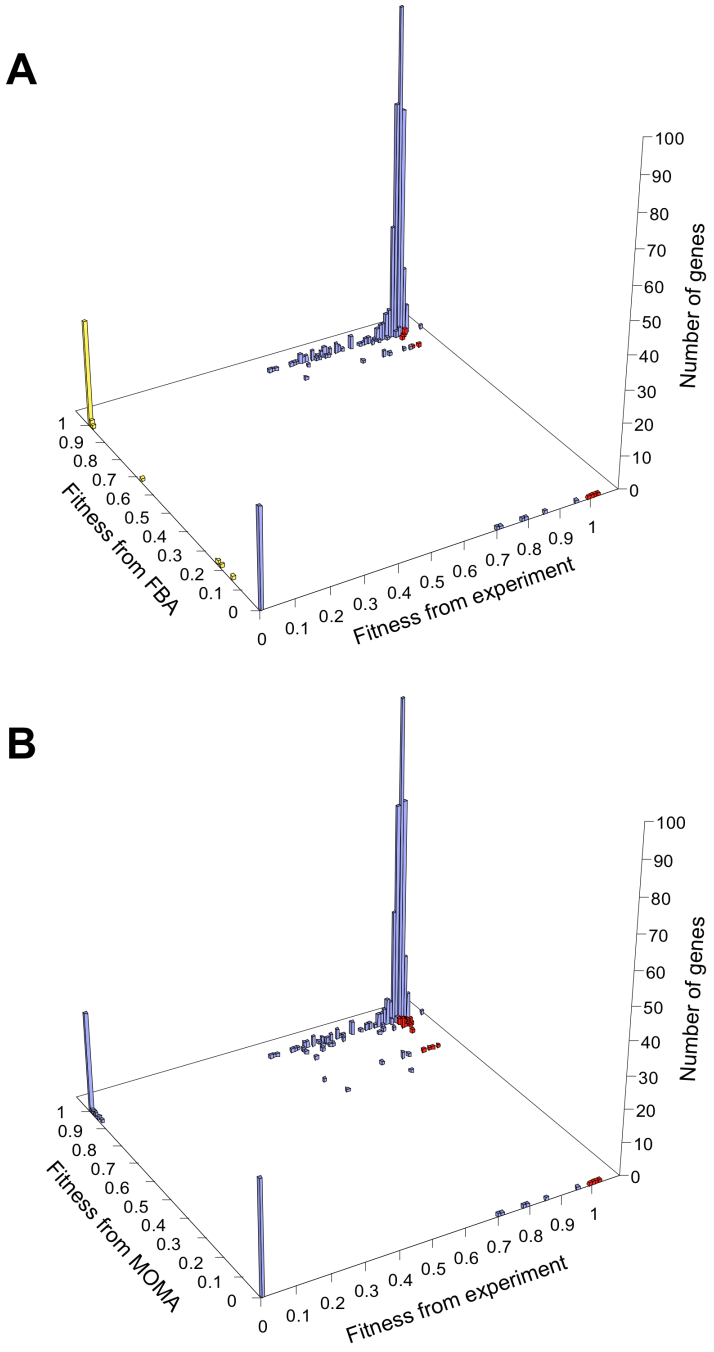


Figure 3.4 Accuracy of flux balance analysis (FBA) and minimization of metabolic adjustment (MOMA). The predicted the fitness values of single-gene-deletion yeast strains in rich medium by (**A**) FBA and (**B**) MOMA were compared to the previously experimentally determined fitness values (see Materials and Methods). A total of 485 genes encoding metabolic enzymes are examined here. Yellow bars are genes that are essential by experimental determination, but nonessential by FBA prediction. Red bars are single-gene-deletion strains that have fitness of >0.99 in experiments, but <0.99 by either FBA or MOMA prediction.



3.7 REFERENCES

Almaas E, Oltvai ZN, Barabasi AL. 2005. The activity reaction core and plasticity of metabolic networks. PLoS Comput Biol 1:e68.

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2:727-738.

Blank LM, Kuepfer L, Sauer U. 2005. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6:R49.

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301-312.

Clark AG. 1994. Invasion and maintenance of a gene duplication. Proc Natl Acad Sci U S A 91:2950-2954.

Conant GC, Wagner A. 2004. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271:89-96.

de Visser JA, Hermisson J, Wagner GP, et al. 2003. Perspective: Evolution and detection of genetic robustness. Evolution Int J Org Evolution 57:1959-1972.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4:e1000113.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. Nat Genet 40:676-681.

Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169:1915-1925.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667-1686.

Duarte NC, Herrgard MJ, Palsson BO. 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14:1298-1309.

Dudley A, Janse D, Tanay A, Shamir R, Church G. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. Molecular Systems Biology doi: 10.1038/msb4100004.

Edwards JS, Covert M, Palsson BO. 2002. Metabolic Modeling of Microbes: the Flux Balance Approach. Environmental Microbiology 4:133-140.

Edwards JS, Ibarra RU, Palsson BO. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125-130.

Famili I, Forster J, Nielsen J, Palsson BO. 2003. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci U S A 100:13134-13139.

Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO. 2005. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng 91:643-648.

Forster J, Famili I, Palsson BO, Nielsen J. 2003. Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics 7:193-202.
Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226-2238.
Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.
Harrison R, Papp B, Pal C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci U S A 104:2307-2312.
Hartman JL, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.
He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885-1891.
Hillenmeyer ME, Fung E, Wildenhain J, et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320:362-365.
Ibarra RU, Edwards JS, Palsson BO. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420:186-189.
Kafri R, Dahan O, Levy J, Pilpel Y. 2008. Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy. Proc Natl Acad Sci U S A.
Kandror O, DeLeon A, Goldberg AL. 2002. Trehalose synthesis is induced upon exposure of Escherichia coli to cold and is essential for viability at low temperatures. Proc Natl Acad Sci U S A 99:9727-9732.
Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet 20:287-290.
Lenski RE, Barrick JE, Ofria C. 2006. Balancing robustness and evolvability. PLoS Biol 4:e428.
Lercher MJ, Pal C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25:559-567.
Liang H, Li WH. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet 23:375-378.
Liao BY, Zhang J. 2007. Mouse duplicate genes are as essential as singletons. Trends Genet 23:378-381.
Mahadevan R, Schilling CH. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264-276.
Musso G, Costanzo M, Huangfu M, et al. 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res 18:1092-1099.
Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. Nature 388:167-171.
Ochman H, Wilson AC. 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74-86.
Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37:1372-1375.

Papin JA, Price ND, Edwards JS, Palsson BB. 2002. The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol 215:67-82.
Papp B, Pal C, Hurst LD. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429:661-664.
Podlaha O, Zhang J. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci U S A 100:12241-12246.
Price ND, Papin JA, Palsson BO. 2002. Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome Res 12:760-769.
Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.
Reed JL, Vo TD, Schilling CH, Palsson BO. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4:R54.
Segre D, Deluna A, Church GM, Kishony R. 2005. Modular epistasis in yeast metabolism. Nat Genet 37:77-83.
Segre D, Vitkup D, Church GM. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99:15112-15117.
Sliwa P, Korona R. 2005. Loss of dispensable genes is not adaptive in yeast. Proc Natl Acad Sci U S A 102:17670-17674.
Steinmetz LM, Scharfe C, Deutschbauer AM, et al. 2002. Systematic screen for human disease genes in yeast. Nat Genet 31:400-404.
Wagner A. 2005a. Distributed robustness versus redundancy as causes of mutational robustness. Bioessays 27:176-188.
Wagner A. 2005b. Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365-1374.
Wagner A. 2005c. Robustness and Evolvability in Living Systems. Princeton, NJ: Princeton University Press.
Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? PLoS Genetics in press.
Zhang J, Webb DM. 2003. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. Proc Natl Acad Sci U S A 100:8337-8341.

CHAPTER 4

Genomic Patterns of Pleiotropy and the Evolution of Complexity

4.1 ABSTRACT

Pleiotropy refers to the phenomenon of a single mutation or gene affecting multiple distinct phenotypic traits and has broad implications in many areas of biology. Due to its central importance, pleiotropy has also been extensively modeled, albeit with virtually no empirical basis. Analyzing phenotypes of large numbers of yeast, nematode, and mouse mutants, we here describe the genomic patterns of pleiotropy. We show that the fraction of traits altered appreciably by the deletion of a gene is minute for most genes and the gene-trait relationship is highly modular. The standardized size of the phenotypic effect of a gene on a trait is approximately normally distributed with variable standard deviations for different genes, which gives rise to the surprising observation of a larger per-trait effect for genes affecting more traits. This scaling property counteracts the pleiotropy-associated reduction in adaptation rate (i.e., the "cost of complexity") in a nonlinear fashion, resulting in the highest adaptation rate for organisms of intermediate complexity rather than low complexity. Intriguingly, the observed scaling exponent falls in a narrow range that maximizes the optimal complexity. Together, the genome-wide observations of overall low pleiotropy, high modularity, and larger pre-trait effects from genes of higher pleiotropy necessitate major revisions of theoretical models of pleiotropy and suggest that pleiotropy not only allowed but may have also promoted the origin of complexity.

4.2 INTRODUCTION

Pleiotropy occurs when a single mutation or gene affects multiple distinct phenotypic traits (Tyler et al. 2009). Pleiotropy has broad implications in genetics (Wright 1968; Barton 1990; Tyler et al. 2009), development (Hodgkin 1998; Carroll 2008), senescence (Williams 1957), disease (Albin 1993; Brunner and van Driel 2004), and many evolutionary processes such as adaptation (Fisher 1930; Wright 1968; Waxman and Peck 1998; Orr 2000; Otto 2004; Carroll 2005), maintenance of sex (Hill and Otto 2007), preservation of redundancy (Wang and Zhang 2009), and stabilization of cooperation (Foster et al. 2004). For example, the antagonistic pleiotropy theory of senescence asserts that alleles beneficial to development and reproduction are deleterious after the reproductive age and cause senescence, which may explain why all species have a limited life span (Williams 1957). Pleiotropy is the main theoretical reason behind the hypothesis that morphological evolution occurs more frequently through *cis*-regulatory changes than through protein sequence changes (Carroll 2005), as the former are less pleiotropic than the latter. Pleiotropy also has important implications in human disease, because many genetic defects each affect multiple phenotypic traits. For instance, mutations in the homeobox gene *ARX* cause ambiguous genitalia and lissencephaly (whole or parts of the surface of the brain appear smooth) (OMIM #300215).

Due to pleiotropy's central importance in biology, several mathematical models of pleiotropy have been developed and important theoretical results have been derived from the analyses of these models (Fisher 1930; Turelli 1985; Wagner 1988; Waxman and Peck 1998). For example, Fisher proposed that every mutation affects every trait and the effect size of a mutation on a trait is uniformly distributed (Fisher 1930). Based on this model and the assumption that the total effect size of a mutation is constant in different organisms, Orr derived that the rate of adaptation of a population to an environment quickly declines with the increase of the organismal complexity defined by the total number of traits (Orr 2000). This "cost of complexity" would likely prohibit the origins of complex organisms and hence is puzzling to evolutionary biologists (Welch and Waxman 2003; Haygood 2006).

Although pleiotropy has been examined in detail in a few genes (Brown, Barlow, and Wynshaw-Boris 1999; Hekerman et al. 2005), its genomic pattern is largely

unknown, which seriously limits us from evaluating the mathematical models of pleiotropy, verifying the theoretical inferences from these models, and testing various pleiotropy-related hypotheses in many fields of biology. In this work, we compile from existing literature and databases phenotypes of large numbers of yeast, nematode, and mouse mutants. We describe the genomic patterns of pleiotropy in these organisms and show that these patterns drastically differ from any mathematical model of pleiotropy. We further demonstrate that the "cost of complexity" is substantially alleviated when the empirical patterns of pleiotropy are taken into consideration and that the observed value of a key parameter of pleiotropy falls in a narrow range that maximizes the optimal complexity.

4.3 RESULTS AND DISCUSSION

4.3.1 Most genes affect only a small fraction of traits

To uncover the genomic patterns of pleiotropy, we complied three large datasets of gene pleiotropy for the baker's yeast *Saccharomyces cerevisiae*, one for the nematode worm *Caenorhabditis elegans*, and one for the house mouse *Mus musculus*. The first dataset, yeast morphological pleiotropy, is based on the measures of 279 morphological traits in haploid wild-type cells and 4,718 haploid mutant strains that each lack a different nonessential gene (Ohya et al. 2005). The second dataset, yeast environmental pleiotropy, is based on the growth rates of the same collection of yeast mutants relative to the wild-type in 22 different environments (Dudley et al. 2005). The third dataset, yeast physiological pleiotropy, is based on 120 literature-curated physiological functions of genes recorded in Comprehensive Yeast Genome Database (CYGD). The fourth dataset, nematode pleiotropy, is based on the phenotypes of 44 early embryogenesis traits in *C. elegans* treated with genome-wide RNA-mediated interference (Sonnichsen et al. 2005). The fifth dataset, mouse pleiotropy, is based on the phenotypes of 308 morphological and physiological traits in gene-knockout mice recorded in Mouse Genome Informatics (MGI). These five datasets provide qualitative information about the traits that are affected appreciably by each gene. In addition, the first dataset also includes the quantitative information of the effect size of each gene on each trait. Even after the

removal of genes that do not affect any trait and traits that are not affected by any gene, these five datasets each include hundreds to thousands of genes and tens to hundreds of traits (Figure 4.1). They are thus suitable for examining genome-wide patterns of pleiotropy.

In all five datasets, we observed that most genes affect only a small fraction of traits and only a minority of genes affect many traits (Figure 4.1). The median degree of pleiotropy varies from 1 to 7 traits (or 1-9% of the traits considered) in these datasets. The degree of pleiotropy measured by the percentage of examined traits is expected to be more accurate for the three datasets in which the same set of traits were examined in all mutants (Figure 4.1A, 4.1B, and 4.1D). By bootstrapping traits, we found that the standard deviations of our estimated median and mean degrees of pleiotropy are generally small, indicating that these estimates are precise (Figure 4.1). To examine the impact of the number of examined traits on the estimated pleiotropy, we randomly removed 50% and 90% of the traits from each dataset, respectively. We found that the mean and median degrees of pleiotropy, measured by the percentage of traits examined, remain largely unchanged (Table C.5), suggesting that further additions of traits to our data would not substantially alter our results. We thus predict that the median number of traits affected by a gene is no greater than a few percent of the total number of traits in an organism. Furthermore, because gene pleiotropy is largely owing to the involvement of the same molecular function in multiple different biological processes rather than the presence of multiple molecular functions per gene (He and Zhang 2006), random mutations in a gene will likely affect the same traits as the deletion of the gene does, although the magnitude of the phenotypic effects should be much smaller. Consequently, the observable degree of pleiotropy is expected to be even lower for random mutations than for gene deletions. Our genome-wide results echo recent small-scale observations from fish and mouse quantitative trait locus (QTL) studies (Albert et al. 2008; Wagner et al. 2008) and an inference from protein sequence evolution (Su, Zeng, and Gu 2009), and reveal a general pattern of low pleiotropy in eukaryotes, which is in sharp contrast to some commonly used theoretically models (Fisher 1930; Turelli 1985) that assume universal pleiotropy (i.e., every gene affects every trait) (Table C.6).

4.3.2 Gene-trait relationships are highly modular

The genome-wide data also allow us to test the modular pleiotropy hypothesis, which is important for a number of theories of development and evolution (Wagner, Pavlicev, and Cheverud 2007). Gene-trait relationships can be represented by a bipartite network of genes and traits, in which a link between a gene node and a trait node indicates that the gene affects the trait (Figure 4.2A). Modular pleiotropy refers to the phenomenon that links within modules are significantly more frequent than those across modules (Figure 4.2B). Given that cellular functions are modularly and hierarchically organized (Wagner, Pavlicev, and Cheverud 2007), modular pleiotropy likely exists, although it is not considered in commonly used models of pleiotropy (Fisher 1930; Turelli 1985; Orr 2000) (Table C.6). Employing a bipartite-network-specific algorithm (Barber 2007), we identified modules and estimated the modularity of each gene-trait network. Because random networks of certain structures also have non-zero modularity (Wang and Zhang 2007), we compared the modularity of an observed network with that of its randomly rewired networks, which have randomized links but an unchanged number of links per node (Wang and Zhang 2007) (Figure 4.2C). We then calculated the scaled modularity of a network, which is the difference between the observed modularity of a network and the mean modularity of its randomly rewired networks in terms of the number of standard deviations (Wang and Zhang 2007). Our results show large scaled modularity (34 to 238) in each of the five gene-trait networks examined (Figure 4.2D-4.2H), providing definitive evidence for the modular pleiotropy hypothesis. Our results remain qualitatively unchanged even when 50% of the traits in each dataset are removed (Table C.7). The modularity would be overestimated if the genetic correlations among traits are biased upward in our datasets compared to the complete datasets that include all possible traits. Although we do not know if this bias exists, to be conservative, we merged traits whose genetic correlation coefficients are greater than 0.7 (see Methods). We found that highly significant modularity is still present in each of the five gene-trait networks (Table C.7).

4.3.3 Genes affecting more traits have larger per-trait effects

The yeast morphological pleiotropy data contain quantitative information about the phenotypic effect size of mutations, which is another important parameter in genetics that has never been available at the genomic scale. Using a standardized measure of effect size for all traits (*Z*-score, defined by the phenotypic difference between a mutant and the mean of the wild-type for a trait in terms of the number of standard deviations; see Methods), we obtained, for each yeast gene, the frequency distribution of the effect sizes on the 279 morphological traits. As exemplified in Figure 4.3A, this distribution is approximately normal for most genes; the actual distribution is not significantly different from a normal distribution for 85% of the genes examined (5% false discovery rate in the goodness-of-fit test). This is consistent with a commonly used model (Turelli 1985), but is in contrast to another where the distribution is assumed to be uniform (Fisher 1930; Orr 2000) (Table C.6). In fact, the uniform distribution can be rejected for every gene at the significance level of $P = 5\times10^{-7}$ (goodness-of-fit test). It is notable that the standard deviation of the effect size distribution varies greatly among genes (Figure 4.3B), in contrast to models that assume a constant standard deviation among genes (Fisher 1930; Turelli 1985; Orr 2000) (Table C.6). It is also notable that the typical effect size distribution has a nearly zero mean, although a minority of genes exhibit positive or negative means (Figure C.6).

If one considers only those traits that are significantly affected by a gene, the total size of the phenotypic effects of the gene can be calculated by the Euclidean distance $T_E = \sqrt{\sum_{i=1}^{n} Z_i^2}$, where n is the gene's degree of pleiotropy defined by the number of significantly affected traits and Z_i is the gene's effect on trait i measured by the Z-score (Wagner et al. 2008). We estimated from the yeast morphological pleiotropy data that the exponent b in the scaling relationship of $T_E = an^b$ equals 0.601, with its 95% confidence interval of (0.590, 0.612) (Figure 4.3C). This exponent is significantly greater than that assumed in any theoretical model (Table C.6). For example, the invariant total effect model (Orr 2000) assumes a constant total effect size ($b = 0$), whereas the Euclidian superposition model (Turelli 1985; Wagner 1988; Waxman and

Peck 1998) assumes a constant effect size per affected trait ($b = 0.5$). Our results indicate that the per-trait effect of a gene is larger when the gene affects more traits. One can also measure the total effect size by the Manhattan distance (Hermisson and McGregor 2008) $T_M = \sum_{i=1}^{n} |Z_i|$. We found the exponent d in the scaling relationship of $T_M = cn^d$ to be 1.095, with its 95% confidence interval of (1.083, 1.107) (Figure 4.3D). Again, the observed d significantly exceeds that assumed in all current models (0.5 in the invariant total effect model and 1 in the Euclidian superposition model; Table C.6) and indicates larger per-trait effects for genes affecting more traits. To examine the robustness of the above results, we randomly removed 50% and 90% of the traits from the data, respectively. Our results that b and d are slightly smaller when the number of traits used is smaller (Figure C.7A-C.7D) suggest that, when more traits are examined in the future, b and d would become slightly greater than the current estimates. Our results are also robust to merging traits with genetic correlations (Figure C.7E-C.7F). Because 279 morphological traits were measured in each yeast mutant, in the above analyses, a 5% false discovery rate was used as a cutoff to control for multiple testing in determining whether a trait is affected by a gene. Our results remain qualitatively unchanged when the more conservative $P = 5\%$ after Bonferroni correction is used to correct for multiple testing (Figure C.8).

We observed that the phenomenon of larger per-trait effects for genes affecting more traits disappeared when the effect sizes of all genes on all traits are randomly shuffled (Figure C.9). Thus, the phenomenon is a property of the actual data rather than an artifact of our analysis. It turns out that this phenomenon results from two genome-wide features of pleiotropy described above: (i) a normal distribution of effect sizes of a gene on different traits and (ii) variable standard deviations of the normal distributions among different genes. Comparing two genes both having normal distributions of effect sizes but with different standard deviations, we proved mathematically that the gene with the larger standard deviation affects more traits (when a fixed effect-size cutoff is applied) and has on average a larger per-trait effect (Appendix A). In fact, the scaling relationships with the observed b and d values can be largely recapitulated by using randomly generated effect size data, provided that a normal distribution with the actual

standard deviation is used in generating such data for each gene (Figure 4.3E and 4.3F). By contrast, when the same standard deviation is used in generating the random effect size data for all the genes, we no longer observe larger per-trait effects for genes affecting more traits ($b < 0.5$ in Figure 4.3G and $d < 1$ in Figure 4.3H).

It is interesting to note that a recent mouse QTL study (Wagner et al. 2008) also reported $b > 0.5$, but a subsequent analysis (Hermisson and McGregor 2008) showed that, owing to the likelihood of the inclusion of multiple genes per QTL, the data can only establish $b > 0$, but not $b > 0.5$. Because our yeast morphological pleiotropy data were collected from strains that each lack only one gene, they are immune from the above multiple-gene problem. Furthermore, because our data were generated by examining all yeast nonessential genes and a large number of traits, they are more likely to reveal the general patterns of pleiotropy. It is important to recognize that our results are based on pleiotropic effects of genes (i.e., null mutations) rather than random mutations. However, our results likely apply to random mutations because the effect sizes of random mutations in a gene are expected to be proportional to the effect sizes of the gene (see Methods).

4.3.4 The "cost of complexity" is diminished with the actual patterns of pleiotropy

One of the most puzzling results from theoretical analysis of pleiotropy is the "cost of complexity" conundrum (Orr 2000). Using Fisher's geometric model (Fisher 1930), Orr (Orr 2000) showed that the rate of adaptation of an organism in which every mutation potentially affects all of the organism's n traits is $U = \frac{dw}{dt} = -\frac{4kT_E^2}{n} Mw \ln w$, where k is the product of the effective population size and the mutation rate per generation per genome (in the functional part), T_E is the total effect size of a mutation defined earlier, w is the current mean fitness of the population relative to the optimal, and M is a function of T_E and n (see Methods). Empirical evidence suggests that k may increase slightly with the level of organismal complexity, but the exact relationship between them is unclear (Haygood 2006). To be conservative, we here assume k to be independent of n. Note that although we are using the original formula (Orr 2000) for U which was based on a fixed mutation size T_E for a given n, this formula is known to be

robust to variable mutation sizes (Welch and Waxman 2003). If T_E is independent of n, as assumed in the invariant total effect model (Orr 2000), or if T_E is proportional to $n^{0.5}$, as assumed in the Euclidian superposition model (Turelli 1985; Wagner 1988; Waxman and Peck 1998) (Table C.6), adaptation rate U decreases with the degree of pleiotropy (or level of organismal complexity) n (Figure 4.4A), creating the "cost of complexity". Interestingly, the relationship between U and n changes when the scaling exponent b exceeds 0.5. It can be shown mathematically that, when $b > 0.5$, an intermediate level of complexity yields the highest adaptation rate (Appendix A) (Figure 4.4A). The n value that corresponds to the highest adaptation rate ($n_{optimal}$) depends on several parameters, including a (referred to as the mutation size), b, and w. Smaller a and w values lead to larger $n_{optimal}$ (Figure 4.4B). The null mutations in the yeast morphological pleiotropy data yield $a = 2.9$, but we expect natural random mutations to have a much smaller a, because they have on average much smaller phenotypic effects than gene deletions do (see Methods). For example, if $a = 0.01$ for natural random mutations, $b = 0.6$ as we have shown, and $w = 0.9$, $n_{optimal}$ becomes 9 (Figure 4.4A). Numerically, we found that, when a and w are given, $n_{optimal}$ reaches its maximum at an intermediate b value (Figure 4.4C). By examining a large parameter space ($10^{-8} \le a \le 10^{-2}$; $0.3 \le w \le 0.99$), we observed that the b value that offers the maximal $n_{optimal}$ occurs in a narrow range between 0.56 and 0.79 (Figure 4.4D), although b potentially can vary from negative infinity to positive infinity.

4.4 CONCLUSION

In summary, our genome-wide analysis of pleiotropy in yeast, nematode, and mouse revealed a generally low level of pleiotropy for most genes in a eukaryotic genome and a highly modular structure in the gene-trait relationship. Furthermore, the quantitative morphological data from yeast showed that genes affecting more traits tend to have larger per-trait phenotypic effects. Although an organism potentially contains many more traits than our data currently include, several analyses indicated that our results are robust and therefore our conclusions are expected to be largely unchanged even when most or all traits of an organism are considered. These findings necessitate a major revision of the current theoretical models that lack the above three empirical

features of pleiotropy (Table C.6) and require reevaluation of biological inferences derived from these models. For example, these three features substantially alleviate the cost of complexity in adaptive evolution. First, the generally low pleiotropy means that even mutations in organisms as complex as mammals do not normally affect many traits simultaneously. Second, high modularity reduces the probability that a random mutation is deleterious, because the mutation is likely to affect a set of related traits in the same direction rather than a set of unrelated traits in random directions (Welch and Waxman 2003; Martin and Lenormand 2006). Third, the greater per-trait effect size for more pleiotropic mutations (i.e., $b > 0.5$) causes a greater probability of fixation and a larger amount of fitness gain when a beneficial mutation occurs in a more complex organism than in a less complex organism. These effects, counteracting lower frequencies of beneficial mutations in more complex organisms (Fisher 1930), result in intermediate levels of complexity having the highest rate of adaptation. Together, they explain why complex organisms could have evolved in despite the cost of complexity. Whether the intriguing finding that the empirically observed scaling exponent b falls in a narrow range that offers the maximal optimal complexity is the result of natural selection for evolvability or a by-product of other evolutionary processes (Pigliucci 2008) requires further exploration.

4.5 METHODS

4.5.1 Pleiotropy datasets

The yeast morphological pleiotropy dataset (Ohya et al. 2005) includes the phenotypic information acquired by fluorescent imaging of 4,718 yeast nonessential gene deletion haploid strains as well as the wild-type haploid strain. The phenotypes include 501 quantitative traits of yeast cellular morphology such as cell shape, actin cytoskeleton, and nuclear morphology. These traits were first measured in 126 independent wild-type cells to estimate the wild-type variation and were then measured in one cell per deletion strain. The raw data were obtained from http://scmd.gi.k.u-tokyo.ac.jp/datamine/. Following the suggestion of the authors of the dataset (Ohya et al. 2005), we transformed the raw data of the 501 traits by power transformation (Yeo and Johnson 2000) and then

checked for normality in distribution among wild-type cells using the Shapiro-Wilk test. For 222 traits, the phenotypes of the wild-type cells either are not power-transformable or do not follow normal distributions. These traits were thus excluded from subsequent analysis and the remaining 279 traits were considered in our morphological pleiotropy data.

The size of the phenotypic effect of a gene on a trait is measured by the statistical *Z*-score, which is defined by $Z = (m_d - m_{wt})/SD$, where m_{wt} and *SD* are the mean and standard deviation of the transformed measures of the trait from wild-type cells, respectively, and m_d is the transformed measure of the trait from a cell in which the gene is deleted. Note that because m_d can be larger or smaller than m_{wt}, *Z* can be positive or negative. A given $m_d - m_{wt}$ value indicates a greater fitness effect when it occurs in a more important trait than in a less important trait. Because the *SD* of a trait is expected to be negatively correlated with the strength of stabling selection on the trait (i.e., the importance of a trait to organismal fitness), Z-scores effectively standardize phenotypic effects in terms of fitness effects and thus are comparable among traits. In order to determine the number of traits a gene affects, we calculated the statistical *P*-values according to the *Z*-scores using the standard normal distribution. Because we simultaneously tested 279 traits for each gene, we corrected for multiple testing using a 5% false discovery rate (FDR). In other words, if a trait shows a *Q*-value < 5% in a gene-deletion strain, we consider that this trait is affected by this gene. By this cutoff, a gene affects on average 22 traits. Thus, the number of false positives is lower than 1 trait per gene. We also used the more conservative Bonferroni correction of multiple testing, and the results are shown in Figure C.9. After the removal of genes that do not affect any trait and traits that are not affected by any gene, the yeast morphological dataset contains 2,449 genes and 253 traits.

The same collection of yeast gene deletion strains were also screened under 22 different environmental conditions for growth defects (Dudley et al. 2005). A gene is considered to affect growth under a condition when the deletion strain shows significantly slower growth than the wild-type strain. Because the data did not contain quantitative measures of growth rates, the gene-trait relationship is qualitative. That is, a gene either affects or does not affect a trait. In total, 774 genes affect growth in at least

one of the 22 environmental conditions. This dataset is referred to as the yeast environmental pleiotropy dataset.

We also obtained yeast knockout phenotype information from Comprehensive Yeast Genome Database (CYGD) (Guldener et al. 2005), which catalogs literature-curated physiological defects of yeast gene deletion strains from small-scale experiments. After removing phenotypes that are annotated as "unclassified", we obtained our yeast physiological pleiotropy data containing 1,256 genes that affect one or more of 120 traits. As in the yeast environmental pleiotropy dataset, this dataset only has qualitative information about gene-trait relationships.

In order to identify genes required for early embryogenesis in nematodes, a recent study used genome-wide RNA-mediated interference (RNAi) to silence gene expression in early *C. elegans* embryos (Sonnichsen et al. 2005). The targeted RNAi experiment for each gene was repeated in 6 embryos, and 45 phenotypic traits were screened for developmental defects. We consider that a gene affects a trait if at least 2 of the 6 embryos showed phenotypic defects. After the removal of one trait named "complex phenotype", we obtained our nematode pleiotropy dataset including 661 genes that affect one or more of 44 traits. This dataset only provides qualitative information about gene-trait relationships.

The mouse pleiotropy data were derived from annotations of MGI version 4.2 (http://www.informatics.jax.org/) (Bult et al. 2008). At the time of this study, 5,586 mouse genes were annotated with one or more Mammalian Phenotype (MP) IDs indicating the phenotypes when the genes were knocked out, knocked down, mutated by transgenic insertions, or occasionally mutated by point mutations. MP IDs are hierarchically structured. That is, one parent MP ID (e.g., MP:0002102, abnormal ear morphology) represents a phenotype lineage which may include several child MP IDs to describe a more detailed phenotype (e.g., MP:0000026, abnormal inner ear morphology; MP:0002177, abnormal outer ear morphology). Here, we used 308 parent MP IDs to define the pleiotropy of mouse genes. These 308 MP IDs were manually selected using the criterion that each MP ID should be phenotypically distinct, if not independent, from the other MP IDs. If a mouse gene is annotated for a child MP, the parent MP ID that this child MP ID belongs to is used. Consequently, pleiotropy of 4,915 mouse genes

associated with at least one of the 308 MP IDs were obtained. This dataset only provides qualitative information of the gene-trait relationships.

4.5.2 Modularity of gene-trait bipartite networks

Gene-trait relationships can be represented by a bipartite network where the genes form one type of nodes and traits form the second type of nodes. A link between a gene node and a trait node indicates that the gene affects the trait. To separate modules, we used BRIM (Barber 2007), which is modified from the widely used Newman definition of modularity (Newman 2004) for bipartite networks. For a given module partition, this algorithm calculates the difference between the density of within-module links and its random expectation. It then attempts to find the module partition that yields the highest difference, which is called the modularity of the network. Because even a random network may have a non-zero modularity (Guimera, Sales-Pardo, and Amaral 2004; Wang and Zhang 2007), we used scaled modularity (Wang and Zhang 2007) to measure the level of modularity of a network. We also calculated scaled modularity after merging traits whose genetic correlation coefficient is greater than 0.7. We chose this cutoff because, after the merge, no trait can genetically explain more than one half of the variance of another trait (0.7^2=0.49).

4.5.3 The scaling relationships between the degree of pleiotropy and the total effect size

Using the yeast morphological pleiotropy data, we calculated the number (n) of traits that are significantly affected by each gene. We then measured a gene's total phenotypic effect on these n traits, using either the Euclidian distance (T_E) or the Manhattan distance (T_M). We expect the scaling relationships of $T_E = an^b$ and $T_M = cn^d$. We estimated a, b, c, and d using the curve fitting toolbox in MATLAB, which employs a non-linear least-squares method to fit the observations and calculates the confidence intervals of the estimated parameters.

Because gene pleiotropy is largely owing to the involvement of the same molecular function in multiple different biological processes rather than the presence of multiple molecular functions per gene (He and Zhang 2006), random mutations in a gene

will likely affect the same traits as the deletion of the gene does, although the magnitude of the phenotypic effects should be much smaller. For simplicity, let us assume that the effect on a trait from a random mutation in a gene is on average *h* times the effect on the same trait from a null mutation in the gene, where the effect is again measured by *Z*-score and $0 < h << 1$. Let T_E' be the total effect size of the random mutation in Euclidean distance. It can be shown that $T_E' = hT_E = (ah)n^b$. Thus, the scaling relationship between the total effect size of a random mutation and pleiotropy is the same as that between the total effect size of a null mutation and pleiotropy, except that the mutation size parameter for random mutations is *h* times that for null mutations.

4.5.4 Simulating normally distributed phenotypic effects of genes

For a given gene *i*, we first calculated the standard deviation (σ_i) of its phenotypic effect size distribution from the yeast morphological pleiotropy data. Note that, in this calculation, we used the phenotypic effects of the gene on all 279 traits, regardless of whether these effects are statistically significant or not. We then randomly generated this gene's phenotypic effects on each of the 279 traits using a normal distribution with mean equal to 0 and standard deviation equal to σ_i. We did this for all 4,718 genes to produce a 4718 × 279 random effect-size matrix. We then analyzed this simulated dataset following the analysis of the real data.

To examine the impact of different standard deviations of different genes on our results, we conducted the second simulation. The procedure is the same as the above simulation, except that, instead of using different standard deviations for different genes, we used the same standard deviation for all genes. This standard deviation used was the mean standard deviation for all genes in the actual data.

4.5.5 Calculating the rate of adaptation

Assuming Fisher's geometric model, Orr (Orr 2000) derived the formula for the rate of fitness increase during an adaptive walk to the optimal to be

$U = \frac{dw}{dt} = -\frac{4kT_E^2}{n} Mw \ln w,$ where *n* is the degree of pleiotropy, which also measures

organismal complexity, k is the product of the effective population size and the mutation rate per generation per genome (in the functional part), T_E is the total effect size of a mutation measured by the Euclidean distance, and w is the current mean fitness of the population relative to the optimal, $M = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} (y - x)^2 e^{\frac{-y^2}{2}} dy$, and $x = \frac{T_E \sqrt{n}}{2\sqrt{-2 \ln w}}$. In Orr's calculation (Orr 2000), T_E was assumed to be independent of n. In our model, T_E scales with the degree of pleiotropy by $T_E = an^b$, where a is the mutation size parameter that corresponds to the mutation size when the degree of pleiotropy is 1 and b is the scaling exponent. We implemented numerical calculations of the above formulas in MATLAB.

4.6 ACKNOWLEDGEMENTS

We thank Meg Bakewell, Xionglei He, Wenfeng Qian, and Jianrong Yang for valuable comments. This work was supported by US NIH research grants to J.Z. and Taiwan NHRI intramural funding to B.-Y.L.

Figure 4.1 Frequency distributions of degree of gene pleiotropy in different species. (**A**) Yeast morphological, (**B**) yeast environmental, (**C**) yeast physiological, (**D**) nematode, and (**E**) mouse pleiotropy data. Mean and median degrees of pleiotropy and their standard deviations are indicated in each panel. The numbers in the parentheses are the mean and media degrees of pleiotropy divided by the total number of traits. After the removal of genes that do not affect any trait and traits that are not affected by any gene, the total number of genes and traits in these datasets are (**A**) 2449 genes and 253 traits, (**B**) 774 genes and 22 traits, (**C**) 1256 genes and 120 traits, (**D**) 661 genes and 44 traits, and (**E**) 4915 genes and 308 traits.

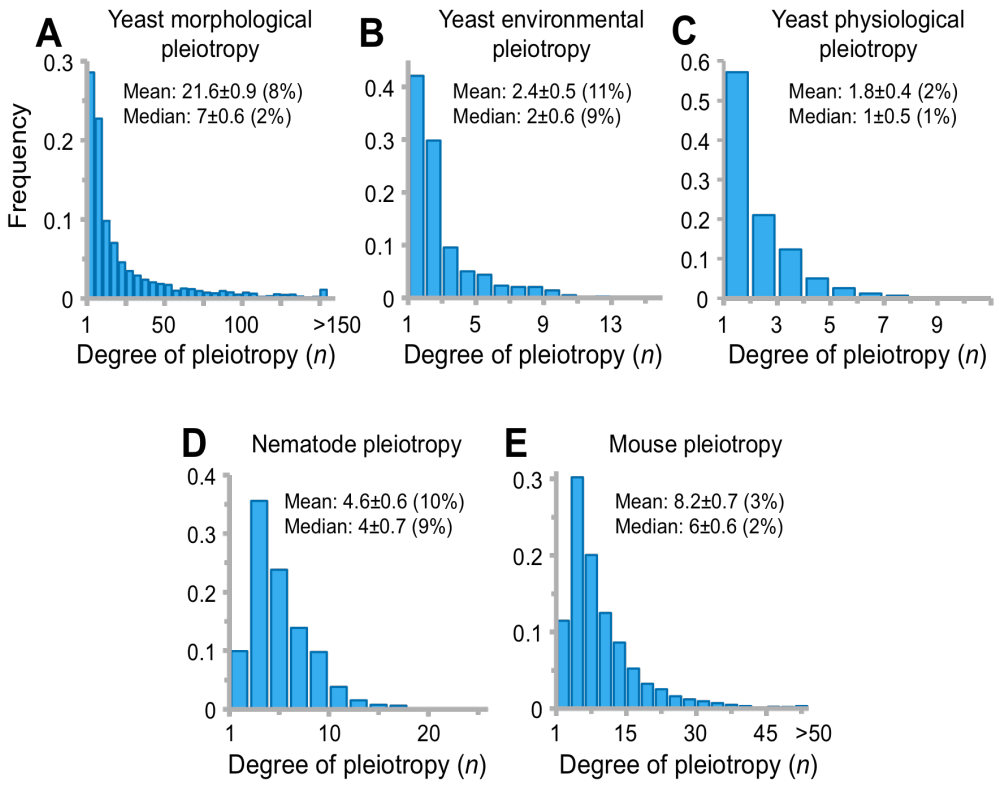


Figure 4.2 High modularity of gene-trait bipartite networks. (**A**) A hypothetical gene-trait bipartite network. A link between a gene and a trait indicates that the gene affects the trait, while the thickness of the link indicates the effect size. (**B**) Two modules are identified in the hypothetical gene-trait network after the quantitative links are transformed to qualitative links (i.e., presence/absence) based on whether an effect size is significantly different from 0. (**C**) A randomly rewired network that has the same degree distribution as the original hypothetical network shows no detectable modular structure. The modularity and scaled modularity of the hypothetical bipartite network are 0.41 and 3.9, respectively. Panels (**D**)-(**H**) show the observed modularity (blue arrows) and distribution of modularity for 250 randomly rewired networks (red histograms) for the gene-trait networks of the (**D**) yeast morphological, (**E**) yeast environmental, (**F**) yeast physiological, (**G**) nematode, and (**H**) mouse pleiotropy datasets.

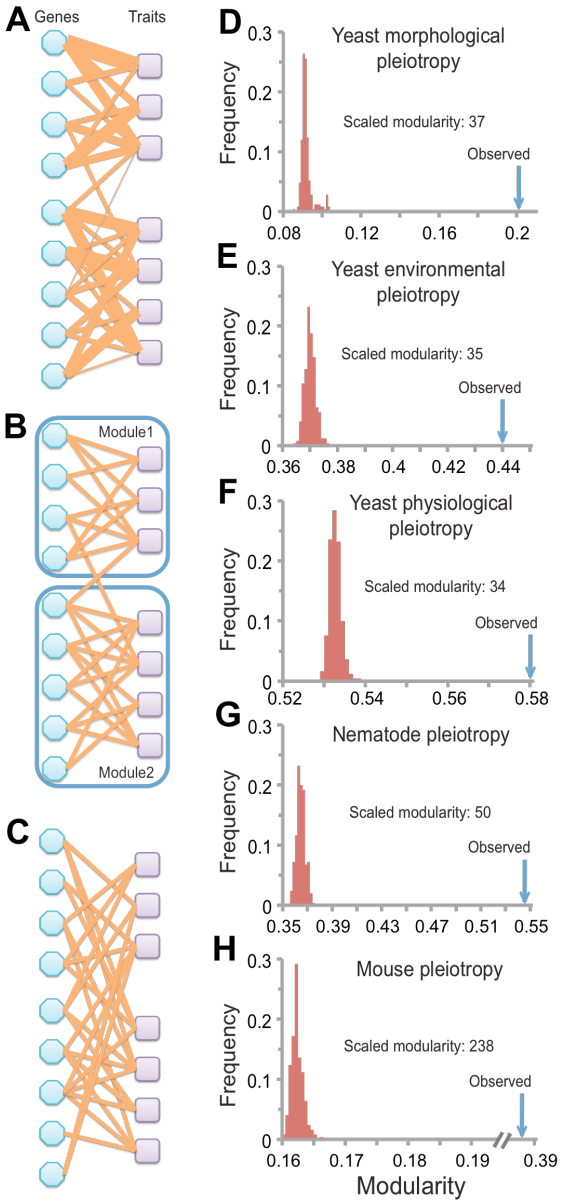


Figure 4.3 Scaling relationships between the total phenotypic effect size of a gene and the degree of pleiotropy in the yeast morphological pleiotropy data. (**A**) Examples showing the normal distribution of effect size over 279 traits. Two genes are chosen to show variable standard deviations of the normal distributions. (**B**) Distribution of the standard deviation (S.D.) of the effect size for all 4718 genes. Observed scaling relationships between the degree of pleiotropy n and the total phenotypic effect of a gene measured by (**C**) Euclidean distance or (**D**) Manhattan distance. The orange curve is the best fit to the power function whose estimated parameters are shown inside the panel. The numbers after ± show the 95% confidence interval for the estimated scaling exponent. R^2 indicates the square of the correlation coefficient. Panels (**E**) and (**F**) are similar to panels (**C**) and (**D**) except that the effect sizes of each gene are randomly generated from a normal distribution with zero mean and observed standard deviation. Panels (**G**) and (**H**) are similar to panels (**C**) and (**D**) except that the effect sizes of each gene are randomly generated from a normal distribution with zero mean and a constant standard deviation, which is the average of all standard deviations of all genes.

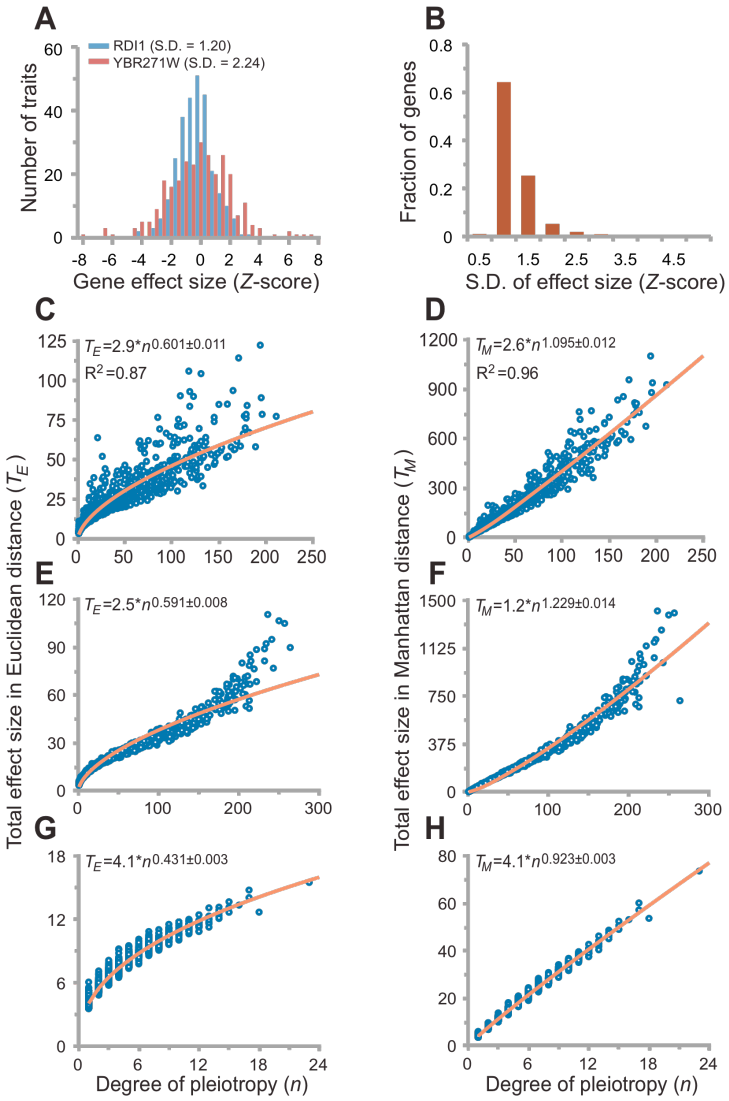
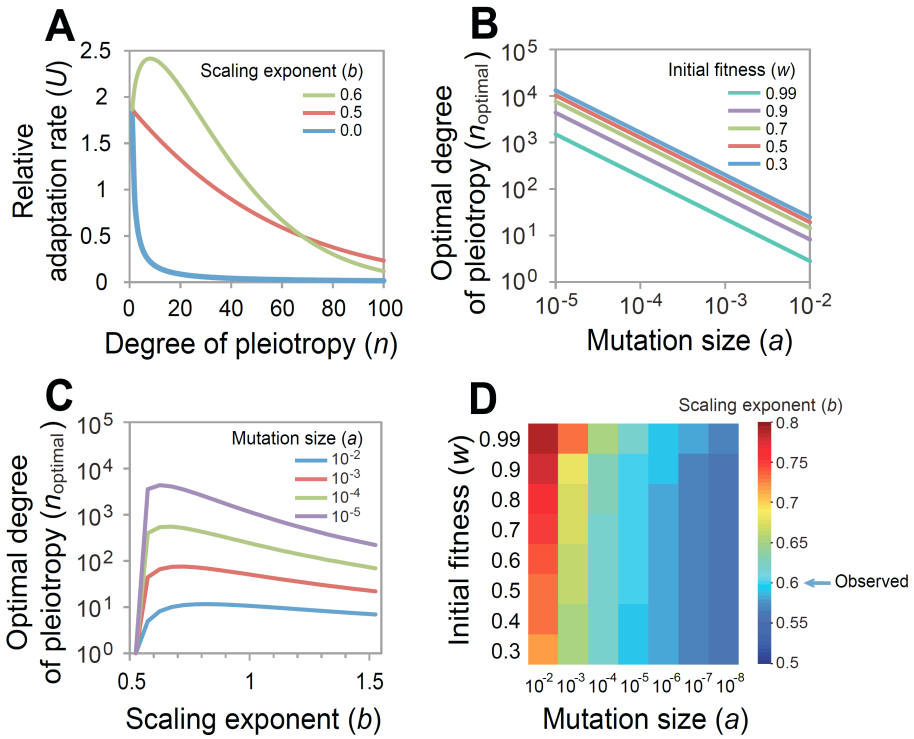


Figure 4.4 The "cost of complexity" is alleviated when the scaling exponent *b* exceeds 0.5. (**A**) The relative adaptation rate as a function of the degree of pleiotropy (n) changes with the scaling exponent b. The relative adaptation rate is calculated using Orr's formula. The initial fitness w is set at 0.9 and the mutation size a is set at 0.01. (**B**) The optimal degree of pleiotropy $n_{optimal}$, defined as the degree of pleiotropy that corresponds to the highest adaptation rate, changes with the mutation size a. Different curves are generated using different initial fitness (w) values but the same $b = 0.6$. (**C**) The optimal degree of pleiotropy $n_{optimal}$ changes with different b. Different curves are generated using different a but the same $w = 0.9$. (**D**) A heat map showing the b value that provides the maximal $n_{optimal}$, given a and w.



4.7 REFERENCES

Albert AY, Sawaya S, Vines TH, Knecht AK, Miller CT, Summers BR, Balabhadra S, Kingsley DM, Schluter D. 2008. The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. Evolution 62:76-85.

Albin RL. 1993. Antagonistic pleiotropy, mutation accumulation, and human genetic disease. Genetica 91:279-286.

Barber MJ. 2007. Modularity and community detection in bipartite networks. Phys Rev E Stat Nonlin Soft Matter Phys 76:066102.

Barton NH. 1990. Pleiotropic models of quantitative variation. Genetics 124:773-782.

Brown KD, Barlow C, Wynshaw-Boris A. 1999. Multiple ATM-dependent pathways: an explanation for pleiotropy. Am J Hum Genet 64:46-50.

Brunner HG, van Driel MA. 2004. From syndrome families to functional genomics. Nat Rev Genet 5:545-551.

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA. 2008. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 36:D724-728.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol 3:e245.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25-36.

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. Mol Syst Biol 1:2005 0001.

Fisher RA. 1930. The Genetic Theory of Natural Selection. Oxford: Clarendon.

Foster KR, Shaulsky G, Strassmann JE, Queller DC, Thompson CR. 2004. Pleiotropy as a mechanism to stabilize cooperation. Nature 431:693-696.

Guimera R, Sales-Pardo M, Amaral LA. 2004. Modularity from fluctuations in random graphs and complex networks. Phys Rev E Stat Nonlin Soft Matter Phys 70:025101.

Guldener U, Munsterkotter M, Kastenmuller G, et al. 2005. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res 33:D364-368.

Haygood R. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Mutation rate and the cost of complexity. Mol Biol Evol 23:957-963.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885-1891.

Hekerman P, Zeidler J, Bamberg-Lemper S, Knobelspies H, Lavens D, Tavernier J, Joost HG, Becker W. 2005. Pleiotropy of leptin receptor signalling is defined by distinct roles of the intracellular tyrosines. FEBS J 272:109-119.

Hermisson J, McGregor AP. 2008. Pleiotropic scaling and QTL data. Nature 456:E3; discussion E4.

Hill JA, Otto SP. 2007. The role of pleiotropy in the maintenance of sex in yeast. Genetics 175:1419-1427.

Hodgkin J. 1998. Seven types of pleiotropy. Int J Dev Biol 42:501-505.

Martin G, Lenormand T. 2006. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60:893-907.
Newman ME. 2004. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69:066133.
Ohya Y, Sese J, Yukawa M, et al. 2005. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A 102:19015-19020.
Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13-20.
Otto SP. 2004. Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc Biol Sci 271:705-714.
Pigliucci M. 2008. Is evolvability evolvable? Nat Rev Genet 9:75-82.
Sonnichsen B, Koski LB, Walsh A, et al. 2005. Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434:462-469.
Su Z, Zeng Y, Gu X. 2009. A preliminary analysis of gene pleiotropy estimated from protein sequences. J Exp Zool B Mol Dev Evol.
Turelli M. 1985. Effects of pleiotropy on predictions concerning mutation-selection balance for polygenic traits. Genetics 111:165-195.
Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays 31:220-227.
Wagner GP. 1988. The influence of variation and developmental constraints on the rate of multivariate phenotypic evolution. J. Evol. Biol. 1:44-66.
Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the 'cost of complexity'. Nature 452:470-472.
Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nat Rev Genet 8:921-931.
Wang Z, Zhang J. 2007. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 3:e107.
Wang Z, Zhang J. 2009. Abundant indispensable redundancies in cellular metabolic networks. Genome Biol. Evol. 1:23-33.
Waxman D, Peck JR. 1998. Pleiotropy and the preservation of perfection. Science 279:1210-1213.
Welch JJ, Waxman D. 2003. Modularity and the cost of complexity. Evolution 57:1723-1734.
Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescene. Evolution 11:398-411.
Wright S. 1968. Evolution and the Genetics of Populations: University of Chicago Press.
Yeo I-K, Johnson RA. 2000. A new family of power transformations to improve normality or symmetry Biometrika 87:954-959.

CHAPTER 5

Why is the Correlation between Gene Importance and Gene Evolutionary Rate So Weak?

5.1 ABSTRACT

One of the few commonly believed principles of molecular evolution is that functionally more important genes (or DNA sequences) evolve more slowly than less important ones. This principle is widely used by molecular biologists in daily practice. However, recent genomic analysis of a diverse array of organisms found only weak negative correlations between the evolutionary rate of a gene and its functional importance typically measured under a single benign lab condition. A frequently suggested cause of the above finding is that gene importance determined in lab differs from that in an organism's natural environment. Here we test this hypothesis in yeast using gene importance values experimentally determined in 418 lab conditions or computationally predicted for 10,000 nutritional conditions. In no single condition or combination of conditions did we find a much stronger negative correlation, which is explainable by our subsequent finding that always-essential (enzyme) genes do not evolve significantly slower than sometimes-essential or always-nonessential ones. Furthermore, we verified that functional density, approximated by the fraction of amino acid sites within protein domains, is uncorrelated with gene importance. Thus, neither the lab-nature mismatch nor a potentially biased among-gene distribution of functional density explains the observed weakness of the correlation between gene importance and evolutionary rate. We conclude that the weakness is factual rather than artifactual. In addition to population genetic reasons, the correlation is likely to have been further weakened by the presence of multiple nontrivial rate determinants that are independent from gene importance. These findings notwithstanding, we show that the principle of

slower evolution of more important genes does have some predictive power when genes with vastly different evolutionary rates are compared, explaining why the principle can be practically useful despite the weakness of the correlation.

5.2 INTRODUCTION

When referring to any DNA sequence, a popular textbook of cell and molecular biology (Karp 2008) states that "if it's conserved, it must be important" and calls this "one of the foremost principles of molecular evolution" (p. 416). Here, the word "conserved" means that the sequence has a low rate of evolution such that its orthologs from distantly related species are detectable and alignable. The word "important" means that the sequence has relevance to the wellbeing and fitness of the organism bearing the sequence. The above principle is often used in a comparative context, asserting that functionally more important DNA sequences evolve more slowly. Despite the fact that thousands of biologists accept this principle and use it daily in identifying functionally important DNA sequences, its validity had not been systematically examined until a few years ago when gene importance could be measured at the genomic scale (Hurst and Smith 1999; Hirsh and Fraser 2001; Jordan et al. 2002; Yang, Gu, and Li 2003; Rocha and Danchin 2004; Wall et al. 2005; Zhang and He 2005; Liao, Scott, and Zhang 2006; Wolf, Carmel, and Koonin 2006). Unexpectedly, however, genomic studies of bacteria, fungi, and mammals showed that although the evolutionary rate of a gene is significantly negatively correlated with its importance, the latter only explains a few percent of the total variance of the former (Krylov et al. 2003; Wall et al. 2005; Zhang and He 2005; Liao, Scott, and Zhang 2006). The striking contrast between the wide acceptance and apparent utility of the principle and the weakness of the correlation revealed from genomic analysis of a diverse array of organisms is perplexing.

The perceived theoretical basis of this simple principle is the neutral theory of molecular evolution, which asserts that most nucleotide substitutions during the evolution of a gene are due to random fixations of neutral mutations (Kimura 1968; King and Jukes 1969; Kimura 1983). Based on this theory, Kimura and Ohta first predicted that functionally more important genes should evolve slower than less important ones because

the former have a lower rate of neutral mutation than the latter (Kimura and Ohta 1974), although their use of "functional importance" appears to mean "functional constraint on the gene" rather than "importance to the fitness of the organism". A few years later, Wilson *et al.* separated the two meanings and decomposed the substitution rate of a gene (k) into two factors: the probability (P) that a random mutation will be compatible with the function of the gene and the probability (Q) that an organism can survive and reproduce normally without the gene (i.e., gene dispensability) (Wilson, Carlson, and White 1977). Under the simple assumption that a mutation either completely abolishes the function of a gene (with a probability of $\alpha = 1\text{-}P$) or does not affect it at all (with a probability of $1\text{-}\alpha$), we can write the substitution rate of a gene as the sum of the rate of fixation of neutral mutations and that of null mutations. Here, α can also be interpreted as functional density, the effective fraction of sites in a gene (or protein) that are required for its function. Let u be the total mutation rate, $\beta = 1\text{-}Q$ be the probability that an organism cannot survive or reproduce without the gene (i.e., gene importance or the coefficient of selection against null mutations), N be the organism's population size, and N_e be the effective population size. For diploid organisms, we have

$$k = (1-\alpha)u + \alpha(2Nu)f = u[1-\alpha(1-2Nf)], \qquad (5.1)$$

where $f = \dfrac{1-e^{\beta N_e/N}}{1-e^{2\beta N_e}}$ is the probability of fixation of a new null mutation with fitness $0<Q<1$, under genic selection (i.e, the selection against the null allele is β in homozygotes and $\beta/2$ in heterozygotes) (Kimura 1983). Because $f < 1/(2N)$, k is a monotonically decreasing function of α. It is obvious that k is also a monotonically decreasing function of β, because the stronger the selection against null mutations, the lower f and k are. However, note that the above formula also indicates that in large populations, f and hence k should be relatively insensitive to β except when β is extremely small (i.e., on the order of $1/N_e$). In other words, under the simplistic model assumed here, a strong negative correlation between gene importance and evolutionary rate is not expected (Hurst and Smith 1999) (see also Appendix B and Figure C.10). However, under a more realistic model with the presence of slightly and moderately deleterious mutations, a much stronger correlation between gene importance and evolutionary rate becomes

theoretically possible (Hirsh and Fraser 2001). The strength of the correlation depends on the distribution of the deleterious functional effects of random mutations (Appendix B and Figure C.10). Because the true distribution is currently unknown, theories cannot predict precisely the strength of the correlation between gene importance and evolutionary rate. These considerations notwithstanding, the apparent utility of the principle in daily practice and its lack of empirical support from genomewide studies require an explanation.

There are two simple, yet untested, hypotheses that potentially explain the weakness of the observed correlation between gene importance and evolutionary rate. First, the importance of a gene to an organism is now commonly measured by the fitness reduction caused by the deletion of the gene from the genome in a benign lab condition; deleting an important gene reduces the fitness of the organism more than deleting a less important one. But, because lab conditions differ significantly from the natural environments of organisms, gene importance determined in lab may be quite different from that in nature (Hurst and Smith 1999; Wolf 2006). For example, in rich media, ~80% of yeast genes are not essential for growth (Papp, Pal, and Hurst 2004). However, metabolic network analysis and experimental studies showed that most of these dispensable genes are important for growth under other conditions (Papp, Pal, and Hurst 2004; Hillenmeyer et al. 2008), some of which may resemble the natural environments of the species better than rich media. Hence, it is plausible that the weakness of the correlation between gene importance and evolutionary rate is due to inaccuracy in measuring genes' natural importance, which we refer to as the lab-nature mismatch hypothesis. But, measuring gene importance in a species' natural environment is difficult because many species such as the yeast *Saccharomyces cerevisiae* are found in diverse environments that are poorly characterized (Fay and Benavides 2005). Moreover, even if we know the present-day natural environments of a species, they may not reflect the environments where the species lived in the past. These historical environments are crucial because the gene evolutionary rate that is being correlated to gene importance is determined by comparison between species. Nonetheless, if gene importance is measured in many different conditions, we can examine whether the correlation between gene importance and evolutionary rate is much stronger in some conditions than in the

benign lab condition, which could at least demonstrate the plausibility of the lab-nature mismatch hypothesis. Here we test this hypothesis in yeast using gene importance measures from both experimental data and computational predictions. The experimental data came from a set of recently published fitness measurements of yeast single-gene-deletion strains under 418 lab stress conditions (Hillenmeyer et al. 2008). We complemented this dataset with *in silico* predictions of importance for metabolic enzyme genes under 10^4 nutritional conditions, achieved by flux balance analysis (FBA) of reconstructed metabolic networks (Edwards, Covert, and Palsson 2002; Price, Reed, and Palsson 2004).

Another potential factor influencing the correlation between gene importance (β) and evolutionary rate (k) is functional density (α) in Equation 5.1. If α and β are negatively correlated (i.e., more important genes have lower functional density), the correlation between k and β will be weakened. Although there is no reason to believe that α and β are negatively correlated, it is worth verifying using actual data. For a given protein, α may be approximately measured by the fraction of sites in functional domains, which can be computationally predicted.

In this work, we show that neither of the above two hypotheses is correct in yeast. Rather, the weakness of the correlation between gene importance and evolutionary rate is likely to be factual rather than artifactual. We show, however, that the principle of slower evolution of more important genes does have some predictive power when genes with vastly different evolutionary rates are compared, explaining why the principle can be practically useful despite the weakness of the correlation.

5.3 RESULTS AND DISCUSSION

5.3.1 Testing the lab-nature mismatch hypothesis with experimental measures of gene importance

The most frequently used yeast gene importance data came from the measures of relative growth rates of 5936 single-gene-deletion yeast strains in the nutritionally rich YPD medium (Steinmetz et al. 2002). Recently, the same type of measure was taken for

all YPD-viable single-gene-deletion yeast strains under 418 diverse laboratory conditions, of which ~75% are chemical drug treatments and the rest are environmental stress conditions such as different pHs and temperatures (Hillenmeyer et al. 2008). These two datasets of gene importance are used in our analysis.

Evolutionary rates of *S. cerevisiae* genes are estimated by comparing these genes to their orthologs in related species. Because the functional importance of a gene may change during evolution (Zhang and He 2005; Liao and Zhang 2008), it is better to use a closely related species for rate estimation. However, when the species are too close, the number of nucleotide substitutions per gene may be insufficient for precise estimation of evolutionary rates. A previous study found the strongest correlation between gene importance and evolutionary rate when *S. cerevisiae* is compared with *S. bayanus* (Zhang and He 2005). We thus use this species pair and obtain 3999 genes with identifiable orthologs. Our results remain qualitatively unchanged when several other yeast species were compared with *S. cerevisiae* (data not shown). We use the number of nonsynonymous substitutions per nonsynonymous site (d_N) between orthologs to measure the rate of gene evolution (k in Equation 5.1). Because the mutation rate (u in Equation 5.1) may vary among genes, we also use the ratio between d_N and the number of synonymous substitutions per synonymous site (d_S) as a measure of k/u in Equation 5.1.

When gene importance is measured under the nutritionally rich YPD medium, the Spearman's rank correlation coefficient between gene importance (i.e., amount of fitness reduction caused by gene deletion) and d_N is $\rho = -0.2189$ ($P < 10^{-43}$; Figure 5.1A). Our examination of 418 other lab conditions found the strongest correlation to be $\rho = -0.2379$ ($P < 10^{-51}$; Figure 5.1A). Thus, none of the 418 conditions provides a substantially stronger correlation than what is observed with YPD. Similar results were obtained for the correlation between gene importance and d_N/d_S (Figure 5.1B).

Krylov et al. suggested another measure of gene evolutionary rate known as the propensity for gene loss (*PGL*), which is the number of times that a gene is lost during the evolution of a group of species (Krylov et al. 2003). Although *PGL* and d_N are correlated with each other (Krylov et al. 2003), they measure the rate of gene evolution from different angles. The correlation between *PGL* and gene importance is expected to be weaker than that between d_N and gene importance, because mutations that impair gene

function only slightly do not matter to gene loss. We estimated *PGL* for each *S. cerevisiae* gene by counting the number of gene loss events on the known phylogeny of 12 fungal species (see Materials and Methods). Consistent with our expectation, the correlation between gene importance and *PGL* is weaker than that between gene importance and d_N (or d_N/d_S) for both YPD and the other 418 lab conditions (Figure 5.1C). Regardless, the examination of the 418 lab conditions does not substantially improve the strength of the correlation between gene importance and *PGL*.

5.3.2 Testing the lab-nature mismatch hypothesis with computationally predicted gene importance values

Because the 418 experimentally examined conditions contain mostly artificial chemical treatments and hence may not cover the diverse natural environments of the yeast, we decide to complement the experimental data with computationally predicted gene importance values for 546 metabolic enzyme genes under 10^4 conditions generated by random combinations of different nutrients following a sampling strategy that mimics the potential nutritional environments of the wild yeast (see Materials and Methods). We then used two different experimentally validated computational methods to predict the fitness reduction caused by the deletion of each enzyme gene. These methods rely on the reconstructed high-quality yeast metabolic network (Duarte, Herrgard, and Palsson 2004), which contains 632 biochemical reactions associated with 546 enzyme genes after the removal of dead-end reactions (Burgard et al. 2004). The first method we used is flux balance analysis (FBA). Under the assumption of steady state of every cellular metabolite, FBA maximizes the rate of biomass production under the stoichiometric constraints of all metabolic reactions (Edwards, Covert, and Palsson 2002). Simulation of different nutritional conditions is achieved by setting the boundaries of uptake reaction fluxes and simulation of gene deletion is achieved by constraining the flux of corresponding enzymatic reaction to zero (see Materials and Methods). In our analysis, we consider the FBA-optimized rate of biomass production as the wild-type Darwinian fitness of the cell under the condition specified. The relative fitness of a cell lacking a gene is the FBA-optimized rate of biomass production of the cell, divided by that of the wild-type cell. Previous studies demonstrated that FBA makes excellent qualitative

predictions of yeast gene essentiality under typical experimental conditions (Duarte, Herrgard, and Palsson 2004; Papp, Pal, and Hurst 2004). A recent study further showed consistent performances of FBA across many different conditions (Snitkin et al. 2008). Following a previous study (Forster et al. 2003), we approximated the YPD condition in the FBA model and predicted the fitness values of single-gene-deletion yeast strains. We found that the FBA-predicted fitness values correlate well with the experimentally determined fitness values under YPD (Pearson's $r = 0.562$, $P < 10^{-41}$). We were not able to verify FBA for the other 418 lab conditions because these conditions are difficult to specify in FBA.

Our extensive analysis of 10^4 simulated conditions identified the strongest correlation between FBA-predicted gene importance and d_N to be $\rho = -0.2186$ ($P = 10^{-6}$; Figure 5.1D) for 546 enzyme genes. Although this correlation is 34% stronger than that estimated using experimentally determined gene importance under YPD ($\rho = -0.1636$, $P = 6\times10^{-4}$; Figure 5.1D) for the same set of genes, the fraction of variance in d_N that is explainable by gene importance is still as low as $(-0.2186)^2 = 4.8\%$. Similar results are obtained when either d_N/d_S (Figure 5.1E) or *PGL* (Figure 5.1F) is used as a measure of gene evolutionary rate. One interesting observation is that the standard deviation of ρ from the 10^4 simulated conditions (0.042, 0.037, and 0.037 in Figure 5.1D, 5.1E, and 5.1F, respectively) is much greater than that for the 418 experimental conditions (0.013, 0.009, and 0.008 in Figure 5.1A, 5.1B, and 5.1C, respectively). Part of this difference is due to the use of essentially all genes in Figure 5.1A-5.1C but only enzyme genes in Figure 5.1D-5.1F. However, even when only enzyme genes are considered, the standard deviation of ρ is still smaller for lab conditions (d_N: 0.024; d_N/d_S: 0.021; *PGL*: 0.020) than for the 10^4 simulated conditions, suggesting that the simulated conditions represent a more diverse set of conditions than the experimental conditions.

FBA assumes that a cell can readjust its metabolic fluxes to achieve the highest possible biomass production immediately after the deletion of any gene, which is probably unrealistic. Segre and colleagues proposed a modified method known as the minimization of metabolic adjustment (MOMA) (Segre, Vitkup, and Church 2002). Instead of maximizing biomass production upon gene deletion, MOMA minimizes the changes of fluxes from those of the wild-type cell. Empirical examples suggested that

MOMA outperforms FBA in predicting gene essentiality and metabolic fluxes (Segre, Vitkup, and Church 2002). We found that MOMA-predicted fitness values of single-gene-deletion strains are slightly better than FBA-predicted values in correlating with the experimentally determined fitness values in YPD (Pearson's $r = 0.571$, $P < 10^{-43}$). However, none of the 10^4 simulated conditions provide a better correlation between MOMA-predicted gene importance and evolutionary rate than the correlation found using experimentally measured gene importance in YPD (Figure 5.1G-5.1H).

Although we examined 10^4 simulated conditions, it is possible that they still do not cover the natural conditions of yeast. We simulated 10^5 additional conditions and found that the distribution of the correlation coefficient ρ (Figure C.11) is virtually identical with that from the initial 10^4 conditions. Because the distribution of ρ is approximately normal, statistically speaking, it is extremely unlikely to obtain a much stronger correlation by examining even 10^6 conditions. Due to the large amount of computational time required for examining large numbers of conditions and the similarity of the results from 10^4 and 10^5 conditions, we used the gene importance values predicted from the 10^4 conditions in subsequent analysis.

5.3.3 Testing the lab-nature mismatch hypothesis using combinations of individual conditions

Because under no single condition, either experimentally examined or computationally simulated, did we find a strong correlation between gene importance and evolutionary rate, and because yeast may have had experienced diverse natural conditions during its evolution, we ask whether we can find combinations of single conditions for which the correlation between gene importance and evolutionary rate is much stronger than that under any single condition. We consider a simple scenario in which gene importance values under different conditions are weighted and linearly combined to form an average gene importance value across all the conditions considered. These weighting coefficients potentially represent the (unknown) relative durations of the conditions where the yeast has lived. We identify these coefficients by mathematically maximizing the correlation between the weighted average gene importance and evolutionary rate. We further constrain the weighting coefficients to be non-negative because negative

coefficients are biologically meaningless. Employing the least squared method in statistics, we can transform this maximization task into a quadratic programming problem. The mathematical representation of the problem is

$$\text{minimizing object: } Z = \sum_i (f_i - k_i)^2 \text{ , where } f_i = \sum_j c_j f_{ij},$$

$$\text{subject to: } c_j \geq 0 \text{ for any } j, \qquad (5.2)$$

where k_i is the evolutionary rate of gene i and f_i is the weighted average importance of gene i in all conditions, calculated by averaging gene importance under each condition (f_{ij}) using non-negative weighting coefficients of the condition (c_j). We solved the quadratic programming problem using the commercial optimization package CPLEX and then calculated the correlation between the weighted average importance of a gene and its evolutionary rate. Note, however, that the above estimation of c guarantees the identification of the strongest Pearson's linear correlation between f_i and k_i, but not Spearman's rank correlation. We know of no method that guarantees the identification of the strongest rank correlation between f_i and k_i.

Our results showed that the improvement of the correlation by combining individual conditions is trivial (Table 5.1). For example, for the 418 experimental conditions, the strongest Pearson's correlation between the weighted average gene importance and d_N is $r = -0.2187$ ($P < 10^{-43}$), only 5% stronger than the strongest correlation found among all single conditions ($r = -0.2082$, $P < 10^{-39}$). Similar results were observed for the other measures of gene evolutionary rate and for combinations of the 10^4 simulated conditions (Table 5.1). These results indicate that even weighted average of gene importance across multiple conditions is not strongly correlated with gene evolutionary rate.

Why doesn't the consideration of so many experimental and simulated conditions and combinations of conditions improve the correlation between gene importance and evolutionary rate? Using FBA, one can classify enzyme genes into three categories according to their importance across multiple conditions: always-essential, sometimes-essential, and always-nonessential. Deleting an always-essential gene causes lethality in all conditions; deleting a sometimes-essential gene causes lethality in some but not all conditions; deleting an always-nonessential gene does not cause lethality in any

condition, although it may reduce the fitness of the organism to a non-zero level. Because always-essential genes are as important as or more important than the other two classes of genes in any condition, it is clear that in order to achieve a strong correlation between gene importance and evolutionary rate in any condition or combination of conditions, the evolutionary rate of always-essential genes must be lower than those of the other two classes of genes. Here the enzyme genes are classified into the above three groups based on the essentiality predicted in the 10^4 simulated conditions. Although the average d_N of always-essential genes is lower than that of sometimes-essential genes and that of always-nonessential genes, the differences are small and not statistically significant (Figure 5.2A). The same is true for d_N/d_S (Figure 5.2B) and *PGL* (Figure 5.2C). These results strongly suggest that no single condition or combination of conditions will show a strong correlation between gene importance and evolutionary rate even when more conditions are examined. Thus, if the conditions under which yeast evolved belong to the 418 experimentally examined conditions or are amenable to the current FBA, the lab-nature mismatch hypothesis must be rejected.

5.3.4 Examining the correlation between functional density and gene importance

Equation 5.1 shows that if functional density (α) and gene importance (β) are independent from each other, evolutionary rate of a gene (k) should decrease with the increase of β. The observed weakness of the correlation between gene importance and evolutionary rate prompts us to examine the presumption of independence between α and β, because the correlation between gene importance and evolutionary rate could have been weakened if there is a negative correlation between α and β. By definition, α is the proportion of mutations that destroy the function of a gene, which may be experimentally determined by large-scale site-directed mutagenesis coupled with gene functional assay, a formidable task even for a few genes. In theory, one can use the average number of allowable alternative states across all amino acid sites of a protein to estimate 1-α. But such a measure is currently difficult to acquire at the genomic scale, because it requires the alignments of orthologs from many (i.e., >> 20) divergent species to assure that all

potentially allowed amino acids have had chance to appear at any given site. Use of many divergent species greatly increases misidentification of paralogs as orthologs and the risk of comparing functionally-different orthologous proteins, leading to potential overestimation of 1-α. A further complication is that the evolution of a site is often dependent on other sites, meaning that an amino acid is allowed at a site only when another site has a particular amino acid (Kondrashov, Sunyaev, and Kondrashov 2002; Gao and Zhang 2003). Thus, the number of allowed amino acids at a site is not a unique number, but rather depends on the genetic background of the same gene or even other genes. Given these difficulties, we decide to use the proportion of amino acid sites within computationally predicted functional domains of a protein to estimate α approximately, because α is expected to be much greater within functional domains than outside domains. This estimation of α is based on the assumption that all sites within functional domains are important to the function of the protein whereas all sites outside domains are unimportant. Although this assumption does not hold in reality, it should not affect our results as long as it does not systematically bias our estimation of α among genes of different β.

Computational algorithms for predicting protein functional domains are based on proteins of known structures and/or amino acid sequences with high evolutionary conservation (Copley et al. 2002). There are many available algorithms for protein domain prediction and they are based on different assumptions. Here we employ two widely used methods. The first is the ProSite prediction algorithm (Hulo et al. 2006), which is based on known conserved functional motif sequences. ProSite predictions are relatively conservative and should contain few false positives, as on average only 10% of amino acid sites in a protein are predicted by ProSite to be within functional domains. The second method we used is InterProScan (Mulder and Apweiler 2008), which integrates 13 well known domain prediction algorithms and databases to look for domains. Because InterProScan uses multiple algorithms, its predictions are more comprehensive. To avoid false positive predictions, we consider only those sites that are identified by at least two algorithms of InterProScan as functional domain sites. Under this criterion, on average 47% of protein sites are identified as functional domain sites.

To examine whether the proportion of sites within predicted domains indeed provide information about functional density, we conducted three tests. First, based on the domains predicted by ProSite, we found that sites within domains evolve more slowly than those outside domains in 89% of the yeast genes. The corresponding number is 77% when the domains are predicted by InterProScan. These percentages are significantly greater than the random expectation of 50 percent ($P < 10^{-100}$, χ^2 test). Second, the mean d_N within domains is 40% and 54% that outside domains in ProSite and InterProScan analysis, respectively, both being significantly different from the random expectation of 100% ($P < 10^{-50}$, paired *t*-test). Finally, we examined if there is a negative correlation between the proportion of sites within domains and the evolutionary rate of the gene, and found the correlation to be $\rho = -0.24$ ($P < 10^{-50}$) and -0.56 ($P < 10^{-50}$), respectively, in ProSite and InterProScan analysis. Taken together, the proportion of sites within predicted domains indeed provide information about functional density and thus may be used as a proxy for α.

Because our results do not support the lab-nature mismatch hypothesis, we here consider only experimentally measured gene importance under YPD (β). We found very weak positive correlation between α estimated by ProSite and β ($\rho = 0.049$, $P = 0.0002$) (Figure 5.3A). If InterProScan predictions are used, there is a stronger positive correlation between α and β ($\rho = 0.15$, $P < 10^{-30}$), suggesting that important genes tend to have a higher fraction of functional sites (Figure 5.3B). We also repeated the analysis under more stringent criteria of InterProScan where a site is considered as a functional domain site only when it is recognized by at least 3 to 6 algorithms. The observed correlation between α and β remains significant ($\rho = 0.08$-0.12, $P < 0.0001$).

However, the above analysis has a confounding factor. Because sequence conservation information is used in predicting functional domains and because important genes tend to be more conserved in sequence (though the correlation is weak), the above observed level of positive correlation between a and b may in part or in total due to the artifact of the analysis. Indeed, we found that after the control of d_N, the partial correlation between a and b becomes $\rho = 0.0190$ ($P = 0.240$) for the ProSite analysis and $\rho = -0.0110$ ($P = 0.497$) for InterProScan analysis ($\geq$ two algorithms). This result

suggests no genuine correlation between α and β. Thus, the weakness of the correlation between gene importance and evolutionary rate is unlikely the result of a potential negative correlation between gene importance and functional density.

5.3.5 Why is the correlation between gene importance and evolutionary rate so weak?

Our analysis rejected two frequently proposed explanations of the weakness of the observed correlation between gene importance and evolutionary rate, raising the question of why the correlation is so weak. As mentioned in Introduction, depending on the distribution of the fitness effect of deleterious mutations, the expected correlation may not be strong (Figure C.10 and Appendix B). In addition, there may be other reasons. Bivariate analysis of yeast data revealed a strong negative correlation between gene expression level and evolutionary rate (Pal, Papp, and Hurst 2001), which led to the recent proposal of the translational robustness hypothesis, asserting that selection against toxicity of misfolded proteins generated by translational errors is the single most important factor governing the rate of protein sequence evolution (Drummond, Raval, and Wilke 2006; Drummond and Wilke 2008). This hypothesis explains several factors known to correlate with the rate of protein sequence evolution (e.g., gene expression level and codon usage bias). However, many other rate determinants are known in yeast, including the number of protein interaction partners and gene length, although their impacts on the evolutionary rate are generally much smaller than that of gene expression level (Pal, Papp, and Lercher 2006). Principal component regression analysis and partial correlation analysis have suggested independent and significant contributions of all these factors (Kim and Yi 2007; Plotkin and Fraser 2007), although it is not always clear how these factors determine the rate of gene evolution independently from the influence of gene importance (Drummond et al. 2005). In bacteria and mammals, independent contributions from multiple factors to gene evolutionary rate are also known (Rocha and Danchin 2004; Liao, Scott, and Zhang 2006). Theoretically speaking, the single most important rate determinant is the fraction of mutations that are unacceptable to the gene (α), but this fraction is affected by many biological factors. The fact that the rate of gene evolution is jointly determined by multiple independent factors, some of which are

stronger determinants than gene importance, is likely an additional reason why the rate is only weakly correlated with gene importance. To simplify the explanation, let us assume that the rate of gene evolution (k) is determined linearly by n independent factors (A_1 to A_n) as $k = a_1A_1 + a_2A_2 + ... + a_nA_n + \varepsilon$, where ε represents the statistical error that cannot be explained by the n factors and a_i's are coefficients. Pearson's correlation coefficient between k and factor A_i is

$$r(k, A_i) = r(k, a_iA_i) = \frac{Cov(k, a_iA_i)}{\sqrt{Var(k)Var(a_iA_i)}}$$
$$= \frac{Var(a_iA_i)}{\sqrt{Var(k)Var(a_iA_i)}} = \sqrt{\frac{Var(a_iA_i)}{Var(k)}} \quad , \qquad (5.3)$$
$$= \sqrt{\frac{Var(a_iA_i)}{Var(\varepsilon) + \sum_{j=1}^{n} Var(a_jA_j)}}$$

where *Var* stands for variance and *Cov* stands for covariance. Because one rate determinant, gene expression, already accounts for >25% of the variance of k (Drummond, Raval, and Wilke 2006; Drummond and Wilke 2008) and several other factors also make independent and nontrivial contributions (Kim and Yi 2007; Plotkin and Fraser 2007), the correlation between gene importance and evolutionary rate is much weakened, compared to that when gene importance is the sole contributor..

5.3.6 Implications for predicting functional importance

Taken together, we showed empirically that the correlation between gene importance and gene evolutionary rate is weak and showed that this weakness may not be inconsistent with theoretical predictions. In fact, if we randomly pick two yeast genes, the probability that the slower evolving of the two is the more important one is only 54% (based on 100,000 pairs of randomly sampled genes under YPD) (Figure 5.4A). That is, the prediction based on one of the foremost principles of molecular evolution has a success rate of only 54%, not much greater than that of a pure guess (50%). When the two genes being compared have a larger difference in evolutionary rate, the prediction about their relative importance becomes more accurate, as expected (Figure 5.4A). For example, we ranked all yeast genes by their evolutionary rates and found that when two

genes are separated in rank by over 95% of all genes, the probability that the slower evolving one is more important than the other is 81% (Figure 5.4A). Essential genes are functionally most important. When the gene importance data from YPD is considered, we found that 55% of the top 5% most conserved genes are essential, whereas only 20% of the remaining 95% of yeast genes are essential (Figure 5.4B). Similar results are found using the gene importance data from the other 418 lab conditions (Figure 5.4B). Note that the above demonstrated predictability may not be entirely due to the causal relationship between gene importance and evolutionary rate, because other confounding factors such as gene expression level have not been controlled for. Regardless, our results show that although the correlation between gene importance and evolutionary rate is weak, the principle does have some predictive power when genes of extreme sequence conservation are considered.

5.3.7 Caveats

There are several caveats in our analysis that warrant discussion. First, experimental measures of gene importance are not without errors. Repeated measures of gene importance under the same conditions showed a correlation as high as 0.92 for the YPD data (Steinmetz et al. 2002) but a reduced mean correlation of 0.72 for the other 418 lab conditions (Hillenmeyer et al. 2008), possibly due to less well controlled experimental procedures in the latter. Thus, the gene importance data we used could potentially explain a maximum of $0.72^2 = 52\%$ of the variance of the evolutionary rate. But the strongest correlation actually observed was only $r^2 = 4.3\%$ among the 418 individual conditions and 4.8% among combinations of the 418 conditions, both being substantially lower than the theoretical maximum. Similar arguments can be made for the analysis based on computationally predicted gene importance values.

Second, a limitation in using d_N and d_N/d_S to measure the rate of gene evolution is that they can be used only for those *S. cerevisiae* genes that have orthologs in the species being compared with (i.e., *S. bayanus*). Our results would not represent a full picture if genes with and without orthologs have drastically different levels of gene importance. To examine this possibility, we compared their importance levels. Because we used reciprocal best hits in BLAST searches to define orthologs, a *S. cerevisiae* gene would

not have its operational *S. bayanus* ortholog, if (i) the gene evolved extremely fast, (ii) the gene has been lost in *S. bayanus*, or (iii) the gene has been duplicated in *S. cerevisiae* such that its *S. bayanus* best hit happens to find its paralog to be the best hit. Thus, we separated *S. cerevisiae* genes into singletons and duplicates. We found no significant difference in gene importance between *S. cerevisiae* genes with and without *S. bayanus* orthologs, for either singletons ($P = 0.11$, Mann-Whitney U test; Table C.8) or duplicates ($P = 0.63$, Table C.8). Hence, the potential bias of studying only *S. cerevisiae* genes that have *S. bayanus* orthologs is negligible.

Third, we used three different measures of gene evolutionary rate: d_N, d_N/d_S, and *PGL*. They all have pros and cons, aside from the above consideration. In principle, d_N/d_S would be the best measure, because it best measures k/u, which is determined by α and β only, according to Equation 5.1. Estimates of d_N/d_S, however, suffer from two problems. First, d_S values may have been saturated because the average d_S between *S. cerevisiae* and *S. bayanus* is as high as 1.24. Although using more closely related species could improve the estimation of d_S, it would increase the estimation error of d_N and that of d_N/d_S, due to a reduced number of nonsynonymous substitutions per gene. Second, codon usage bias, prevalent in highly expressed genes of yeast, could lead to underestimation of neutral substitution rates and thus overestimation of k/u. Because of the positive correlation between the importance of a gene and its expression level (Zhang and He 2005), codon usage bias causes greater overestimation of k/u for more important genes, weakening the negative correlation between k/u and gene importance. If there is little variation in mutation rate among genes, d_N would be a better index of evolutionary rate for our purpose than d_N/d_S, because estimates of d_N have smaller sampling errors than those of d_N/d_S. Our results show stronger correlations between gene importance and d_N, compared to that between gene importance and d_N/d_S, suggesting that the disadvantages of using d_N/d_S outweigh its advantages. Propensity for gene loss (*PGL*) treats each gene as a unit and does not consider the number of substitutions per nucleotide or amino acid site. It is thus conceptually different from the evolutionary rate that Kimura and Ohta (Kimura and Ohta 1974) and Wilson *et al.* (Wilson, Carlson, and White 1977) referred to. There are three reasons underlying our observation that gene importance correlates more poorly with *PGL* than with d_N and d_N/d_S. First, because *PGL*

is determined by the fixation of null mutations but not slightly deleterious mutations, it should be less influenced by gene importance, as explained in Introduction and Figure C.10. Second, estimation of *PGL* requires genome sequences from a number of species related to the focal species of interest (*S. cerevisiae*). In the present case, *PGL* is estimated from 12 diverse fungi and thus may not accurately reflect the propensity of gene loss in *S. cerevisiae*, because the importance of a gene can change in evolution (Zhang and He 2005; Liao and Zhang 2008). Third, estimates of *PGL* potentially have large sampling errors, because the estimated number of losses per gene is quite small.

Fourth, to understand why no single condition or combination of single conditions provides gene importance values that correlate strongly with evolutionary rates, we classified enzyme genes into three groups (always-essential, sometimes-essential, and always-nonessential) and compared their respective evolutionary rates. Due to computational intensity, our classification was based on the FBA analysis of 10^4 simulated conditions, while in theory it should have been based on all possible conditions. This limitation potentially caused misclassification of some truly sometimes-essential genes as always-essential genes or always-nonessential genes and hence blurred the differences among the three groups. To rectify this problem, we used a strategy that guarantees the identification of all always-essential genes. The metabolic model of yeast allows us to know all nutrients that can be used by this metabolic model. If a gene is essential when all these nutrients are present, it must be essential when one or more of these nutrients are absent. We find that in fact the always-essential genes thus identified are identical to those identified from the 10^4 simulated conditions. There is, however, no systematic way to guarantee the exact separation of sometimes-essential and always-nonessential genes. We thus merged them and compared this combined group with always-essential genes. Again, we do not find the combined group to have significantly greater d_N, d_N/d_S, or *PGL* than always-essential genes (Figure 5.2). Thus, our result is true not only for the 10^4 simulated conditions, but also for all possible combinations of nutrients usable by the yeast metabolic model. Our result differs from that of Papp *et al.* (Papp, Pal, and Hurst 2004) where they found that enzyme genes active in more conditions have lower probabilities of presence in the genomes of 133 diverse species. At least five reasons may account for this difference. First, we counted *PGL* on a known

phylogeny of related species using the parsimony method whereas these authors simply calculated the percentage of species that do not have the gene without considering the species phylogeny (Papp, Pal, and Hurst 2004). Second, most of the species they used are distantly related to yeast and their result is expected to be highly dependent on the choice of species. Third, we considered gene essentiality, a more relevant measure of gene importance than gene activity, because deleting an active gene may or may not have any fitness consequence, depending on alternative pathways in the metabolic network. Fourth, we used a more recent reconstruction of the yeast metabolic network, which is more complete and accurate than the one they used. Fifth and most importantly, because only nine conditions were examined, their result could simply be due to small sample size.

Fifth, Hirsh and Fraser suggested that the correlation between gene importance and evolutionary rate should exist only among genes with relatively low importance (Hirsh and Fraser 2001). This is because, in Equation 5.1, f quickly declines to virtually 0 when β increases from 0 to 0.1 and any further increase in β has negligible effects on f and k, although Hirsh and Fraser came to this conclusion using a more complex model (Hirsh and Fraser 2001). However, we found that the correlation for genes with $\beta < 0.1$ is extremely weak ($\rho = -0.05$ for YPD and the strongest $\rho = -0.04$ among the 418 experimental conditions). We cannot test genes with even smaller β because the accuracy of the estimated β decreases and the number of useable genes decreases. The contradiction between Hirsh and Fraser's prediction and our empirical observation can be understood using Figure C.10. Apparently, when there are many slightly and moderately deleterious mutations, use of all genes provides a stronger correlation than using only unimportant genes, because the expected evolutionary rates can still be different between a gene with $\beta = 0.2$ and a gene with $\beta = 0.3$ (Figure C.10K). For example, in Figure C.10L, using only genes with $\beta < 0.1$ gives $\rho = -0.36$, whereas using all genes gives $\rho = -0.83$.

Sixth, the correlation between gene importance and evolutionary rate reported here may be in part caused by other co-varying factors. For three reasons, we did not control for confounding factors in our analysis. First, previous authors already determined that the correlation is statistically significant even after the control of

confounding factors (Wall et al. 2005; Zhang and He 2005). Second, our goal here is to discern why the correlation is so weak even when part of it may come from confounding factors. Third, we study the difference in the magnitude of the correlation when various gene importance measures are used; confounding factors such as gene expression level would not affect this difference.

5.4 CONCLUSIONS AND IMPLICATIONS

Despite the general belief and wide application of the principle that important genes evolve more slowly than less important ones, genomic analysis showed that the correlation between gene importance and evolutionary rate is quite weak. Our analysis does not support the hypothesis that the weakness of the observed correlation is due to the difference between gene importance in the lab and in nature. Furthermore, we found no evidence for the possibility that the correlation is weakened by the potential presence of a smaller fraction of functional sites in more important genes. We conclude that the weakness of the correlation is factual, rather than artifactual. This conclusion is not inconsistent with population genetic predictions, because the predictions vary depending on the prevalence and distribution of the fitness effect of deleterious mutations.

Our result cautions molecular biologists from predicting relative functional importance of genes directly from their relative levels of evolutionary conservation. Nevertheless, our finding that extremely conserved genes are highly likely to be functionally very important may explain the universal perception that the principle of slower evolution of more important genes (or DNA sequences) works well. For example, substantial amount of comparative genomic work aims at using the principle to identify functional non-coding sequences based on their extremely low rates of nucleotide substitution (Boffelli et al. 2003; Kellis et al. 2003; Xie et al. 2005; Pennacchio et al. 2006). An ultra-conserved non-coding sequence is a segment of DNA of over 200 nucleotides with no variation among human, mouse, and rat. Pennacchio et al. found that such ultra-conserved sequences, when they are also conserved between mouse and fish, have a probability of 62% to be actual enhancers during mouse embryonic development (Pennacchio et al. 2006). Compared to the virtually zero probability with which a

random segment of DNA in the mouse genome is an enhancer, the principle appears to work well. This success is not surprising, because only extremely conserved non-coding sequences are considered. Nevertheless, it should be noted that although a large fraction of extremely conserved non-coding sequences are functional, many functional sequences are not extremely conserved. In other words, the current application of the principle in detecting functional non-coding sequences has a high false-negative rate. Thus far, there has been no evidence that the correlation between sequence importance and evolutionary rate is stronger for non-coding regions than for coding regions. One reason for a potentially stronger correlation for non-coding regions is that several rate determinants in coding sequence evolution simply do not exist in non-coding sequence evolution (e.g., codon usage bias, amount of translation, gene length, and number of protein-interacting partners). In addition, the fraction of mutations that are slightly deleterious may be greater for non-coding regions than for coding regions, given the high modularity of regulatory sequences. In the future when relative importance of many functional non-coding sequences is measured, it will be interesting to examine whether non-coding sequences exhibit a greater correlation between importance and evolutionary rate.

5.5 MATERIALS AND METHODS

5.5.1 Yeast gene importance values under YPD and other 418 lab conditions

The fitness values of homozygous-single-gene-deletion yeast strains in the YPD medium (Steinmetz et al. 2002) were downloaded from http://www-deletion.stanford.edu/YDPM/YDPM_index.html. The corresponding data from the other 418 lab conditions (Hillenmeyer et al. 2008) were obtained from http://chemogenomics.stanford.edu:16080/supplements/global/download.html. The microarray raw data were processed by the author-provided Perl scripts and were then normalized to the central mean to yield the relative fitness values of the deletion strains under each condition.

5.5.2 Yeast metabolic network

The metabolic network model of *S. cerevisiae* (iND 750) (Duarte, Herrgard, and Palsson 2004) used in this study was downloaded from the BiGG database (http://bigg.ucsd.edu) and parsed by the COBRA toolbox (Becker et al. 2007). The network is composed of 1149 reactions, associated with 750 known genes. Some reactions do not have associated genes because the genes whose protein products catalyze these reactions have yet to be identified. The network model also provides information about stoichiometry, direction of reaction, and gene-reaction association. We followed an established protocol (Burgard et al. 2004) to identify dead-end reactions, which are reactions that must have zero flux under a steady state. These reactions are involved in the generation of metabolites that are neither included in biomass nor transported outside the cell, and may reflect the incompleteness of the metabolic network model. After the removal of dead-end reactions, the yeast metabolic network used in our analysis contains 632 biochemical reactions with 546 associated enzyme genes.

5.5.3 Flux balance analysis (FBA) and minimization of metabolic adjustment (MOMA)

Details of FBA have been described in the literature (Edwards, Covert, and Palsson 2002; Price, Reed, and Palsson 2004). Briefly, the flux of each reaction is determined by maximizing the rate of biomass production under the assumption of steady state and the constraints of stoichiometry. We used the optimization package CPLEX (www.ilog.com) to solve the linear programming problem. Gene deletion is modeled by constraining the flux of the corresponding reaction to zero.

MOMA has been previously described in detail (Segre, Vitkup, and Church 2002). Briefly, MOMA predicts the maximal biomass production rate upon deletion of a reaction by minimizing the differences in all metabolic fluxes between the deletion strain and the wild-type strain. All the constraints used in FBA are still enforced in MOMA. The quadratic programming problem is also solved by CPLEX. As in FBA, deletion of a gene is realized by constraining the flux of the corresponding reaction to zero.

5.5.4 Simulation of nutritional conditions

The natural environments of yeast may change frequently. It is also likely that yeast usually faces nutritionally poor conditions but occasionally encounter rich conditions. To mimic their natural environments, we simulate random nutritional conditions in the following manner. For each condition, we generate a random number g from an exponential distribution with a mean of $m = 0.1$ for each of the 103 usable carbon-source nutrients. Here, g is the probability that the carbon-source nutrient is available. The actual presence or absence of each nutrient is then determined stochastically using g. We then add all required inorganic metabolites. Use of other m values (0.05 or 0.5) does not change our results. For each available nutrient, we fix the uptake rate at a random value between 0 and $D = 20$. The actual D value used is unimportant and does not alter our result. Only conditions that support the growth of the wild-type cell, as shown by FBA, are considered.

5.5.5 Separation of singleton from duplicate genes

Singleton and duplicate genes of yeast *S. cerevisiae* are identified by BlastP searches of each gene against all other genes in the genome. A gene is considered as a duplicate if it hits at least one other gene in the genome with the criteria of an E-value = 10^{-10} and an alignable region > 50% of the longer sequence. Otherwise, it is treated as a singleton.

5.5.6 Gene evolutionary rates

Following (Zhang and He 2005), we used the maximum likelihood method to estimate synonymous (d_S) and nonsynonymous (d_N) substitution rates of yeast genes by comparing the orthologous genes of *S. cerevisiae* and *S. bayanus*, which were identified by reciprocal best BLAST hits. The *PGL* information was obtained from a previous study (Wang and Zhang 2007), which used the parsimony principle to estimate the number of gene losses on the phylogeny of 12 fungi (*S. cerevisiae*, *S. bayanus*, *S. paradoxus*, *S. mikatae*, *Candida glabrata*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, *Yarrowia lipolytica*, *Neurospora crassa*, *Kluyveromyces waltii*, and *Schizosaccharomyces pombe*).

5.5.7 Protein domain identification

We downloaded the latest release (Release 20.27) of protein domain scan algorithm ProSite (Hulo et al. 2006) from ftp://ca.expasy.org/databases/prosite/, where an executable program and a compiled domain motif database were available. InterProScan (Mulder and Apweiler 2008) was downloaded from http://www.ebi.ac.uk/Tools/InterProScan/ with the current-release database, and was set up to run locally to identify protein domains.

5.6 ACKNOWLEDGEMENTS

We thank Meg Bakewell, Wenfeng Qian, and three anonymous reviewers for valuable comments.

Figure 5.1 Frequency distributions of Spearman's rank correlation coefficient ρ between gene importance (i.e., fitness reduction upon gene deletion) and evolutionary rate across many conditions. Gene importance is measured by experiments in 418 lab conditions (panels **A-C**), predicted by FBA for enzyme genes in 10^4 simulated nutritional conditions (**D-F**), or predicted by MOMA for enzyme genes in the same 10^4 conditions (**G-I**). Gene evolutionary rate is measured by nonsynonymous substitution rate d_N (**A**, **D**, **G**), nonsynonymous/synonymous rate ratio d_N/d_S (**B**, **E**, **H**), or propensity for gene loss *PGL* (**C, F**, **I**). The yellow arrow in each panel indicates the observed correlation using gene importance values experimentally determined in the YPD medium and the red arrow indicates the strongest correlation across the conditions examined. The numbers of genes used are 3999 for panels **A-C**, 478 for panels **D**, **E**, **G**, and **H**, and 546 for panels **F** and **I**. The gene number is lower than 546 for panels **D**, **E**, **G**, and **H**, because some *S. cerevisiae* genes do not have orthologs in *S. bayanus*. The yellow arrow is on the left-hand side of the red arrow in panels **G**, **H**, and **I**, because, under all simulated conditions, MOMA-predicted fitness values have weaker correlations with the evolutionary rates than that observed under YPD.

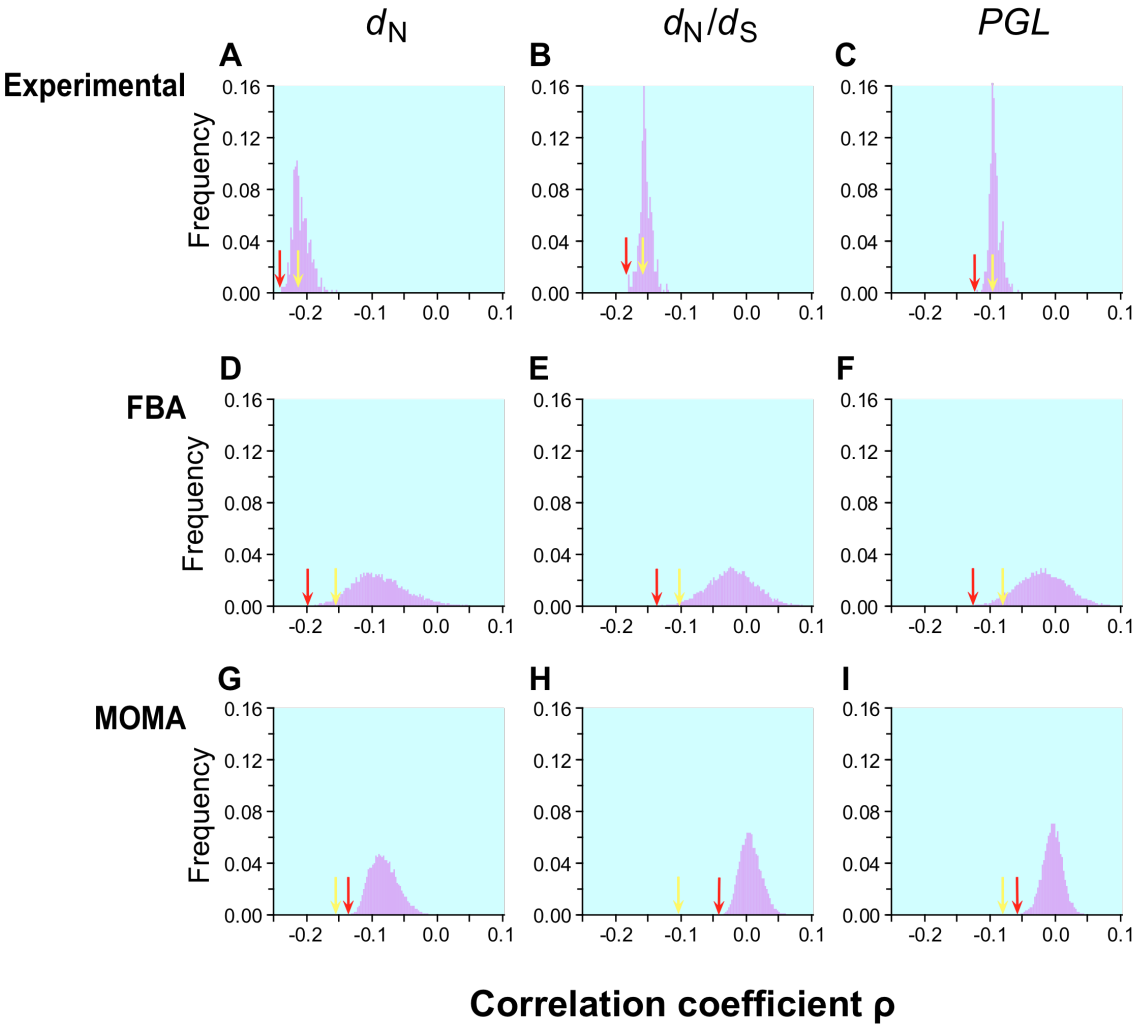


Figure 5.2 Always-essential enzyme genes do not evolve significantly slower than sometimes-essential and always-nonessential ones, regardless of the measure of the evolutionary rate. Error bars show one standard error. *P*-values are from Mann-Whitney *U* test between groups of genes. The numbers of genes used are 478 for panels **A** and **B** and 546 for panel **C**.

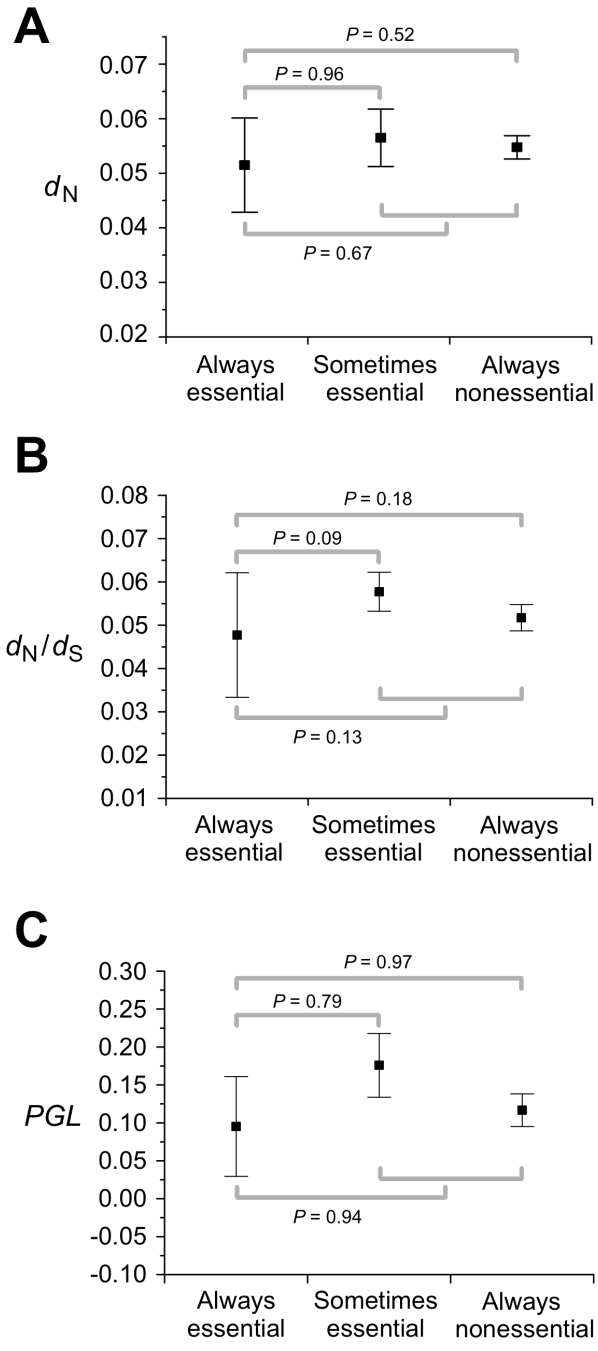


Figure 5.3 Relationship between the importance (β) and functional density (α) of genes. Gene importance is measured by the experimentally determined fitness reduction upon gene deletion in YPD. Functional density is measured by the proportion of amino acid sites within functional domains predicted by (**A**) the ProSite algorithm or (**B**) InterProScan. In InterProScan, a site is considered a domain site when predicted by at least two algorithms. A total of 5936 yeast genes are used in this analysis.

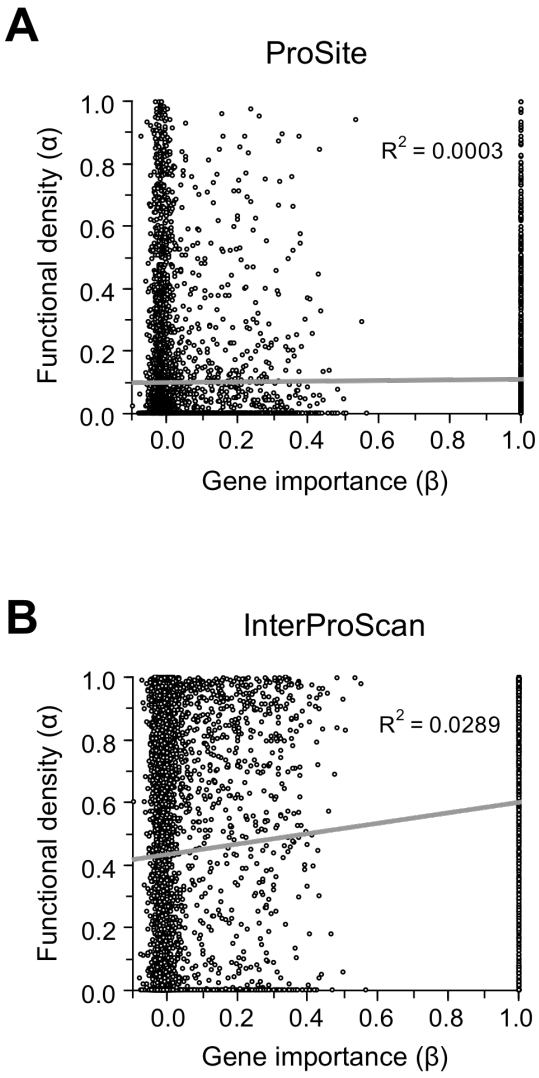


Figure 5.4 Predictability of the principle of slower evolution of more important genes. (**A**) Predictions of relative gene importance are more likely to be correct when the difference in evolutionary rate between the two genes under comparison increases. Rank difference shows the minimal fraction of genes in the genome whose ranks in d_N are between the two genes under comparison. Gene importance is measured by the amount of fitness reduction caused by the deletion of the gene under YPD. For each rank difference criterion, 100,000 random pairs of genes satisfying the criterion are used to estimate the prediction accuracy. (**B**) Extremely conserved genes (measured by d_N) are more likely to be essential. For the 418 lab conditions, the average proportion of essential genes among the 418 lab conditions and its standard error are shown.

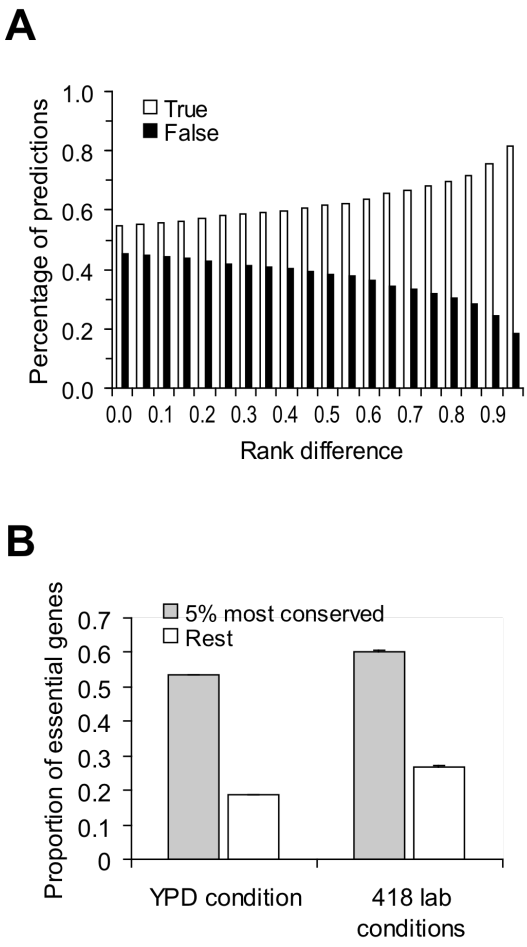


Table 5.1 Strongest correlations between gene evolutionary rate and importance measured at different conditions.

Conditions (methods)	Measures of evolutionary rate		
	d_N	d_N/d_S	*PGL*
418 individual lab conditions (experimental)	-0.2082[a] (1E-39[b])	-0.1520 (1E-21)	-0.1122 (1E-12)
Combined lab conditions (experimental)	-0.2187 (1E-43)	-0.1580 (1E-23)	-0.1185 (1E-13)
10,000 individual simulated conditions (FBA)	-0.1193 (0.009)	-0.0747 (0.14)	-0.0868 (0.04)
Combined simulated conditions (FBA)	-0.1252 (0.006)	-0.0767 (0.12)	-0.0937 (0.03)
10,000 individual simulated conditions (MOMA)	-0.1354 (0.003)	-0.0748 (0.13)	-0.0941 (0.03)
Combined simulated conditions (MOMA)	-0.1442 (0.002)	-0.0786 (0.12)	-0.1021 (0.02)

[a] Pearson's correlation coefficient

[b] *P*-value

5.7 REFERENCES

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2:727-738.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394.

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301-312.

Copley RR, Doerks T, Letunic I, Bork P. 2002. Protein domain analysis in the era of complete genomes. FEBS Lett 513:129-134.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102:14338-14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327-337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341-352.

Duarte NC, Herrgard MJ, Palsson BO. 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14:1298-1309.

Edwards JS, Covert M, Palsson BO. 2002. Metabolic Modeling of Microbes: the Flux Balance Approach. Environmental Microbiology 4:133-140.

Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. PLoS Genet 1:66-71.

Forster J, Famili I, Palsson BO, Nielsen J. 2003. Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics 7:193-202.

Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet 19:678-681.

Hillenmeyer ME, Fung E, Wildenhain J, et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320:362-365.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046-1049.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. 2006. The PROSITE database. Nucleic Acids Res 34:D227-230.

Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol 9:747-750.

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

Karp G. 2008. Cell and Molecular Biology. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-254.

Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131:151-156.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624-626.
Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.
Kimura M, Ohta T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A 71:2848-2852.
King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 164:788-798.
Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99:14878-14883.
Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:2229-2235.
Liao BY, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23:2072-2080.
Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci U S A 105:6987-6992.
Mulder NJ, Apweiler R. 2008. The InterPro database and tools for protein domain analysis. Curr Protoc Bioinformatics Chapter 2:Unit 2.7.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927-931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat. Rev. Genet. **7**:337-348.
Papp B, Pal C, Hurst LD. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429:661-664.
Pennacchio LA, Ahituv N, Moses AM, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499-502.
Plotkin JB, Fraser HB. 2007. Assessing the determinants of evolutionary rates in the presence of noise. Mol Biol Evol 24:1113-1121.
Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.
Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108-116.
Segre D, Vitkup D, Church GM. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99:15112-15117.
Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segre D. 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biol 9:R140.
Steinmetz LM, Scharfe C, Deutschbauer AM, et al. 2002. Systematic screen for human disease genes in yeast. Nat Genet 31:400-404.
Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. U S A 102:5483-5488.
Wang Z, Zhang J. 2007. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 3:e107.
Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem 46:573-639.

Wolf YI. 2006. Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. Trends Genet 22:354-357.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci 273:1507-1515.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338-345.
Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol 20:772-774.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

CHAPTER 6

Conclusions

6.1 CONCLUDING REMARKS

In this dissertation, I addressed several complex genetic and evolutionary phenomena by systems biology approaches. Using large-scale functional genomic data and network analysis methods, I uncovered the genome-wide patterns and genetic mechanisms of modularity, redundancy, and pleiotropy. Additionally, this systemic approach allows me to investigate the rate determinants of protein evolution from a new angle. Partly because of the use of new approaches, the conclusions from my studies challenge many common believes and deepen our understandings of the organization and evolution of some complex genetic systems.

In Chapter 2, I studied the modular organization of the protein interaction network. I showed that although the protein interaction network shows a much higher modularity than a random network with a similar network topology, the identified structural modules do not correspond to known functional units. In addition, I found that these structural modules are unconserved across species. These results contradict the common belief that modularity in cellular function may have arisen from the modularity in the underlying molecular interaction networks (Hartman, Garvik, and Hartwell 2001; Wagner, Pavlicev, and Cheverud 2007). Using computer simulation, I further demonstrated that the network modules can originate simply as a byproduct of the process of gene duplication. My results echo some recent studies of motifs in transcriptional regulatory networks showing that the structural motifs do not represent functional units, are not subject to natural selection, and might be the result of random gene duplication and mutation (Mazurie, Bottani, and Vergassola 2005; Cordero and Hogeweg 2006; Dwight Kuo, Banzhaf, and Leier 2006).

Biological systems are also known to be redundant, but the level of redundancy has never been quantitatively assessed at the systems level (Wagner 2005a; Wagner 2005b). In Chapter 3, I examined redundancies in the yeast and *E. coli* metabolic networks using flux balance analysis (FBA) (Price, Reed, and Palsson 2004). Although my definition of functional redundancy is stringent, the results showed that the amount of redundancy is surprisingly high. A more interesting and challenging question is how the redundancies are maintained in the system during evolution. Many biologists believe that redundancies are favored by natural selection (de Visser et al. 2003; Wagner 2005b). However, my direct test of the key prediction of the adaptive backup hypothesis of higher redundancy for more important functions did not produce positive results. My theoretical population genetic analysis and empirical metabolic network analysis demonstrated that the majority of the redundant reactions can be maintained if the species alternates among different nutritional environments, which is plausible for both *E. coli* and yeast. The remaining redundant reactions are largely explained by the pleiotropic effect. Together, these and a few other minor mechanisms explain the evolutionary preservation of 95% of the identified redundant reactions. Thus, I gave an answer to the long-standing puzzle of how redundant reactions are evolutionarily maintained. Although my analyses are limited to the metabolic network, the obtained biological principles and insights are likely to be applicable to redundancies in other biological systems, because all biological systems can be treated as complex networks.

In Chapter 4, I studied the genomic patterns of gene pleiotropy. In spite of its broad implications in genetics and evolutionary biology, the genomic patterns of pleiotropy are unclear (Carroll 2008; Tyler et al. 2009). I found that most genes have only low degree of pleiotropy and that the gene-trait relationship is highly modular. Furthermore, the quantitative pleiotropy data showed that genes affecting more traits tend to have larger per-trait phenotypic effects. Because of its central importance, pleiotropy has been extensively modeled, albeit with virtually no empirical basis (Waxman and Peck 1998; Orr 2000). My results from the analysis of five large datasets of three species with varying degrees of complexity indicate that general patterns of pleiotropy exist and that the observations are likely to be true for all eukaryotes. The “cost of complexity” conundrum, which is based on a previous model of pleiotropy, has long puzzled

evolutionary biologists (Orr 2000). Now I show that the conundrum disappears when correct parameters of pleiotropy are used. More interestingly, I found that the observed value of a key parameter of pleiotropy is in a narrow range that maximizes the optimal complexity, which suggests that pleiotropy not only allowed but may have also promoted the origin of complexity. Whether these pleiotropy patterns are the results of natural selection for evolvability or byproducts of other processes will likely become a stimulating question for future exploration.

In the past few years, it has been of great interest to identify nonrandom patterns in biological systems, because such nonrandomness is often interpreted as having functional significance and having been favored by natural selection (Alon 2003). While this may be true in some cases, a nonrandom pattern can also originate as a byproduct of other processes without having its own function or advantage. In the first section of my dissertation, modularity and redundancy, two genetic properties widely believed to be biologically important and adaptively selected, are added to this growing list of byproducts of other evolutionary processes. The three genomic features I studied, modularity, redundancy, and pleiotropy, are also inter-related in their contribution to genetic robustness and evolvability (Wagner 2005b; Lenski, Barrick, and Ofria 2006). Robustness measures the persistency of a genetic system when facing environmental perturbations and mutations. Redundancy is one of the most important mechanisms contributing to genetic robustness, as high redundancy ensures high robustness against mutations. My results suggest that the genetic robustness of metabolism is an evolutionary byproduct rather than something that has been enhanced directly by natural selection. On the other hand, evolvability measures the tendency of a genetic system to adapt to a new environment. My results from the pleiotropy study suggest that pleiotropy can increase the evolvability of an organism, because pleiotropy promotes the evolution of complexity to some degree. The prevailing view is that genetic robustness constrains evolvability, because robust systems are resistant to changes and thus are less likely to evolve new traits (Lenski, Barrick, and Ofria 2006). By contrast, my studies imply that genetic robustness actually enhances evolvability. First, the genetic robustness of metabolic networks directly resulted from the requirement of the organisms to survive in multiple different environments. Second, it is interesting to note that functional

redundancy in metabolic networks can also result from gene pleiotropy, which is positively associated with evolvability. Third, although I found that modularity in protein interaction networks lacks biological significance, modularity in metabolic networks and gene pleiotropic networks can enhance both the robustness and evolvability of the system.

In the second section and my last major chapter, Chapter 5, I applied the systems biology approach to the study of protein evolution. It is commonly believed that functionally more important genes evolve more slowly than less important ones (Karp 2008). However, empirical data revealed only weak negative correlations between gene importance and evolutionary rate (Wall et al. 2005; Zhang and He 2005). After surveying 10,000 different nutritional conditions using FBA, I showed that neither the lab-nature mismatch nor a potentially biased among-gene distribution of functional density explains the observed weakness of the correlation between gene importance and evolutionary rate. I thus conclude that the weakness is factual rather than artifactual and suggest that it is likely due to the presence of many rate-determinants that are independent from gene importance. My results caution molecular biologists from predicting relative functional importance of genes directly from their relative levels of evolutionary conservation. Nevertheless, the demonstration that the principle of slower evolution of more important genes does have some predictive power when genes with vastly different evolutionary rates are compared may explain why the principle can be practically useful despite the weakness of the correlation. In particular, substantial amount of comparative genomic work has successfully identified functional non-coding sequences based on their extremely low rates of nucleotide substitution (Boffelli et al. 2003; Pennacchio et al. 2006). However, it remains to be seen whether non-coding sequences exhibit a stronger correlation between importance and evolutionary rate.

On the other hand, for all of my studies, I acknowledge that the results and interpretations are dependent on the quality of the datasets. Because high throughput functional genomics techniques are still at the early stage of development, and in particular, the protein-protein interaction data are highly incomplete and contain large number of false positives, my conclusions will need to be further verified when more and better quality datasets are available.

In summary, systems biology is a relatively new field. However, I was able to apply the systems concepts and approaches to address some of the most fundamental yet difficult questions in evolutionary biology. I believe that my results provide deepened understandings and fresh perspectives on these and other fundamental characteristics of life. More importantly, they offer brightened prospects for systems approaches to these questions. The continuing generation of large-scale biological data and development of new high-throughput techniques are going to provide more comprehensive measures of biological systems. Therefore, I believe that systems approaches will become more powerful in explaining the genetic and evolutionary mechanisms of life.

6.2 REFERENCES

Alon U. 2003. Biological networks: the tinkerer as an engineer. Science 301:1866-1867.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25-36.

Cordero OX, Hogeweg P. 2006. Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23:1931-1936.

de Visser JA, Hermisson J, Wagner GP, et al. 2003. Perspective: Evolution and detection of genetic robustness. Evolution Int J Org Evolution 57:1959-1972.

Dwight Kuo P, Banzhaf W, Leier A. 2006. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. Biosystems 85:177-200.

Hartman JL, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Karp G. 2008. Cell and Molecular Biology. Hoboken, New Jersey: John Wiley & Sons, Inc.

Lenski RE, Barrick JE, Ofria C. 2006. Balancing robustness and evolvability. PLoS Biol 4:e428.

Mazurie A, Bottani S, Vergassola M. 2005. An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6:R35.

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13-20.

Pennacchio LA, Ahituv N, Moses AM, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499-502.

Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays 31:220-227.

Wagner A. 2005a. Distributed robustness versus redundancy as causes of mutational robustness. Bioessays 27:176-188.
Wagner A. 2005b. Robustness and Evolvability in Living Systems. Princeton, NJ: Princeton University Press.
Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nat Rev Genet 8:921-931.
Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102:5483-5488.
Waxman D, Peck JR. 1998. Pleiotropy and the preservation of perfection. Science 279:1210-1213.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

APPENDICES

APPENDIX A

Mathematical Analysis of Pleiotropic Scaling

A.1 GENES AFFECTING MORE TRAITS HAVE LARGER PER-TRAIT EFFECT

Comparing two genes both with normal distributions of effect sizes but with different standard deviations, here we prove mathematically that the gene with the larger standard deviation affects more traits (when an effect-size cutoff is applied) and has on average a larger per-trait effect.

For a given gene, let $f(x)$ be the probability density function of the distribution of effect size, where effect size is measured by Z-scores. Based on empirical observations, we assume that $f(x)$ is a normal distribution with mean equal to 0 and standard deviation equal to t, or

$$f(x) = \frac{1}{\sqrt{2\pi}t} e^{-x^2/(2t^2)} . \quad \text{(A.1)}$$

Let $g > 0$ be the cutoff used to determine whether a trait is regarded as being affected significantly by the gene. The mean effect size per trait $F(t)$ can be expressed as

$$F(t) = \frac{\int_g^{+\infty} xf(x)\,dx}{\int_g^{+\infty} f(x)\,dx} = \frac{u(t)}{v(t)} \quad , \quad \text{(A.2)}$$

where $u(t) = \int_g^{+\infty} xf(x)\,dx$ and $v(t) = \int_g^{+\infty} f(x)\,dx$. Below, we prove that $F(t)$ is a monotonically increasing function of t, or $F'(t) > 0$.

We have

$$F'(t) = \frac{u'(t)v(t) - u(t)v'(t)}{v^2(t)} = \frac{v(t) - u(t)v'(t)/u'(t)}{v^2(t)} = \frac{A}{B} . \quad \text{(A.3)}$$

where $A = v(t) - u(t)v'(t) / u'(t)$ and $B = v^2(t)$. We can derive that

$$
\begin{aligned}
u(t) &= \int_g^{+\infty} \frac{x}{\sqrt{2\pi t}} e^{\frac{-x^2}{2t^2}} dx = \frac{t}{\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}}; \\
u'(t) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}} (1 + \frac{g^2}{t^2}); \\
v(t) &= \int_g^{+\infty} \frac{1}{\sqrt{2\pi t}} e^{\frac{-x^2}{2t^2}} dx = \frac{1}{\sqrt{2\pi}} \int_g^{+\infty} \frac{1}{t} e^{\frac{-x^2}{2t^2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{\frac{g}{t}}^{+\infty} e^{\frac{-y^2}{2}} dy, \text{ where } y = x / t; \\
v'(t) &= \frac{g}{t^2\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}}.
\end{aligned}
\tag{A.4}
$$

Therefore,

$$
\begin{aligned}
A &= \frac{1}{\sqrt{2\pi}} \int_{\frac{g}{t}}^{+\infty} e^{\frac{-y^2}{2}} dy - (\frac{t}{\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}})(\frac{a}{t^2\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}}) / [\frac{1}{\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}} (1 + \frac{g^2}{t^2})] \\
&= \frac{1}{\sqrt{2\pi}} \int_{\frac{g}{t}}^{+\infty} e^{\frac{-y^2}{2}} dy - \frac{gt}{(g^2 + t^2)\sqrt{2\pi}} e^{\frac{-g^2}{2t^2}} \\
&= \frac{1}{\sqrt{2\pi}} (\int_h^{+\infty} e^{\frac{-y^2}{2}} dy - \frac{h}{1 + h^2} e^{-h^2/2}), \text{ where } h = g / t.
\end{aligned}
\tag{A.5}
$$

Because it can be shown that

$$
\begin{aligned}
\int_h^{+\infty} e^{-y^2/2} dy &= \frac{h^2}{1 + h^2} \int_h^{+\infty} (1 + \frac{1}{h^2}) e^{-y^2/2} dy \\
&> \frac{h^2}{1 + h^2} \int_h^{+\infty} (1 + \frac{1}{y^2}) e^{-y^2/2} dy \\
&= -\frac{h^2}{1 + h^2} \int_h^{+\infty} d(\frac{1}{y} e^{-y^2/2}) \\
&= \frac{h}{1 + h^2} e^{-\frac{h^2}{2}}
\end{aligned}
\tag{A.6}
$$

A is positive. Because B is also positive, $F'(t)$ is positive. In other words, $F(t)$ is a monotonically increasing function of t.

Let N be the total number of traits considered. Then the number of traits affected by a gene is $n(t) = Nv(t)$. Because $v'(t)$ is positive, n is a monotonically increasing

function of t. Thus, both $F(t)$ and $n(t)$ increase with t. In other words, when t is larger, both the number of affected traits and the mean effect size increase, which creates the phenomenon of larger per-trait effect sizes for genes affecting more traits. Although in the above proof only traits with Z-scores larger than a positive cutoff g are considered to be affected by a gene, the result is the same when traits with Z-scores smaller than a negative cutoff g are considered to be affected, because $f(x)$ is symmetrical to 0. Thus, when all traits with absolute Z-scores larger than a cutoff $g > 0$ are considered to be affected, which is what we did in actual data analysis, the above proof is also valid.

Note that our proof assumes that we use a constant cutoff $g > 0$ for all genes. In the actual data analysis, the cutoff may vary for different genes when the same false discovery rate is used to determine the cutoff. However, the small variation in cutoff apparently did not affect the general trend of larger per-trait effect sizes for genes affecting more traits.

A.2 EXISTENCE OF NON-ZERO OPTIMAL PLEIOTROPY

Let T_E be the total phenotypic effect size of a mutation measured by the Euclidian distance and n be the degree of pleiotropy (or organismal complexity). Here we prove that when the exponent $b > 0.5$ in the scaling relationship of $T_E = an^b$, the highest adaptation rate occurs at an intermediate n. Based on Orr (Orr 2000), the adaptation rate of a population is

$$U(n) = \frac{dw}{dt} = -\frac{4kT_E^2}{n} Mw \ln w = -4ka^2 n^{2b-1} Mw \ln w, \qquad (A.7)$$

where k is a positive constant dependent on population size and mutation rate, $0 < w < 1$ is the current mean fitness of the population, $M = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} (y-x)^2 e^{\frac{-y^2}{2}} dy$, and $x = \frac{T_E\sqrt{n}}{2\sqrt{-2\ln w}} = \frac{an^{b+0.5}}{2\sqrt{-2\ln w}}$. We can show that

$$
\begin{aligned}
U'(n) &= \left(\frac{-4ka^2 w \ln w}{\sqrt{2\pi}}\right)\frac{d[n^{2b-1}(\sqrt{2\pi}M)]}{dn} \\
&= \left(\frac{-4ka^2 w \ln w}{\sqrt{2\pi}}\right)[(2b-1)n^{2b-2}(\sqrt{2\pi}M) + n^{2b-1}\frac{d(\sqrt{2\pi}M)}{dx}\cdot\frac{dx}{dn}] \\
&= \left(\frac{-4ka^2 n^{2b-2} w \ln w}{\sqrt{2\pi}}\right)[(2b-1)(\sqrt{2\pi}M) + n\frac{d(\sqrt{2\pi}M)}{dx}\cdot\frac{x}{n}(b+\frac{1}{2})] \\
&= \left(\frac{-4ka^2 n^{2b-2} w \ln w}{\sqrt{2\pi}}\right)[(2b-1)(\sqrt{2\pi}M) + x(b+\frac{1}{2})\frac{d(\sqrt{2\pi}M)}{dx}].
\end{aligned}
\tag{A.8}
$$

It can be shown by Maxima (http://maxima.sourceforge.net/), a computer algebra system, that

$$
\begin{aligned}
\frac{d(\sqrt{2\pi}M)}{dx} &= \frac{d[-xe^{-x^2/2} + \sqrt{\frac{\pi}{2}}(1+x^2)Erfc(\frac{x}{\sqrt{2}})]}{dx}, \\
&= -2e^{-x^2/2} + \sqrt{2\pi}xErfc(\frac{x}{\sqrt{2}})
\end{aligned}
\tag{A.9}
$$

where $Erfc(x)$ is the complimentary error function:

$$
Erfc(x) = 1 - Erf(x) = \frac{2}{\sqrt{\pi}}\int_x^{+\infty} e^{-t^2}\,dt\,. \tag{A.10}
$$

Combining (8) and (9), we have

$$
U'(n) = \frac{-4ka^2 n^{2b-2} w \ln w}{\sqrt{2\pi}}\,m(x), \tag{A.11}
$$

where

$$
m(x) = \sqrt{\frac{\pi}{2}}Erfc(\frac{x}{\sqrt{2}})(4bx^2 + 2b - 1) - 4bxe^{-x^2/2}\,. \tag{A.12}
$$

When $b = 0, m(x) = -\sqrt{\frac{\pi}{2}}Erfc(\frac{x}{\sqrt{2}}) < 0$. Thus, $U'(n) < 0$. This means that $U(n)$ decreases with n. Let n_{optimal} be the n with the largest U. Our results indicate that $n_{\text{optimal}} = 0$.

When $b = 0.5$, we can show that

$$
\begin{aligned}
m(x) &= 2x(\sqrt{\frac{\pi}{2}}xErfc(\frac{x}{\sqrt{2}}) - e^{\frac{-x^2}{2}}) \\
&= 2x(x\int_x^{+\infty} e^{-t^2/2}\,dt - e^{-x^2/2}) \\
&= 2x(\int_x^{+\infty} xe^{-t^2/2}\,dt - \int_x^{+\infty} te^{-t^2/2}\,dt) < 0.
\end{aligned}
\tag{A.13}
$$

The last step is true because x is biologically meaningful only when it is positive. This means that $U'(n) < 0$ and $U(n)$ decreases with n. In other words, $n_{optimal} = 0$.

When $b > 0.5$, we have

$$
\begin{aligned}
&m(0) = \sqrt{\frac{\pi}{2}}Erfc(0)(2b-1) - 0 > 0 \text{ and} \\
&m(2) = 4b[\sqrt{\frac{\pi}{2}}Erfc(\frac{x}{\sqrt{2}})(x^2 + \frac{1}{2} - \frac{1}{4b}) - xe^{-x^2/2}] \\
&< 4b[\sqrt{\frac{\pi}{2}}Erfc(\frac{x}{\sqrt{2}})(x^2 + \frac{1}{2}) - xe^{-x^2/2}] \\
&= 4b[\sqrt{\frac{\pi}{2}}Erfc(\sqrt{2})(\frac{9}{2}) - 2e^{-2}] = -0.0564b < 0
\end{aligned}
\tag{A.15}
$$

Because $m(x)$ is a continuous function, there exists $0 < x_{optimal} < 2$ for which $m(x) = 0$ and $U'(n) = 0$. As x moves from 0 to 2, U' changes from positive to negative, indicating that $x_{optimal}$ corresponds to a peak of U. The n value determined by $x_{optimal}$ thus corresponds to a peak of U and is positive. Thus, we proved that, when $b > 0.5$, there exists a positive $n_{optimal}$.

A.3 REFERENCES

Orr HA. 2000. Adaptation and the cost of complexity. Evolution 54:13-20.

APPENDIX B

Theoretical Analysis of Protein Evolutionary Rate

B.1 THEORETICAL EXPECTATIONS OF THE CORRELATION BETWEEN GENE IMPORTANCE AND EVOLUTIONARY RATE

In the simplest model of neutral evolution, a mutation is either completely neutral or null (Kimura 1983). Equation 5.1 in the main text describes the relationship between substitution rate and gene importance under this model. Figure A10A shows the cumulative probability distribution of the deleterious effect of random mutations on gene function, when 80% of mutations are null ($\alpha = 0.8$) and 20% are neutral. This distribution predicts an L-shape curve with a square angle for the relationship between d_N/d_S (i.e., k/u in Equation 5.1) and gene importance (Figure C.10B), under the assumption of $N_e = 10^7$ for yeast. We can then estimate the correlation between gene importance and evolutionary rate that can be observed under this model for a large set of genes (e.g., 1000), by considering the distribution of gene importance in yeast, errors in the estimation of gene importance, and errors in the estimation of d_N/d_S. We first classify yeast genes into 10 uniform bins according to their experimentally determined fitness upon deletion (i.e., fitness = 0.0-0.1, 0.1-0.2, 0.2-0.3, …>0.9). We then randomly sample 1000 genes from these bins to represent the genome. For each gene, a uniform random number in the fitness range of the bin to which the gene belongs is assigned to the gene as its true fitness. The true importance value of the gene is one minus the fitness. We assume that the measurement error for gene importance follows a normal distribution with the mean equal to 0 and standard deviation equal to 0.05. To generate an "observed" gene importance value, we randomly generate an error variable following the above distribution and add it to the true gene importance value assigned to the gene. The expected evolutionary rate for every gene can be calculated by Equation 5.1 using the

true importance value. We assume that the measurement error of d_N/d_S follows a normal distribution with the mean equal to 0 and standard deviation equal to 10% of the expected value. We similarly generated an "observed" d_N/d_S value for the gene. After generating these values for 1000 genes, we measure Spearman's rank correlation between "observed" d_N/d_S and "observed" gene importance. Our result shows that a significant correlation between "observed" gene importance and "observed" evolutionary rate is not expected under this simple neutral model (Figure C.10C).

However, the situation could be different when there is a large fraction of mutations that only slightly or moderately impair the function of a gene (Ohta 1973; Ohta 1992). Let the selection coefficient against a slightly/moderately deleterious mutation be $e\beta$, where e is the deleterious effect of the mutation on gene function and β is the importance of the gene as defined in the main text. Following a previous study (Hirsh and Fraser 2001), we assume that e follows a beta distribution. A beta distribution has two parameters, a and b. The mean of the distribution is $a/(a+b)$ and the variance is $ab/[(a+b)^2(a+b+1)]$. We here examine three sets of parameters for the beta distribution because the mean functional effect of slightly/moderately deleterious mutations in real genes is unknown. As in the previous section, we still assume that 20% of the mutations are completely neutral. We further assume that 20% of the mutations are null. The remaining 60% of the mutations follow the beta distribution (Figure C.10D, C.10G, and C.10J). As the functional effects of most non-null deleterious mutations get smaller, the L-shape curve for the relationship between d_N/d_S and gene importance starts to have a round angle. Following the same simulation strategy described above, we sample 1000 genes under this model for each parameter set. We found that for all three parameter sets examined, the correlation between gene importance and evolutionary rate is statistically significant (Figure C.10F, C.10I, and C.10L). In particular, when the beta distribution has the parameters of $a = 10^6$ and $b = 1$, this correlation can reach $\rho = -0.83$, suggesting that a strong correlation is theoretically possible under a more realistic model of evolution. Note that under the above parameter set, the deleterious functional effects of non-neutral, non-null mutations are concentrated between 10^{-7} and 10^{-5} (Figure C.10J). Because the mean gene importance is 0.3 in yeast, these mutations have a deleterious fitness effect between $N_e s = 0.3$ and 30 for an average gene, where s is the selection

coefficient against the mutation. This range of $N_e s$ fits the classic definition of slightly to moderately deleterious mutations. The model with the other two parameter sets has little slightly deleterious mutations but much more moderately deleterious mutations.

Because the proportion (x) of slightly/moderately deleterious mutations in yeast genes is unknown, we further examine the effect of x on the correlation between gene importance and evolutionary rate. To achieve this goal, we still assume that 20% of the mutations are neutral, but x varies between 0 and 0.8, whereas the proportion of null mutations equals 0.8-x. Similar to the above analysis, we simulate 1,000 genes and calculate their "observed" gene importance and "observed" evolutionary rate with different fractions of slightly/moderately deleterious mutations. Here, the parameter set of $a = 10^6$ and $b = 1$ is used for the beta distribution because this parameter set best reflects slightly and moderately deleterious mutations for yeast, as aforementioned. Our result shows that as long as there are at least 5-10% mutations that are slightly/moderately deleterious, the correlation between gene importance and evolutionary rate is substantial (Fig. C.10M).

B.2 REFERENCES

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046-1049.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96-98.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Ann Rev Ecol Evol 23:263-286.

APPENDIX C

Supplimentary Figures and Tables

Figure C.1 Lack of obvious correspondence between structural modules and protein cellular locations. In (**A**) and (**B**), each cellular location is indicated by a letter (A to U). In parentheses next to the letter is the percentage of proteins in the network that belong to that cellular location. Note that one protein may belong to more than one location. The circles next to the grid show the statistical significance of nonrandom distributions of genes of the same cellular locations across modules. Each small square in the grid shows the statistical significance of enrichment of a particular location in a module. For the circles and squares, significance levels are indicated by different colors. Panels (**C**) and (**D**) show the correlation between co-membership in structural modules and co-localization in cellular components for all pairs of proteins in the PIC and PEC networks, respectively. The circle size is proportional to the number of protein pairs. The line shows the linear regression and *r* is the correlation coefficient.

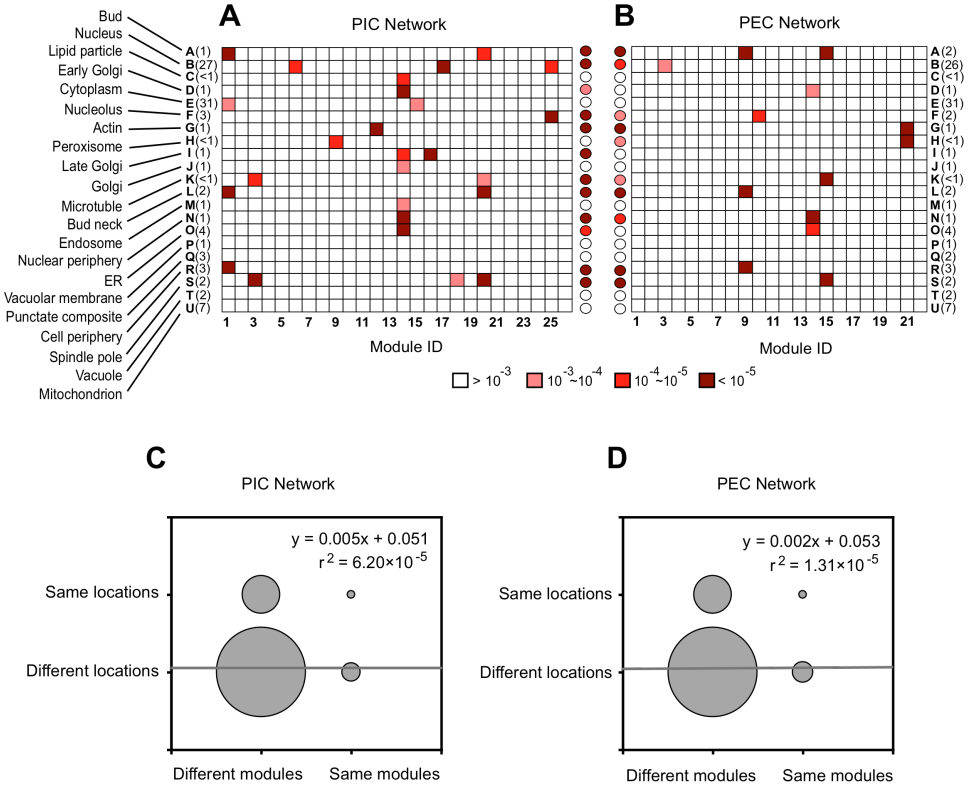


Figure C.2 Lack of evolutionary conservation between the yeast and nematode PPI modules. (**A**) The observed *NMI* (normalized mutual information) between yeast and nematode modules is not significantly different from the chance expectation. The bars show the distribution of *NMI* between the yeast and nematode modules when the yeast modules are randomly separated. (**B**) The observed CI_P (conservation index for pairs of proteins) between yeast and nematode modules is not significantly different from the chance expectation. The bars show the distribution of CI_P between the yeast and nematode modules when the yeast modules are randomly separated.

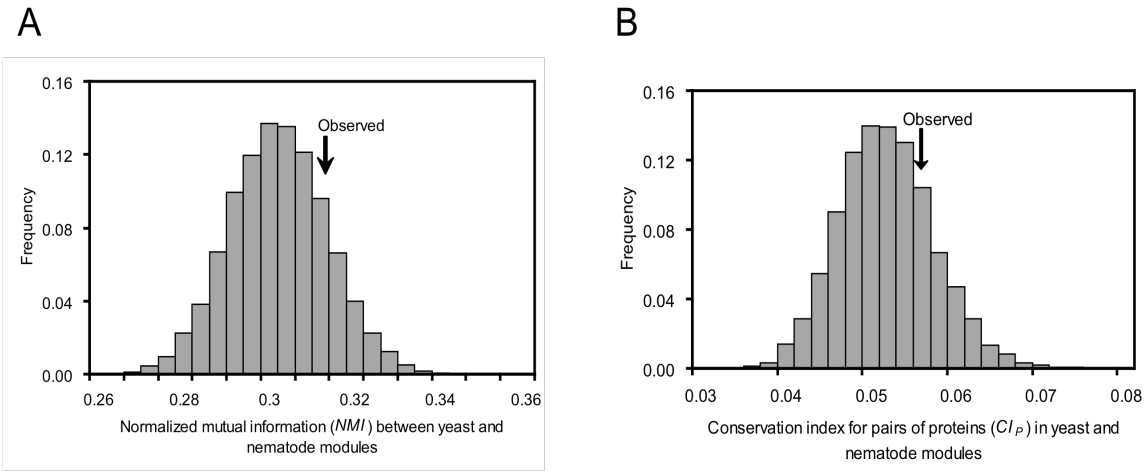


Figure C.3 Fractions of redundant, sometimes-essential, and always-essential reactions in different metabolic functional categories. Error bars show one standard error. In the parentheses following each function are the total numbers of *E. coli* and *S. cerevisiae* reactions belonging to this category, respectively. The vertical lines show the average fractions of the three types of reactions among all metabolic reactions. The distributions of the three types of reactions are not random among the functional categories, which to some degree can be intuitively explained. For instance, always-essential reactions are overrepresented in cell lipid membrane synthesis for both organisms. Lipids are important components of the biomass and essential for cell growth. However, cells lack transporters for most lipids. Consequently, most lipid synthesis reactions are always-essential. Always-essential reactions are also overrepresented in vitamin and cofactor metabolism in *E. coli*, but not in *S. cerevisiae*. This is because for *E. coli*, vitamin and cofactors are essential biomass components that can only be synthesized from intermediate metabolites, while for *S. cerevisiae*, vitamin and cofactors are not considered as biomass components, potentially due to the limitation of the current yeast metabolic model. Carbon and amino acid metabolites are important biomass constituents. However, most carbon metabolites and all amino acids can be transported from outside the cell. So, depending on the medium, reactions in carbon and amino acid metabolisms may or may not be essential. Therefore, we expect enrichment of sometimes-essential react ions in these functional categories, as is observed here. The reactions of cellular respiration are also highly redundant for both species, which is probably because both species can produce all biomass constituents under anaerobic conditions and cellular respiration reactions are only used when there is oxygen in the environment.

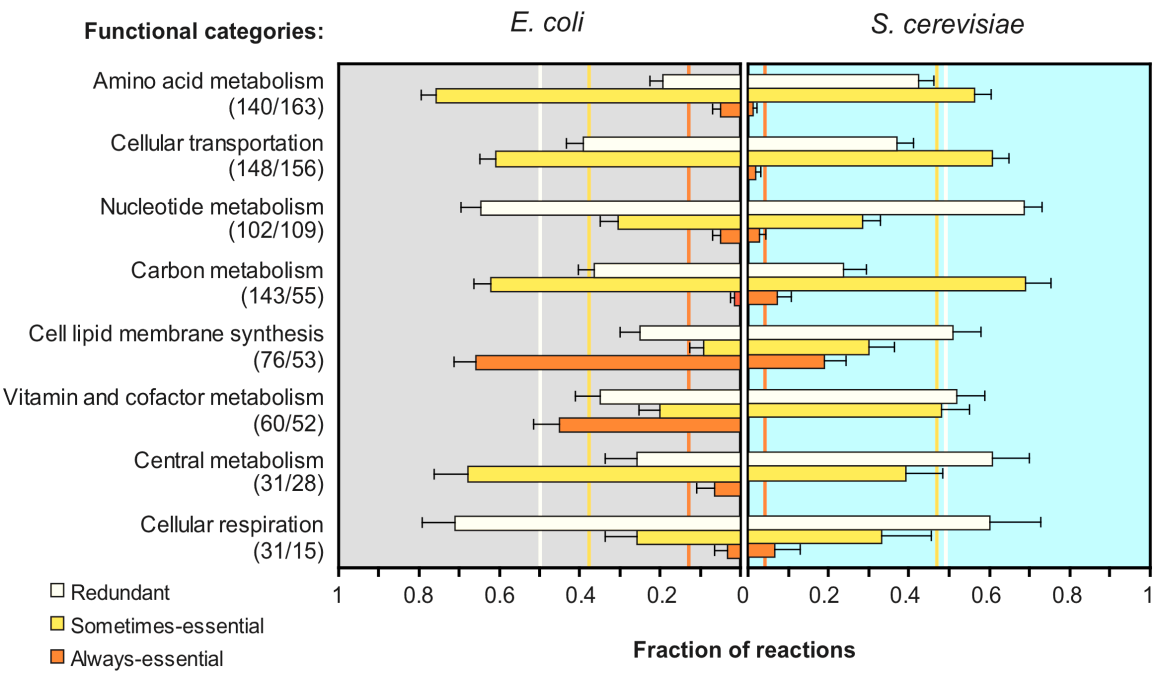


Figure C.4 Frequency distribution of the null mutation rate. The null mutation rate *u* is measured per gene per generation for the genes associated with (**A**) *E. coli* and (**B**) S. *cerevisiae* metabolic reactions concerned in this work. A total of 704 genes in *E. coli* and 542 genes in *S. cerevisiae* were used.

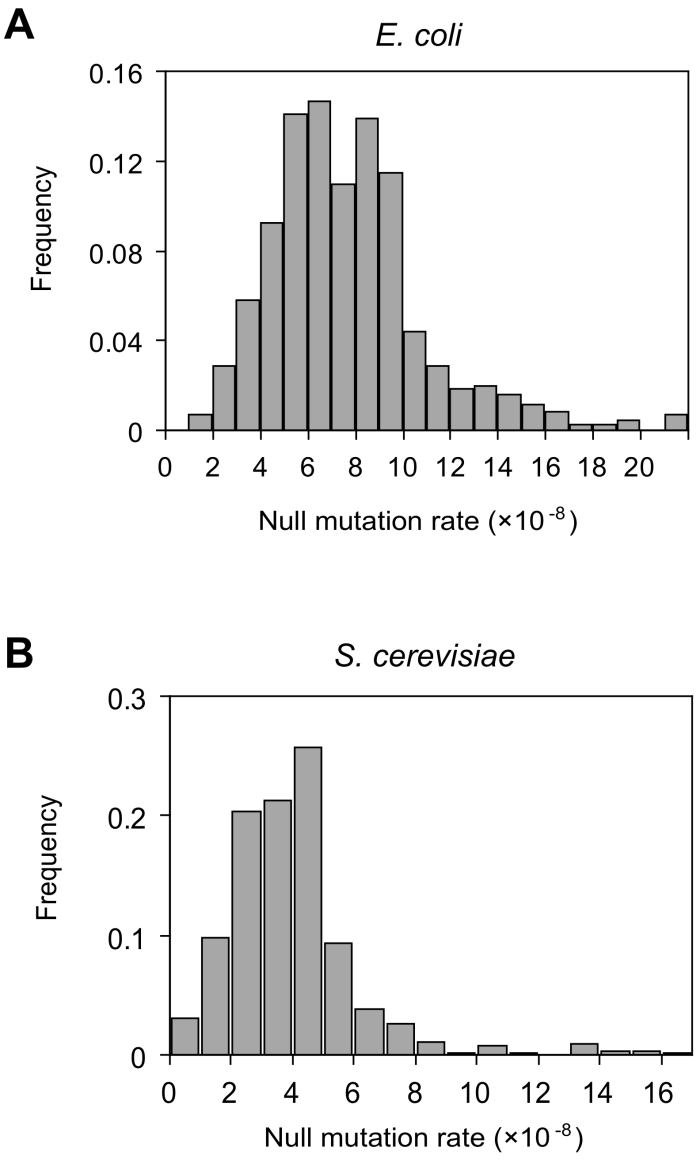


Figure C.5 Frequency distribution of the number of reactions in zero-redundancy metabolic networks. (**A**) *E. coli* and (**B**) *S. cerevisiae*. For each species, the main figure is obtained from examining 250 random zero-redundancy networks each in 10^3 conditions, while the inset is obtained from examining 10 random zero-redundancy networks each in 10^4 conditions.

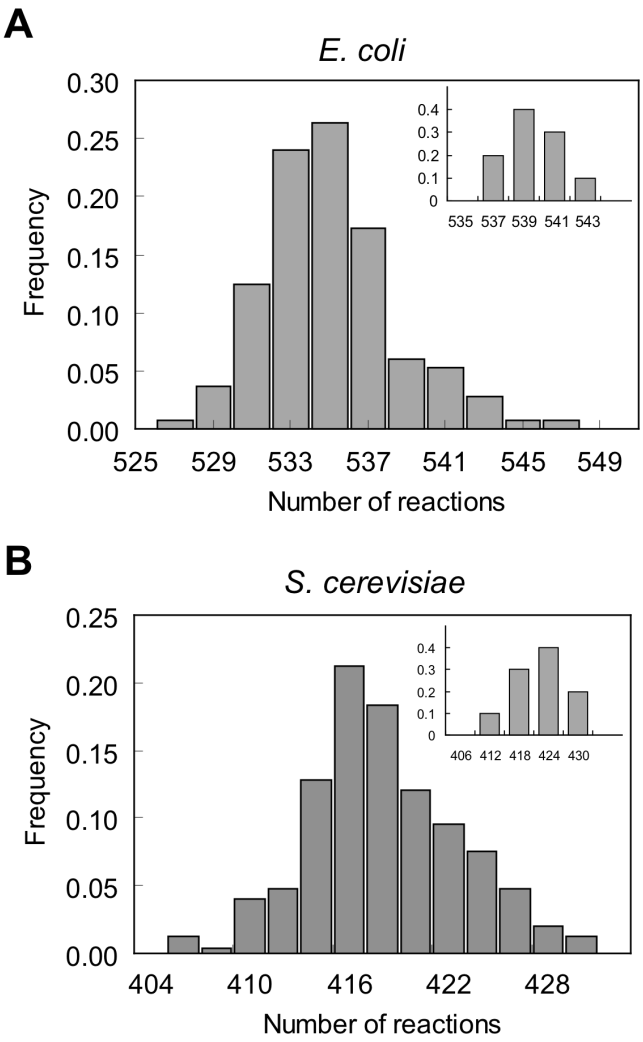


Figure C.6 Frequency distribution of the mean effect size (measured by *Z*-score) of a gene on the 279 morphological traits for all 4718 yeast genes. Note that the effect of a gene on a trait can be either positive or negative.

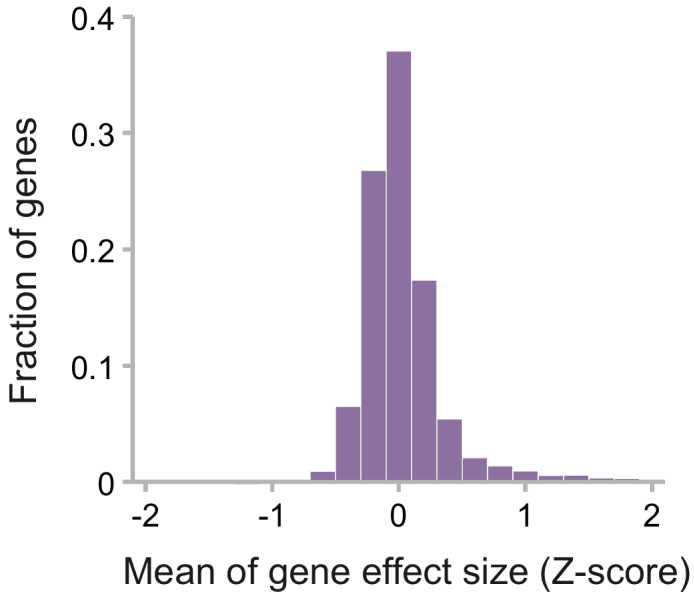


Figure C.7 The phenomenon of larger per-trait effects from genes affecting more traits is robust. Observed scaling relationships between the degree of pleiotropy and (**A**) Euclidean distance or (**B**) Manhattan distance, based on the yeast morphological pleiotropy data from which a random 50% of the traits are removed. The orange curve is the best fit to the power function whose estimated parameters are shown inside the panel. The numbers after ± show the 95% confidence interval for the estimated scaling exponent. Panels (**C**) and (**D**) are similar to panels (**A**) and (**B**) except that the dataset used is generated after the random removal of 90% of the traits. Panels (**E**) and (**F**) are similar to panels (**A**) and (**B**) except that the dataset used is generated by merging traits with a Pearson's correlation coefficient in gene effects greater than 0.7.

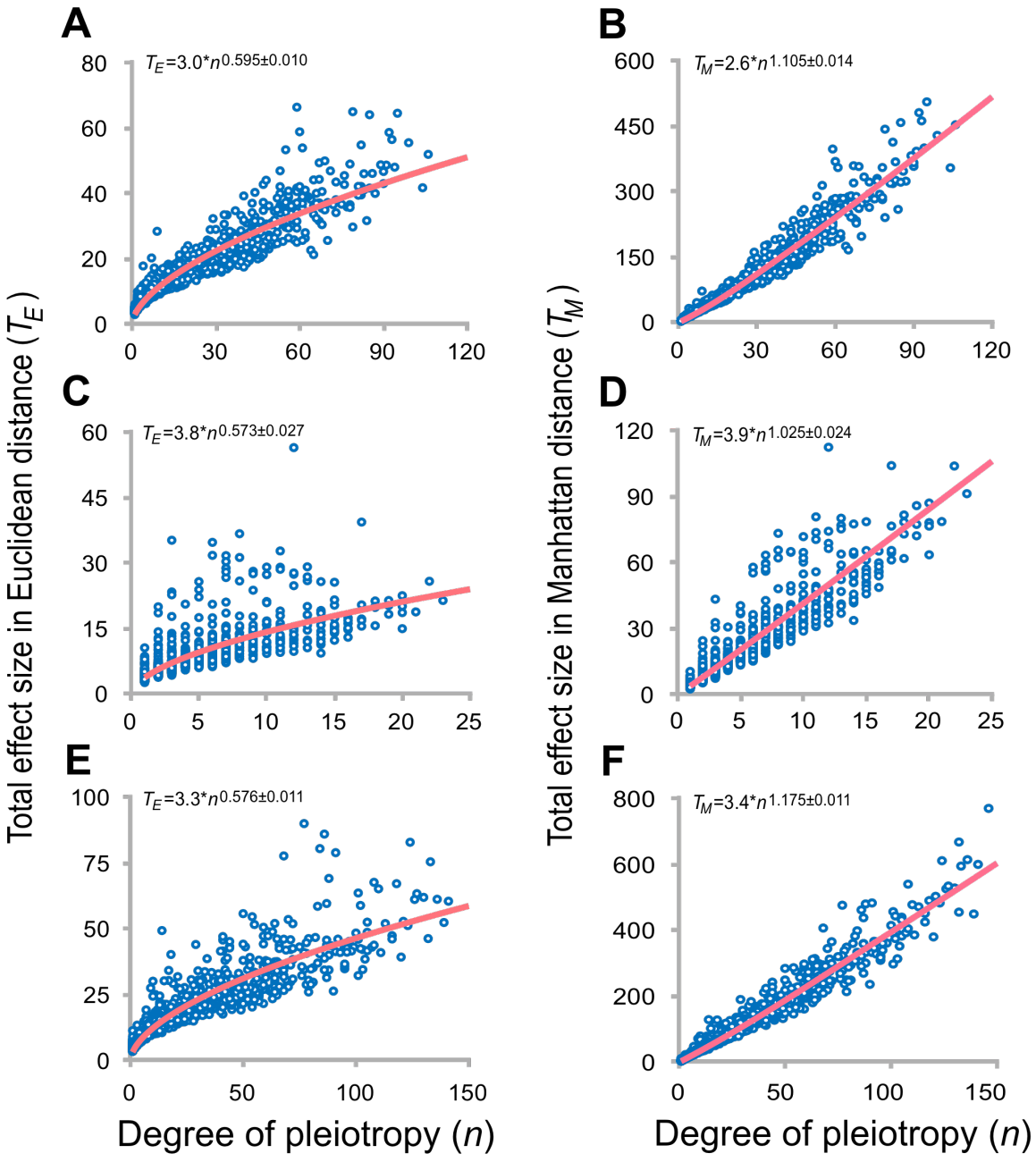


Figure C.8 Yeast morphological pleiotropy data analyzed using the conservative Bonferroni method to correct for multiple testing. (**A**) Genome-wide frequency distribution of the degree of gene pleiotropy. The numbers in the parentheses are the mean and media degrees of pleiotropy divided by the total number of traits. After the removal of genes that do not affect any trait and traits that are not affected by any gene, there are 2091 genes and 264 traits. (**B**) Observed modularity (blue arrow) and the distribution of modularity for 250 randomly rewired networks (red histograms). Observed scaling relationships between the degree of pleiotropy and the total effect size measured by (**C**) Euclidean distance or (**D**) Manhattan distance. The orange curve is the best fit to the power function whose estimated parameters are shown inside the panel. The numbers after ± show the 95% confidence interval for the estimated scaling exponent. R^2 indicates the square of the correlation coefficient.

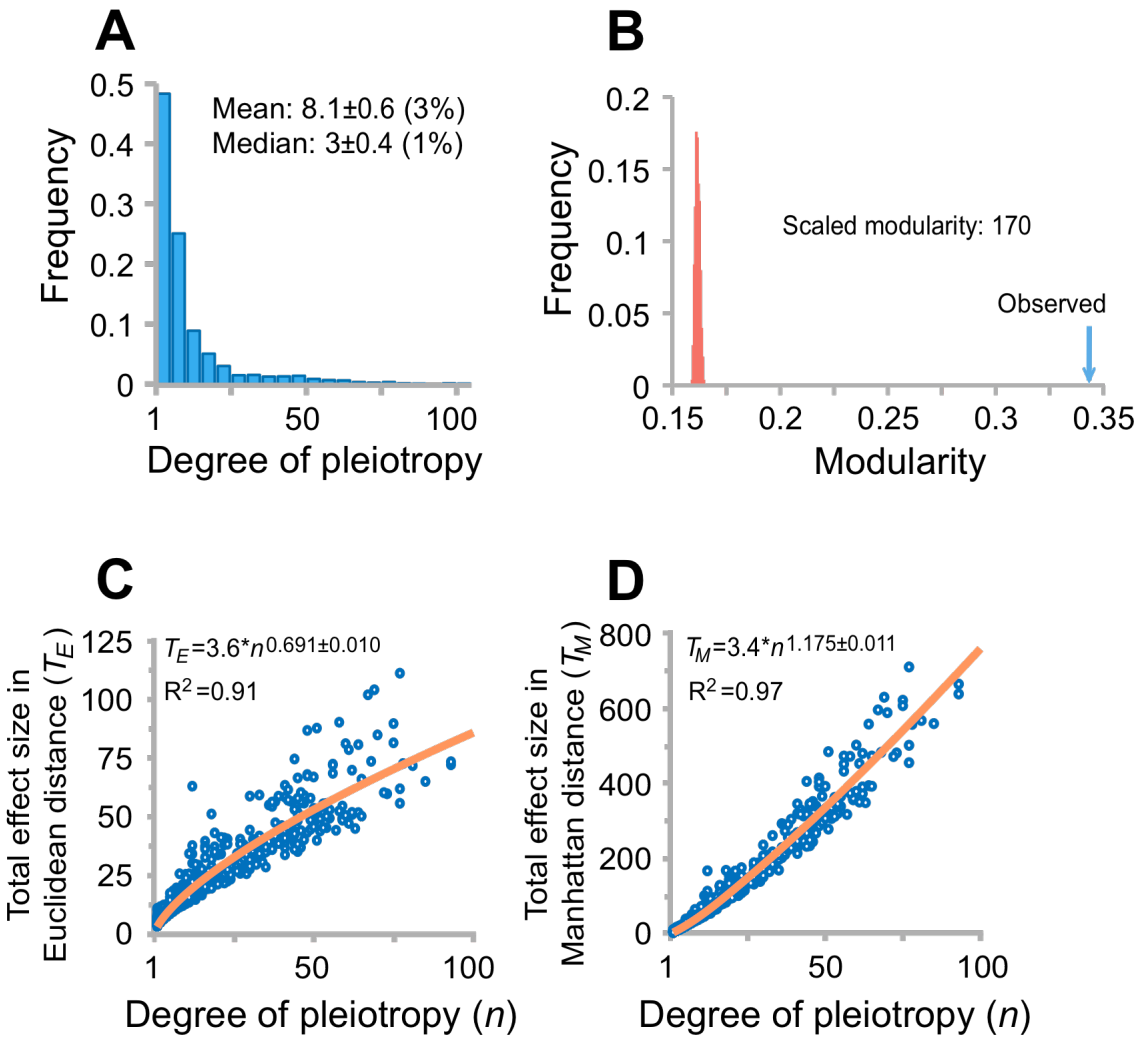


Figure C.9 Observed scaling relationships between the degree of pleiotropy and the total effect size. The total effect size is measured by (**A**) Euclidean distance or (**B**) Manhattan distance, when the effect sizes of all genes on all traits in the actual data are randomly shuffled. The orange curve is the best fit to the power function whose estimated parameters are shown inside the panel. The numbers after ± show the 95% confidence interval for the estimated scaling exponent.

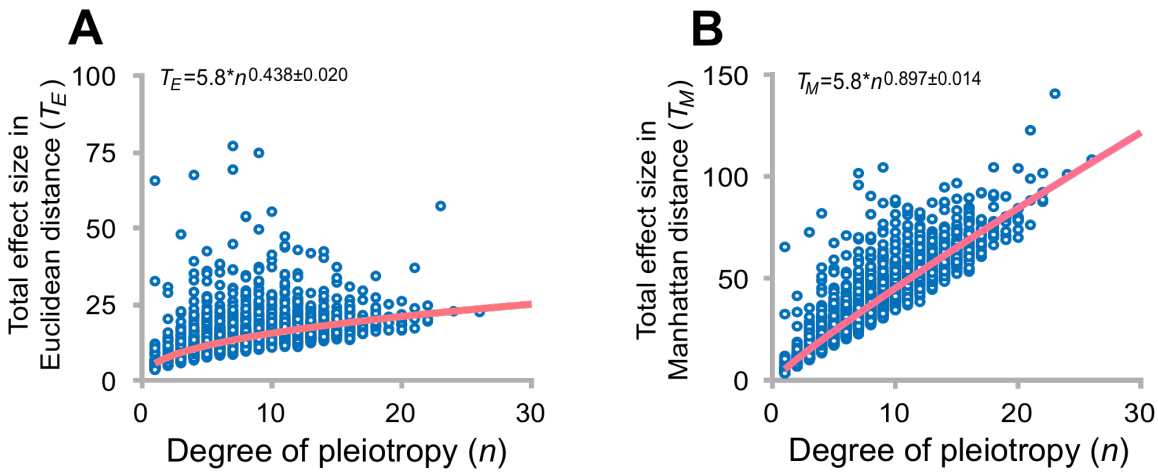


Figure C.10 Theoretical expectations of the correlation between gene importance and evolutionary rate under neutral and nearly neutral models. The cumulative probability functions of deleterious effects of random mutations on gene function are shown for the neutral model (**A**) and the nearly neutral model with three sets of parameters (**D**), (**G**), and (**J**). The expected relationships between d_N/d_S and gene importance under the four situations are shown in panels (**B**), (**E**), (**H**), and (**K**), respectively. When 1000 genes are simulated with measurement errors, the observed relationships between d_N/d_S and gene importance under the four situations are shown in panels (**C**), (**F**), (**I**), and (**L**), respectively, with the blue lines showing the linear regressions. The beta distribution that describes the deleterious functional effect of mutations used in panels (**D**), (**G**), and (**J**) all have the parameter $b = 1$. The parameter $a = 10^4$, 10^5, and 10^6, respectively, for (**D**), (**G**), and (**J**). Panel (**M**) shows Spearman's rank correlation coefficient under different fractions of slightly deleterious mutations.

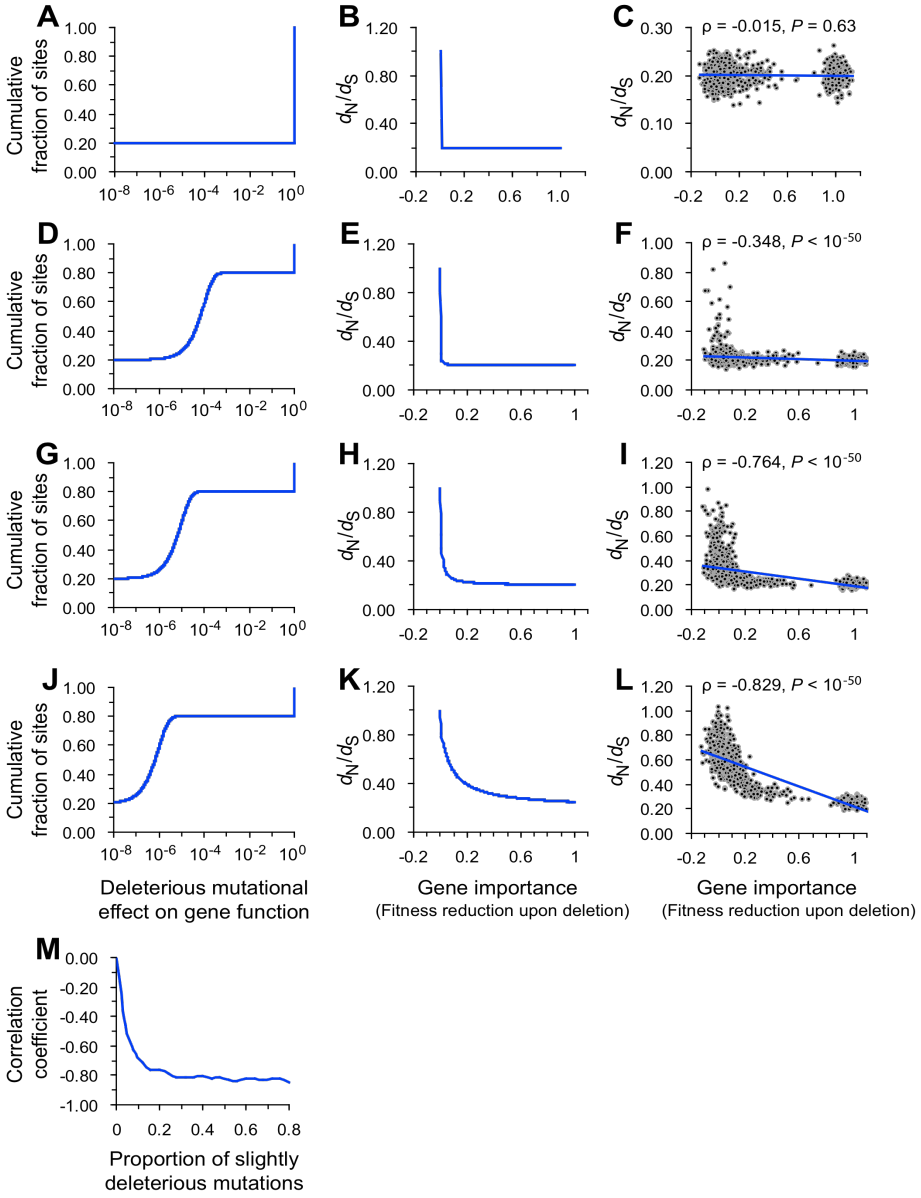


Figure C.11 Frequency distributions of Spearman's rank correlation coefficient ρ between gene importance and evolutionary rate across 10^5 simulated nutrient conditions. Gene importance is predicted by FBA. Gene evolutionary rate is measured by (**A**) nonsynonymous substitution rate d_N, (**B**) nonsynonymous/synonymous rate ratio d_N/d_S, or (**C**) propensity for gene loss *PGL*. The yellow arrow in each panel indicates the observed correlation using gene importance values experimentally determined in the YPD medium and the red arrow indicates the strongest correlation across the conditions examined.

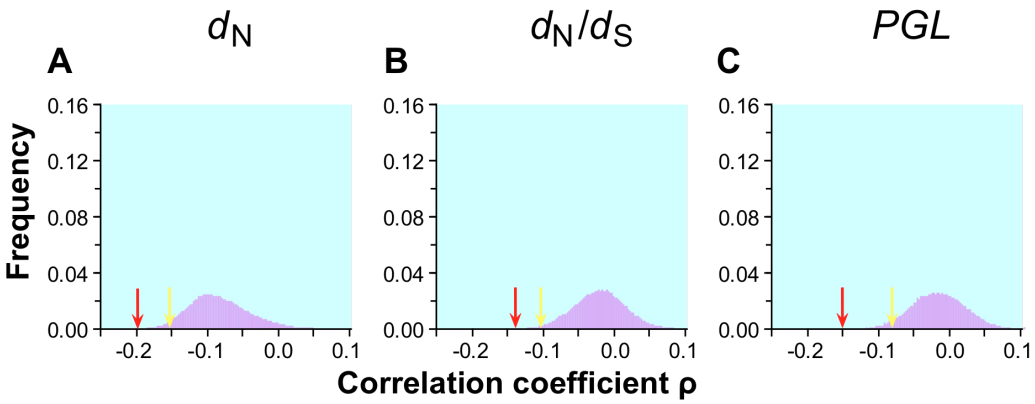


Table C.1 Summary statistics of the giant component in the random networks generated by gene duplication followed by subfunctionalization.

Networks	No. of nodes	No. of interactions	Modularity	Mean modularity of randomly rewired networks	Standard deviation of modularity in randomly rewired networks	Scaled modularity
1	212	1028	0.6717	0.4485	0.0077	29
2	186	1088	0.6001	0.3936	0.0063	33
3	155	864	0.5884	0.3979	0.0074	26
4	136	634	0.5917	0.4402	0.0091	17
5	205	1032	0.6193	0.4289	0.0064	30
6	185	860	0.6498	0.4605	0.0077	25
7	221	1194	0.6466	0.4208	0.0072	32
8	146	760	0.6525	0.4247	0.0093	25
9	148	754	0.5934	0.4044	0.0067	28
10	118	564	0.6164	0.4202	0.0091	22
Average	171.2	877.8	0.6230	0.4240	0.0077	26

Table C.2 Relative importance of redundant and non-redundant reactions in 10 simulated metabolic networks of *E. coli* and *S. cerevisiae*.

			Non-redundant reactions		
	Mean fitness reduction[a]	Standard error	Mean fitness reduction[a]	Standard error	
E. coli					
Simulated network #1	0.278	0.022	0.643	0.024	2.60E-22
Simulated network #2	0.254	0.023	0.611	0.023	1.60E-20
Simulated network #3	0.290	0.024	0.627	0.023	1.50E-19
Simulated network #4	0.290	0.025	0.625	0.024	6.20E-17
Simulated network #5	0.296	0.024	0.609	0.023	2.50E-13
Simulated network #6	0.314	0.024	0.638	0.024	3.10E-16
Simulated network #7	0.320	0.026	0.603	0.024	2.40E-12
Simulated network #8	0.317	0.025	0.608	0.024	1.90E-11
Simulated network #9	0.292	0.024	0.605	0.024	1.10E-14
Simulated network #10	0.282	0.025	0.582	0.023	3.50E-13
S. cerevisiae					
Simulated network #1	0.672	0.026	0.580	0.022	3.50E-03
Simulated network #2	0.669	0.026	0.566	0.022	1.50E-02
Simulated network #3	0.678	0.024	0.574	0.022	6.10E-03
Simulated network #4	0.705	0.025	0.556	0.022	7.27E-06
Simulated network #5	0.717	0.025	0.573	0.022	2.73E-05
Simulated network #6	0.649	0.026	0.572	0.022	1.37E-01
Simulated network #7	0.614	0.026	0.580	0.022	1.11E-01
Simulated network #8	0.671	0.027	0.579	0.022	1.31E-02
Simulated network #9	0.697	0.024	0.570	0.022	6.24E-04
Simulated network #10	0.665	0.026	0.567	0.022	1.90E-03

[a] Mean fitness reduction upon removal of the reaction, measured in zero-redundancy networks by FBA.

[b] Mann-Whitney U test.

Table C.3 Numbers (and percentages) of reactions that are always-essential, sometimes-essential, or redundant in 10^5 conditions examined.

	Total	Always-essential	Sometimes-essential	Redundant
E. coli				
All reactions	737	95 (13%)	366 (50%)	276 (37%)
Reactions with non-zero fluxes in aerobic glucose minimal medium	293	95 (33%)	153 (52%)	45 (15%)
Reactions with non-zero fluxes in anaerobic glucose minimal medium	288	95 (33%)	156 (54%)	37 (13%)
S. cerevisiae				
All reactions	632	24 (4%)	313 (49%)	295 (47%)
Reactions with non-zero fluxes in YPD medium	209	24 (11%)	112 (54%)	73 (35%)
Reactions with non-zero fluxes in aerobic glucose minimal medium	293	24 (8%)	207 (71%)	62 (21%)

Table C.4 Numbers of various types of redundant reactions in *E. coli* and *S. cerevisiae*.

	E. coli		*S. cerevisiae*	
Redundant reactions	f=0.99[a]	f=0.90[b]	f=0.99[a]	f=0.90[b]
Efficient reactions	64	61	89	56
Non-efficient, active reactions	158	160	166	176
Non-efficient, non-active reactions	54	55	40	63
Explained by pleiotropic effect	27	27	13	17
Reactions acquired by recent HGT[c]	6	6	0	0
FBA limitation	7	8	14	16
Unexplained reactions (genes[d])	14 (8)	14 (8)	13 (4)	30 (9)
Sum	276	276	295	295

[a] A reaction is regarded as indispensable if its removal renders FBA or MOMA predicted fitness lower than 0.99

[b] A reaction is regarded as indispensable if its removal renders FBA or MOMA predicted fitness lower than 0.90.

[c] HGT, horizontal gene transfer.

[d] The number of unexplained reactions that have known associated genes is given in parentheses.

Table C.5 Robustness of pleiotropy estimates.

Gene pleiotropy datasets	No. of traits	Mean pleiotropy	95% confidence interval of mean pleiotropy	Mean pleiotropy divided by no. of traits	Median pleiotropy	95% confidence interval of median pleiotropy	Median pleiotropy divided by no. of traits
Yeast morphological pleiotropy							
100% of traits used	279	21.6	[20.2, 23.1]	0.077	7	[6, 8]	0.025
50% of traits used	140	10.9	[9.9, 11.9]	0.078	4	[3, 4]	0.029
10% of traits used	28	2.2	[1.7, 2.6]	0.079	1	[0, 1]	0.036
Yeast environmetnal pleiotropy							
100% of traits used	22	2.4	[1.7, 3.2]	0.109	2	[1, 3]	0.091
50% of traits used	11	1.2	[0.7, 1.8]	0.109	1	[0, 1]	0.091
10% of traits used	2	0.2	[0.1, 0.5]	0.100	0	[0, 0]	0.000
Yeast physiological pleiotropy							
100% of traits used	120	1.8	[1.3, 2.5]	0.015	1	[1, 2]	0.008
50% of traits used	60	0.9	[0.5, 1.4]	0.015	1	[0, 1]	0.017
10% of traits used	12	0.2	[0.1, 0.4]	0.017	0	[0, 0]	0.000
Nematode pleiotropy							
100% of traits used	44	4.6	[3.7, 5.6]	0.105	4	[3, 6]	0.091
50% of traits used	22	2.3	[1.7, 3.0]	0.105	2	[1, 3]	0.091
10% of traits used	4	0.5	[0.2, 0.7]	0.125	0	[0, 1]	0.000
Mouse pleiotropy							
100% of traits used	308	8.2	[7.1, 9.3]	0.027	6	[5, 6]	0.019
50% of traits used	154	4.1	[3.3, 4.9]	0.027	3	[2, 3]	0.019
10% of traits used	31	0.8	[0.5, 1.2]	0.026	1	[0, 1]	0.032

Table C.6 Comparison between the observed genomic patterns of pleiotropy and assumptions made in the existing theoretical models of pleiotropy.

Features	Invariant total effect model[1]	Euclidean superposition model[2]	Observed genomic patterns
Proportion of traits affected by a gene	100%	100%	1% to 9%
Modularity of the gene-trait network	None	None	High
Distribution of effect size on a trait	Uniform	Normal	Normal
Among-gene variation in standard deviation of the effect-size distribution	Absent	Absent	Present
Total effect size	Constant	Increase with pleiotropy	Increase with pleiotropy
Per-trait effect size	Decrease with pleiotropy	Constant	Increase with pleiotropy
Scaling exponent b	0	0.5	0.6
Scaling exponent d	0.5	1	1.1
Degree of pleiotropy (complexity) that offers the highest adaptation rate	1	1	Intermediate level of pleiotropy

[1] Fisher (1930) and Orr (2000).

[2] Turelli (1985), Wagner (1988), Wagner (1989), and Waxman & Peck (1998).

Table C.7 High modularity of gene-trait networks.

Gene pleiotropy datasets	No. of traits	No. of genes	Modularity	Scaled modularity
Yeast morphological pleiotropy (all traits)	279	2449	0.204	36.8
Yeast morphological pleiotropy (random half of the traits)	140	1902	0.221	37.7
Yeast morphological pleiotropy (after the merge of related traits*)	197	2272	0.209	45.0
Yeast environmetnal pleiotropy (all traits)	22	774	0.440	35.1
Yeast environmetnal pleiotropy (random half of the traits)	11	448	0.579	13.6
Yeast environmetnal pleiotropy (after the merge of related traits*)	22	774	0.440	35.1
Yeast physiological pleiotropy (all traits)	120	1256	0.580	34.2
Yeast physiological pleiotropy (random half of the traits)	60	712	0.673	62.6
Yeast physiological pleiotropy (after the merge of related traits*)	118	1256	0.575	27.0
Nematode pleiotropy (all traits)	44	661	0.544	50.4
Nematode pleiotropy (random half of the traits)	22	579	0.473	48.2
Nematode pleiotropy (after the merge of related traits*)	44	661	0.544	50.4
Mouse pleiotropy (all traits)	308	4915	0.384	237.5
Mouse pleiotropy (random half of the traits)	154	4901	0.449	197.4
Mouse pleiotropy (after the merge of related traits*)	307	4915	0.376	202.8

*Traits with Pearson's correlation coefficient >0.7 are merged. Some datasets do not contain such correlated traits.

Table C.8 No significant difference in importance between *S. cerevisiae* genes with and without *S. bayanus* orthologs.

Genes	Mean fitness reduction upon gene deletion in YPD condition		*P*-values
	With orthologs	Without orthologs	
Singletons	0.288 (2668[a])	0.275 (883)	0.114[b]
Duplicates	0.171 (1331)	0.158 (1104)	0.625

[a] Number of genes in this category

[b] Mann-Whitney U test