FUNCTIONAL EVOLUTIONARY GENOMICS: YEAST AS A MODEL SYSTEM

by

Wenfeng Qian

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
2012

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor L. Lacey Knowles
Associate Professor Anuj Kumar
Associate Professor Patricia Wittkopp
Assistant Professor Barry L. Williams, Michigan State University

---

2012

DEDICATION

This thesis is dedicated

To

My parents and my wife

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my advisor Prof. George Zhang for giving me an opportunity to work with him and for his continual guidance, great encouragement, and strong support in all aspects of my graduate study. He is the type of scientist that I want to be in the future.

I would like to thank members in the Zhang lab. They are a group of great people that I love to work with every day. Some of them came to the lab earlier than I did, e.g., Meg Bakewell, Soochin Cho, Wendy Grus, Xionglei He, Ben-Yang Liao, Ondrej Podlaha, Peng Shi, Xiaoxia Wang, and Zhi Wang. They helped me in almost every aspect of my research. They especially took care of me in the first few years when I came to the U.S. from a different country with a different language. Some lab members came to the lab later than I did, e.g., Xiaoshu Chen, Wei-Chin Ho, Chuan Li, Ying Li, Calum Maclean, Chungoo Park, Seong Hwan Park, Nathaniel Pearson, Huailiang Xu, Jinrui Xu, Jianrong Yang, Zhihua Zhang, and Huabin Zhao. They also helped me significantly in my research, provided important comments, and supported me all the time. In particular, I would like to thank my undergraduate associates, Edwin Chan and Che Xiao. Both participated in my research and made significant contributions. I enjoy my six years in the Zhang Lab and cherish my experience here.

I thank my committee members, Profs. Lacey Knowles, Anuj Kumar, Barry Williams, and Trisha Wittkopp, for their insightful and critical comments on my research. I especially appreciate the time Barry spent on traveling to Ann Arbor.

I also thank many colleagues at University of Michigan who are not members of the Zhang lab. This is a very long list of people who have helped me in my research: Arielle Cooley, Brian Metzger, Dan Chang, Di Ma, Gizem Kalay, Huateng Huang, Ju Huang, Jun Ma, Joesph Coolon, Jonathan Gruber, Kai Mao, Nike Bharucha, Qixin He, and Ya Yang.

I would like to thank faculty members, staff, and students in the Department of Ecology and Evolutionary Biology at University of Michigan, for giving me the chance to study here, for establishing this great research environment, and for all the support.

My research is funded by departmental fellowships (Block grant, Edwin H. Edwards award, and Peter Olaus Okkelberg Award) and Rackham graduate school fellowships (Rackham graduate student research grant and Rackham Predoctoral Fellowship), in addition to U.S. National Institutes of Health research grants to my advisor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

A long-standing and central question in evolutionary biology is the molecular genetic basis of phenotypic evolution (Maynard Smith 1998). This genotype-phenotype map is essential for comprehending evolutionary processes, including evolutionary rates and trajectories (Weinreich et al. 2006; Poelwijk et al. 2007). The genotype-phenotype map, however, is complex and has been difficult to discern.

With more and more genomes sequenced in the last two decades, many genes have been identified by sequence analysis to have been under positive Darwinian selection. For example, genes under positive selection in human and chimpanzee evolution have been identified by analyzing the two genome with the use of the macaque genome as an outgroup (Bakewell et al. 2007). These positively selected genes potentially underlie the morphological, physiological, and behavioral differences between human and chimpanzee. Other studies aimed to estimate the fraction of substitutions that are under positive Darwinian selection (Fay et al. 2002; Smith and Eyre-Walker 2002), but the selective agents as well as the phenotypes affected are typically unknown.

To understand the phenotypic effects of the detected positive selection, a number of authors introduced mutations individually and tested their functional effects. For example, one study investigated the effects of amino acid substitutions on the catalytic activities following the duplication of a digestive ribonuclease gene in colobine monkeys (Zhang et al. 2002). Another study examined the evolutionary paths of an antibiotic resistant gene by reconstructing all possible intermediate mutants followed by an investigation of their functional effects (Weinreich et al. 2006). Thornton and colleagues analyzed the evolution of vertebrate hormone receptors and examined how the amino acid replacements in evolution have impacted the structure and function of the receptors (Ortlund et al. 2007). Several other studies also shed light on the effects of mutations on

protein function, gene expression, or organismal fitness (Patwardhan et al. 2009; Lind et al. 2010; Loh et al. 2010; Shultzaberger et al. 2010; Hietpas et al. 2011; Cuevas et al. 2012; Gruber et al. 2012). However, these studies are based on one to a few genes and thus do not provide a genomic perspective.

My study aims to understand at the genomic scale how variations at the DNA sequence level influence the functions and expressions of genes as well as the phenotype. My chief approach is functional genomics (Eisenberg et al. 2000; Lockhart and Winzeler 2000), which studies gene expression, protein-protein interaction, protein subcellular localization, metabolic network, genetic interaction, phenotypic effects upon gene deletion, and so on at the genomic scale. The integration of functional genomics and evolution improves our understanding of evolutionary processes by building a bridge between the genotype and phenotype and by providing a genomic picture of the genetic mechanisms of evolution.

In my studies, I used the budding yeast *Saccharomyces cerevisiae* as a model organism. *S. cerevisiae* is particularly useful for functional evolutionary genomics because of (i) the availability of extensive functional genomic data that allow potential connections between the genotype and phenotype at many different levels, (ii) the availability of genome sequences of multiple strains and related species, and (iii) the relative ease of various genetic manipulations including gene deletion and gene replacement.

In Chapter 2, I studied the evolutionary rate of protein-protein interaction. Protein sequence changes may result in protein function changes, which can in turn lead to phenotypic changes. Because many proteins function by interacting with other proteins, investigating protein-protein interaction is crucial to fill the gap between protein sequence evolution and phenotypic evolution. Although it is widely assumed that orthologous genes of different species have similar functions, this hypothesis has never been critically evaluated systematically. Furthermore, while genome-wide patterns of protein sequence evolution have been extensively studied, genome-wide patterns of protein function evolution are unclear. To address these questions, I used experimental assays to examine the potential interactions of pairs of *Kluyveromyces waltii* proteins whose *S. cerevisiae* orthologs are known to interact with each other. I then calculated the

evolutionary rate of protein function using the information that *K. waltii* and *S. cerevisiae* diverged approximately 150 million years ago. My results show that protein interactions are extremely conserved in evolution when there is no gene duplication and suggest a complex relationship between protein function evolution and sequence evolution. These results validate the widely held view that orthologous genes of different species have similar functions (at least at the level of protein interaction) and provide a baseline for further examination of the evolutionary rate of protein function in duplicate genes.

In Chapter 3, I studied genetic interactions (i.e. epistasis) in metabolic networks**.** Evolutionary changes often occur at multiple loci and the interaction of these changes is a common phenomenon with widespread and profound evolutionary implications. Although high-throughput epistasis data from model organisms are being generated and used to construct genetic networks, the extent to which epistasis reflects functional relationship of involved genes remains unclear. We addressed this question using metabolic network analysis. We found that negative epistasis in fitness occurs mainly between nonessential reactions with overlapping functions, whereas positive epistasis usually involves essential reactions, is highly abundant, and surprisingly, often occurs between reactions without overlapping functions. I validated these theoretical results experimentally in yeast by introducing yeast partial deletion alleles on a set of selected loci, measuring the fitness values of single mutants and double mutants, and calculating the epistasis for each allele pair. Furthermore, we identified the mechanistic basis of our observations. Our findings necessitate the distinction of the concept of genetic interaction from non-independent gene effects and call for reevaluation of evolutionary theories that depend on prevalent negative epistasis.

In Chapter 4, I studied the genetic interaction between tRNA and codon usage, and how such interaction improves the efficiency of protein translation, which is an important fitness determinant in rapidly growing organisms. It is widely believed that synonymous codons are translated with unequal speeds and that translational efficiency is maximized by the exclusive use of rapidly translated codons. Using next-generation-sequencing-based ribosomal profiling data, we estimated for the first time the *in vivo* translational speeds of all 61 sense codons from the budding yeast *S. cerevisiae*. Surprisingly, preferentially used synonymous codons are not translated faster than

unpreferred ones, and no correlation exists between the translational speed of a codon and the concentration of its cognate tRNA. We hypothesize that the phenomenon of similar translational speeds of different synonymous codons is a result of proportional use of synonymous codons according to their cognate tRNA concentrations, the optimal strategy in enhancing translational efficiency under tRNA shortage, which is a cellular condition that is supported by circumstantial empirical evidence but has not been seriously considered in the codon usage literature. Our hypothesis predicts that, for each amino acid, the fractional use of a codon among its synonymous codons equals the fractional concentration of its cognate tRNA among all isoaccepting tRNAs, and this is indeed the case in all eukaryotic model organisms examined (*S. cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*). We further tested our hypothesis by a manipulative experiment in which multiple synonymous versions of a heterologous red florescent protein gene were highly expressed to induce different levels of codon-tRNA imbalance in yeast. We measured the expression level of a yellow florescent protein gene, which serves as a reporter that indicates the overall cellular translational efficiency. This inducer-reporter experimental system excluded multiple confounding factors such as the potentially different translational accuracies of synonymous codons. Our results unambiguously support the hypothesis that codon-tRNA balance, rather than exclusive use of preferred codons, optimizes cellular translational efficiency. Our hypothesis also applies to amino acid usage, suggesting that it again is shaped by selection for translational efficiency. Together, our study reveals a previously unsuspected mechanism by which unequal codon usage increases translational efficiency, demonstrates widespread natural selection for translational efficiency, and offers new strategies to improve synthetic biology.

In Chapter 5, I aimed to address a question about gene-environment interactions. The fitness effect of deleting a gene tells us how important the gene is, which is crucial for understanding the gene's evolutionary patterns and rates. However, the genome-wide fitness effect data were previously generated by DNA hybridization based methods that are known to have a low accuracy and a low dynamic range. Furthermore, these data were generated in artificial conditions rather than in natural environments of yeast.

Taking advantage of the high accuracy and large dynamic range of a high-throughput sequencing technology (Illumina), I am now digitally counting the numbers of cells of different deletion strains in a competition assay that includes all deletion strains. I am also using various conditions, including oak tree exudates and grape juice that may mimic the natural environments of yeast. The much improved fitness measurements in biologically relevant environments will allow the test of several important hypotheses (e.g., adaptive gene loss) and provide a solid base for understanding the relation between DNA sequence evolution and phenotypic evolution.

In sum, I capitalize on recent developments in molecular genetics and functional genomics to study the molecular basis of phenotypic variation, and how such molecular basis influences the evolutionary process. By doing so, I am able to test a series of important evolutionary hypotheses and offer new perspectives on the mechanisms of evolution.

1.2 References

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* **104**(18): 7489-7494.

Cuevas JM, Domingo-Calap P, Sanjuan R. 2012. The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol Biol Evol* **29**(1): 17-20.

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. 2000. Protein function in the post-genomic era. *Nature* **405**(6788): 823-826.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature* **415**(6875): 1024-1026.

Gruber JD, Vogel K, Kalay G, Wittkopp PJ. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects and dominance. *PLoS Genet*: e1002497.

Hietpas RT, Jensen JD, Bolon DN. 2011. Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A* **108**(19): 7896-7901.

Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. *Science* **330**(6005): 825-827.

Lockhart DJ, Winzeler EA. 2000. Genomics, gene expression and DNA arrays. *Nature* **405**(6788): 827-836.

Loh E, Salk JJ, Loeb LA. 2010. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci U S A* **107**(3): 1154-1159.

Maynard Smith J. 1998. *Evolutionary genetics*. Oxford University Press.

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**(5844): 1544-1548.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**(12): 1173-1175.

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**(7126): 383-386.

Shultzaberger RK, Malashock DS, Kirsch JF, Eisen MB. 2010. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLoS Genet* **6**(7): e1001042.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* **415**(6875): 1022-1024.

Weinreich DM, Delaney NF, Depristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**(5770): 111-114.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet* **30**(4): 411-415.

CHAPTER 2
MEASURING THE EVOLUTIONARY RATE OF PROTEIN-PROTEIN INTERACTION

2.1 Abstract

Despite our extensive knowledge about the rate of protein sequence evolution for thousands of genes in hundreds of species, the corresponding rate of protein function evolution is virtually unknown, especially at the genomic scale. This lack of knowledge is primarily due to the huge diversity in protein function and the consequent difficulty in gauging and comparing rates of protein function evolution. Nevertheless, most proteins function through interacting with other proteins, and protein-protein interaction (PPI) can be tested by standard assays. Thus, the rate of protein function evolution may be measured by the rate of PPI evolution. Here we experimentally examine 87 potential interactions between *Kluyveromyces waltii* proteins whose one-to-one orthologs in the related budding yeast *Saccharomyces cerevisiae* have been reported to interact. Combining our results with available data from other eukaryotes, we estimate that the evolutionary rate of protein interaction is $(2.6 \pm 1.6) \times 10^{-10}$ per PPI per year, three orders of magnitude lower than the rate of protein sequence evolution measured by the number of amino acid substitutions per protein per year. The extremely slow evolution of protein molecular function may account for the remarkable conservation of life at molecular and cellular levels and allow studying the mechanistic basis of human disease in much simpler organisms.

2.2 Introduction

The rate of protein sequence evolution has been of enduring interest to evolutionary biologists (Zukerkandl and Pauling 1965; Kimura 1968; Li 1997; Nei and Kumar 2000; Koonin and Galperin 2003; Drummond and Wilke 2008; Wolf et al. 2009) ever since the primary sequences of homologous proteins became available about 50 years ago (Doolittle 2010). Estimation and comparison of the rate of protein sequence evolution led to several major discoveries, including the establishment of the molecular clock concept (Zukerkandl and Pauling 1965) and the application of the concept to molecular dating of evolutionary events (Hedges et al. 2006), and the proposal of the neutral theory of molecular evolution (Kimura 1968; King and Jukes 1969; Kimura 1983), a paradigm-shifting episode in the history of evolutionary biology (Zhang 2010). In the last decade, studies of the rate of protein sequence evolution have exploded, due to the availability of hundreds of complete genome sequences from diverse organisms. Despite some controversies, much has been learned from these studies, such as the identification of various determinants of the rate of protein sequence evolution (Hirsh and Fraser 2001; Pal et al. 2001; Fraser et al. 2002; Jordan et al. 2002; Zhang and He 2005; Liao et al. 2006; Drummond and Wilke 2008; Wang and Zhang 2009; Liao et al. 2010b; Liao et al. 2010a; Yang et al. 2010) and the estimation of the fraction of proteins subject to positive selection in human and ape evolution (Clark et al. 2003; Nielsen et al. 2005; Bakewell et al. 2007; Nozawa et al. 2009).

Surprisingly, however, very little is known about the rate of protein function evolution, despite that such information could be invaluable for answering a number of important questions. For instance, if most amino acid changes are adaptive, one would predict a positive correlation between the rate of protein function change and the rate of protein sequence change. By contrast, this correlation is not expected if most amino acid changes do not affect protein function and are neutral. Knowing the rate of protein function evolution also helps us understand the speed and frequency with which new functions originate in evolution. While the rate of protein function evolution can be calculated anecdotally for a few functionally well characterized proteins such as the vertebrate hemoglobin and opsin, there has been no systematic effort to estimate this rate from many proteins. This is probably due to the huge diversity in protein function, which

makes functional characterizations of many proteins from multiple species both technically challenging and financially costly. Furthermore, even if such functional data are available, it would be difficult to quantify functional differences among homologous proteins and compare such differences among nonhomologous proteins. Thus, it would be ideal to have a universal functional measure that can be applied to a large number of proteins for estimating and comparing the rates of protein function evolution. Because most proteins function through interacting with other proteins and protein-protein interactions (PPIs) can be tested by standard methods such as the yeast two-hybrid (Y2H) assay (Fields and Song 1989), we propose to use PPI as one universal measure of protein function in estimating the evolutionary rate of protein function. In this study, we experimentally probe PPI evolution between two yeast species using Y2H assays. Combining our data with existing PPI data from multiple other eukaryotes, we estimate that the evolutionary rate of PPI is three orders of magnitude lower than that of protein sequence. The striking conservation of protein molecular function has important implications for evolutionary biology and biomedicine.

2.3 Measuring the rate of PPI evolution in two yeasts

Because large-scale PPI data are available for a number of model organisms such as the budding yeast *Saccharomyces cerevisiae* (Yu et al. 2008), nematode worm *Caenorhabditis elegans* (Li et al. 2004), fruitfly *Drosophila melanogaster* (Giot et al. 2003), and human (*Homo sapiens*) (Rual et al. 2005), one may think that the rate of PPI evolution can be estimated directly from these existing data. Such estimation, however, would be highly unreliable, because these data were generated by high-throughput methods that have high false-negative and false-positive rates whose exact values are either unknown or not known with any precision (Yu et al. 2008). Instead, we decided to examine individually the interactions between two proteins whose respective one-to-one orthologs in another species are known to interact. The requirement for one-to-one orthologs minimizes the influence of gene duplication, which is known to induce changes of protein function (Ohno 1970; Zhang 2003; Conant and Wolfe 2008), including PPIs (Wagner 2001; He and Zhang 2005). We chose to compare the budding yeast *Saccharomyces cerevisiae* (*Sce*), a genetic model organism with abundant PPI data (He

and Zhang 2009), with its relative *Kluyveromyces waltii* (*Kwa*), which diverged ~150 million years (MY) ago (Wolfe and Shields 1997). This intermediate level of divergence provides time for potential evolutionary changes in PPI, yet ensures accuracy in identifying one-to-one orthologs.

We started by identifying the subset of *Sce* proteins that have one-to-one orthologs in *Kwa* (Figure 2.1). For this subset, we then identified from BioGRID (http://thebiogrid.org/) 335 PPIs with at least two Y2H reports and 481 PPIs with only one small-scale Y2H report. We focused on previous Y2H reports because different types of PPIs have variable rates of detection by different methods (Braun et al. 2009) and Y2H is our method of choice. We disregarded PPIs with only one large-scale Y2H report due to the high false positive rates of high-throughput studies. We refer to these 335+481=836 PPIs as putative *Sce* PPIs. They should have relatively low probabilities to be false positives because they have been reported in either one small-scale experiment or at least two experiments (Figure 2.1). We selected *Kwa* genes orthologous to the *Sce* genes involved in 115 randomly chosen putative PPIs after setting several criteria to lessen the effort for gene cloning (Figure 2.1), and were able to clone 87 pairs of them for a standard Y2H assay.

We found that, among the 87 protein pairs (Table 2.1), one or both members of 11 pairs showed self-activation (e.g., row II, column B in Figure 2.2; Figure 2.1) and hence could not be evaluated for PPI, 33 pairs had no PPI (e.g., column E in Figure 2.2; Figure 2.1), and the remaining 43 had PPIs (e.g., column C in Figure 2.2; Figure 2.1). To validate that the 33 *Kwa* non-interactive pairs represent true evolutionary changes between *Sce* and *Kwa*, we need to confirm that their corresponding *Sce* proteins interact in our Y2H assay. We were able to clone the *Sce* genes for 29 of the 33 pairs (Figure 2.1). Excluding 2 self-activation cases, we found none of the remaining 27 *Sce* pairs to interact (e.g., column F in Figure 2.2; Figure 2.1), which is owing to either false positive errors in the *Sce* PPI database or the known variation in PPI detection by different variants of Y2H (Chen et al. 2010). Thus, none of the 33 *Kwa* non-interactive pairs can be confirmed to have resulted from true evolutionary changes.

If two proteins have been reported to interact in *Sce* and that their orthologs are confirmed by us to interact in *Kwa*, the probability that the reported *Sce* interaction is not

genuine is lower than 0.005 (see 2.8.1). Thus, for the 43 *Kwa* PPIs, their corresponding *Sce* PPIs are most likely true. Indeed, we were able to experimentally validate each of 20 randomly selected corresponding *Sce* PPIs (e.g., column D in Figure 2.2; Figure 2.1; Table 2.1).

Taken together, our experiments showed that 43 of the 43 *Sce* PPIs are conserved in *Kwa*. Although PPIs detected by Y2H may not be biological, those detected here are highly likely to be biological because nonspecific or artificial interactions are not expected to be evolutionarily conserved. Our experiment was designed to identify *Sce* PPIs that are absent in *Kwa* due to either gains of interactions in *Sce* or losses of interactions in *Kwa* after the species separation. Hence, we effectively measured the total gains and losses that occurred in one lineage during 150 MY. With this consideration, we estimated that the 95% confidence interval of the total rate of PPI evolution is between 0 and 4.6×10^{-10} per PPI per year, with the maximum likelihood estimate being 0 (Figure 2.3; see 2.8.2).

2.4 A combined estimate of the rate of PPI evolution

To investigate the generality of our estimate, we analyzed all previously reported between-species PPI differences for which false positive and false negative errors can be excluded. One study (Matthews et al. 2001) reported that 6 of the 19 *Sce* PPIs are conserved in *C. elegans*. However, the authors did not use one-to-one orthologs and thus could not exclude the influence of gene duplication. If only one-to-one orthologs are considered, their data indicate that 2 of 5 *Sce* PPIs are conserved in *C. elegans*. If the two species diverged 1300 MY ago, as suggested by molecular dating (Hedges et al. 2006), the maximum likelihood estimate of the PPI evolutionary rate is 7.0×10^{-10} and the 95% confidence interval is (1.6×10^{-10}, 2.0×10^{-9}) (Figure 2.3; see 2.8.2). Another study used a high-throughput method to examine the PPIs between transcription factors and found 6 of the 23 gold-standard mouse PPIs to be conserved in human (Ravasi et al. 2010). The purpose of using gold-standard PPIs was to avoid false positives. Because the rate of detection of a true PPI in that study (Ravasi et al. 2010) was 0.253, the actual fraction of mouse PPIs conserved in human is $6/(23 \times 0.253) \approx 100\%$ (i.e., 6 out of 6). If human and mouse diverged 90 MY ago (Hedges et al. 2006), the maximum likelihood estimate of the

PPI evolutionary rate is 0, with the 95% confidence interval being (0, 5.5×10^{-9}) (Figure 2.3; see 2.8.2).

The confidence intervals of the PPI evolutionary rate estimates from the three datasets that encompassing yeasts, worm, and mammals overlap (Figure 2.3), although the estimate from our data is most precise, because the size of our data is four times that of the previous data combined. Using the three datasets together, we derived a maximum likelihood estimate of the PPI evolutionary rate of 2.6×10^{-10} per function per year, with a standard error of 1.6×10^{-10} (see 2.8.2). Our estimated rate of PPI evolution is extremely low. As a comparison, the rate of sequence evolution for the yeast, *C. elegans*, mouse, and human proteins involved in the calculation of the rate of PPI evolution in this study is on average 4.1×10^{-7} amino acid substitutions per protein per year (see 2.8.3). That is, ~1558 amino acid substitutions, or ~5.0 per site, will happen in the time required for one PPI change in a protein.

2.5 Caveats

Although our yeast experiment has substantially increased the sample size for estimating the rate of PPI evolution, the number of PPIs examined is still small, compared to the number of all PPIs in yeast (He and Zhang 2009). It is thus important to ask whether the PPIs we studied are a representative sample of all yeast PPIs. For this purpose, we first plotted the frequency distribution of PPI degrees (i.e., the number of BioGRID-recorded Y2H-based PPIs per gene) for all 3152 *Sce* genes that have one-to-one *Kwa* orthologs, and the corresponding distribution for the 74 *Sce* genes involved in the 43 PPIs conserved between *Sce* and *Kwa*. Note that only PPIs among the 3152 genes are counted to avoid the complication of PPI changes after gene duplication. We found that both distributions cover similarly large degree variations among genes, although our sample of 74 genes tend to have higher degrees than the 3152 genes ($P = 5 \times 10^{-8}$, Mann-Whitney test; Figure 2.4A). This disparity, however, is not unexpected, because genes with higher degrees are more likely to be chosen when PPIs are randomly picked. To illustrate this point, we randomly sampled 43 PPIs from all the PPIs among the 3152 genes and calculated the median degree of the genes involved in the sampled PPIs. We repeated this process 1000 times to obtain a frequency distribution of the median degree

(Figure 2.4B). Interestingly, the median degree of the 74 genes studied is even lower than that of randomly selected ones, although their difference is not significant ($P = 0.06$; simulation test).

We also plotted the frequency distribution of gene importance measured by the fitness reduction caused by the deletion of the gene, for all 3152 genes and the 74 genes we studied, respectively. The 74 genes cover the whole range of gene importance, although they tend to be more important than average genes ($P = 0.002$; Figure 2.4C). This finding is expected, because PPI degree and gene importance are known to correlate positively with each other (Jeong et al. 2001; He and Zhang 2006). Nevertheless, there is no significant difference between the median importance of the 74 genes we studied and that of the genes involved in the randomly selected PPIs of the above simulation ($P = 0.13$, simulation test; Figure 2.4D). Thus, the apparent bias in the degree and importance of the 74 genes we studied (Figure 2.4A & C) is the byproduct of random PPI selection. Because the PPIs were randomly selected in our experiment, the above bias in some gene properties is unlikely to affect our estimation of the rate of PPI evolution. Furthermore, there is no indication that important genes or genes with higher degrees tend to have PPIs that are evolutionarily more conserved.

We further compared the frequency distributions for the 3152 genes and the 74 studied genes in terms of protein sequence conservation (Figure 2.4E) and the nonsynonymous/synonymous substitution rate ratio (Figure 2.4F), but found no significant differences. If the rate of PPI evolution is primarily determined by the rate of protein sequence evolution, our results suggest that our sample is unbiased for estimating the rate of PPI evolution.

Because mRNA and protein expression levels affect the evolution of protein structure, stability, and the propensity for nonspecific protein interactions (Drummond and Wilke 2008; Vavouri et al. 2009; Yang et al. 2010), we also compared our 74 studied genes with the 3152 genes in terms of mRNA expression levels (Figure 2.4G) and protein expression levels (Figure 2.4H), but found no significant differences.

We further examined Gene Ontology (GO) (Ashburner et al. 2000) differences between the two groups of genes. Although one to three functional categories were found to be significantly deprived or enriched (at a false discovery rate of 5%) among the

74 genes for each of the three aspects of GO (cellular component, molecular function, and biological process) (Al-Shahrour et al. 2007), the 74 genes are not limited to a small number of GO categories (Figure 2.5). Furthermore, even for the GO categories with significant discrepancies between the 74 genes and the 3152 genes, the discrepancies are moderate when the entire distribution of genes across all GO categories is considered and are thus unlikely to have a major impact on the estimation of the rate of PPI evolution (Figure 2.5).

We found that at least 27 of the 87 putative PPIs of *Sce* cannot be confirmed by our Y2H assay. Because we defined putative PPIs relatively rigorously, with the requirement that they had been reported by two Y2H experiments or one small-scale Y2H experiment, one may wonder why many of them cannot be confirmed in our Y2H assay. One reason is that a PPI may not be detectable by all variants of Y2H (Chen et al. 2010). Further, our Y2H assay uses three reporter genes and only when all three genes are activated will the PPI be scored. This stringent design guards against false positives caused by spurious gene activation without PPI, which can happen occasionally. In fact, our validation rate is much higher than those reported in the literature (Yu et al. 2008; Ravasi et al. 2010), presumably because of the small-scale nature of our experiment.

Because both *Sce* PPIs and *Kwa* PPIs were examined in *Sce* cells, one wonders whether our experimental design would cause an overestimation of the evolutionary rate of PPI, due to the possibility that naturally interacting *Kwa* proteins may not interact well in *Sce* cells. This concern is unnecessary here, because we found no validated *Sce* PPIs whose *Kwa* orthologs do not interact in our Y2H assay.

2.6 Implications

In this work, we estimated the evolutionary rate of protein molecular function by measuring the conservation of PPIs between species, and found the rate to be strikingly low in the absence of gene duplication. Our finding has a number of important implications. First, it suggests a high similarity in molecular function between one-to-one orthologs from even distantly related species. For instance, based on our estimated rate, an interaction between two human proteins is expected to be present between their respective one-to-one orthologs in mouse, fish, fly, worm, fungi, and plants with a

probability of 98%, 89%, 79%, 77%, 71%, and 66%, respectively (Figure 2.6). Life is fundamentally conserved at molecular and cellular levels, as most biological processes at these levels are similar among divergent species (Zhang 2010). Given the prevalence and importance of PPIs in almost all cellular processes, the extreme conservation of PPIs is likely one of the bedrocks of the conservative nature of life. Note, however, that our Y2H assay is qualitative rather than quantitative. Hence, we cannot exclude the possibility that the strength of a protein interaction evolves much faster than the presence/absence of the interaction.

Second, although molecular functions of proteins are conserved in evolution, the physiological roles of proteins and their contributions to organismal fitness can change quickly and substantially, evident from frequent observations of the huge diversity in the phenotypic effect of orthologous gene deletions (McGary et al. 2010) and the great disparity in the dispensability of orthologous genes in different species (Zhang and He 2005; Liao and Zhang 2008; Kim et al. 2010). For instance, some mouse defects in blood vessel formation and yeast hypersensitivity to the hypercholesterolemia drug lovastatin are caused by mutations of orthologous genes (McGary et al. 2010). In another example, *Arabidopsis* orthologs of human genes implicated in the Waardenburg syndrome (deafness and neutral crest anomalies) are involved in gravitropism (McGary et al. 2010). A systematic comparison between phenotypes of human and mouse mutations found that over 20% of mouse one-to-one orthologs of human essential genes are nonessential (Liao and Zhang 2008). Yet, if the molecular functions of one-to-one orthologous genes are highly conserved in evolution, as suggested by the present study, the molecular underpinnings of human disease may be studied in much simpler model organisms that do not even have the disease or relevant tissue/organ.

Third, previous analyses of high-throughput PPI data revealed a substantial amount of subfunctionalization and neofunctionalization after gene duplication (Wagner 2001; He and Zhang 2005). The contrast between these results with the present finding in one-to-one orthologous genes suggests that the majority of molecular function changes in protein evolution are associated with gene duplication. However, due to the unreliability of high-throughput PPI data, previous results on duplicate genes (Wagner 2001; He and Zhang 2005) should be verified in the future. It would be highly desirable

to conduct a study on duplicate genes similar to the present one to quantify the difference in the rate of protein function evolution in the presence and absence of gene duplication (Studer and Robinson-Rechavi 2009), as has been conducted recently on the rate of protein subcellular relocalization (Qian and Zhang 2009). In this respect, the *Sce* and *Kwa* comparison will also be appropriate, because *Sce* retains ~500 pairs of duplicate genes generated by a whole genome duplication that occurred since the separation of *Sce* from *Kwa* (Wolfe and Shields 1997; Kellis et al. 2004).

2.7 Methods

2.7.1 Identification of putative PPIs for experimental tests

Gene sequences of *S. cerevisiae* (*Sce*) were downloaded from *Saccharomyces* Genome Database (SGD, http://yeastgenome.org/) and those of *K. waltii* (*Kwa*) were downloaded from the supplementary materials of (Kellis et al. 2004) (http://www.nature.com/nature/journal/v428/n6983/extref/nature02424-s1.htm). To identify one-to-one orthologous genes between the two species, we combined the genomes of *Sce* and *Kwa*, conducted all-against-all BlastP searches with an E-value cutoff of 1×10^{-20}, and removed self-hits. If (i) proteins A and B are reciprocal best hits in the above search, (ii) they do not belong to the same species, and (iii) the aligned region of the two proteins is longer than 80% of the shorter one, we classify them as a pair of one-to-one orthologs.

Protein interaction data in *Sce* were downloaded from BioGRID (http://thebiogrid.org/) at the beginning of our study in 2006 (GRID-ORGANISM-Saccharomyces_cerevisiae-2.0.20.tab.txt). Among the *Sce* proteins that have one-to-one *Kwa* orthologs, we identified 355 PPIs with at least two independent Y2H reports. Among the remaining PPIs, we identified 481 that had one small-scale Y2H report (i.e., with <30 PPIs per report). These two sets of PPIs were treated as putative *Sce* PPIs subject to further analysis.

To test the interaction between the *Kwa* orthologs of a pair of *Sce* proteins that are known to interact, we need to clone the *Kwa* orthologs. To reduce the difficulty in gene cloning, we selected *Sce* PPIs for which the *Kwa* orthologs have open reading frames between 400 and 1600 nucleotides long. We excluded intron-containing genes so that the

full coding region could be amplified from genomic DNA in one piece. We also eliminated genes incompatible with the restriction sites available on plasmids pGADT7 or pGBKT7. The majority of *Sce* PPIs selected had been discovered in at least two independent Y2H reports (73%), while a minority of them (27%) had been identified in only one small-scale Y2H experiment (Table 2.1).

2.7.2 Yeast two-hybrid assays

After cloning the *Kwa* genes, we performed the Y2H assay using the Matchmaker GAL4 Two-Hybrid System 3 (Clonetch). The two PPI partners were cloned into pGADT7 and pGBKT7 plasmids, respectively, through the following procedure. We first amplified the genes by polymerase chain reaction (PCR), using PfuUltra high-fidelity polymerase (Stratagene) to minimize PCR errors during the amplification. We purified the PCR products (Qiagen), digested them with 2 of the 5 restriction enzymes (*EcoRI*, *BamHI*, *NdeI*, *PstI*, and *ClaI*), and cloned them into pGADT7 and/or pGBKT7 by T4 DNA ligase (Promega). We transformed the ligation products into TOP10 chemically competent cells (Invitrogen), extracted the plasmid (Qiagen) and confirmed the clones by DNA sequencing at University of Michigan DNA Sequencing Core.

We transformed the two plasmids into *S. cerevisiae* AH109 cells (*MAT***a**, *trp1-901*, *leu2-3, 112*, *ura3-52*, *his3-200*, *gal4Δ*, *gal80Δ*, *LYS2::GAL1*$_{UAS}$*-GAL1*$_{TATA}$*-HIS3*, *GAL2*$_{UAS}$*-GAL2*$_{TATA}$*-ADE2*, *URA3::MEL1*$_{UAS}$*-MEL1*$_{TATA}$*-lacZ*, *MEL1*), which were selected on synthetic dextrose plates with leucine and tryptophan dropped out (SD-Leu-Trp). The colonies were further pinned onto synthetic dextrose plates with adenine, histidine, leucine, and tryptophan dropped out, and with 20 μg/ml X-α-gal added (SD-Ade-His-Leu-Trp/X-α-gal). Because *MEL1* encodes a secreted enzyme α-galactosidase, its presence can be assayed directly on X-α-gal-containing plates without cell lysis. If the transformed yeast can grow on the dropout plates (SD-Ade-His-Leu-Trp/X-α-gal) and appears blue, the proteins are considered to be interacting with each other. For a strain to grow on SD-Ade-His-Leu-Trp/X-α-gal and be blue, all three reporter genes (*HIS3*, *ADE2* and *MEL1*, under promoters *pGAL1*, *pGAL2* and *pMEL1*, respectively) must be activated. Hence, our Y2H assay is quite stringent. The high stringency implies that some weak PPIs may not be detected by our Y2H assay. Self-activation was tested by co-

transformation of a gene-containing plasmid (pGADT7 or pGBKT7) and an empty plasmid (pGBKT7 or pGADT7, respectively). We excluded a gene pair from further consideration if either gene showed self-activation.

We selected 10 random protein pairs and tested their interaction [(Kwa_10129, Kwa_23895), (Kwa_12079, Kwa_9492), (Kwa_12518, Kwa_5419), (Kwa_13638, Kwa_23894), (Kwa_15314, Kwa_15321), (Kwa_16145, Kwa_18622), (Kwa_1973, Kwa_21767), (Kwa_2064, Kwa_10342), (Kwa_2079, Kwa_17326), (Kwa_21273, Kwa_23528)] in our Y2H assay, and none of them showed PPI, suggesting a low false positive rate in our experiment.

The potential PPI pairs that did not show positive results in *Kwa* by our Y2H experiment were examined in *Sce*. We cloned the corresponding *Sce* genes into the same plasmids (pGBKT7 and pGADT7) and then conducted the Y2H assay as described above. We also randomly selected a subset of positive *Kwa* PPIs and examined whether their *Sce* orthologs interact in our Y2H assay.

2.7.3 Examination of potential biases of the experimentally studied genes

Sce PPI information was from BioGRID. Only Y2H PPIs between *Sce* genes that have one-to-one *Kwa* orthologs were counted in PPI degree calculation. The fitness effect of gene deletion was obtained from (Deutschbauer et al. 2005). The mRNA expression levels were from (Holstege et al. 1998) and the protein expression levels were from (Ghaemmaghami et al. 2003). One-to-one orthologous genes between *Sce* and *Kwa* were aligned by ClustalW (Larkin et al. 2007) and the nonsynonymous to synonymous substitution rate ratios were calculated by Codeml in PAML (Yang 2007). To examine the impact of random sampling of PPIs on the bias of the selected genes, we randomly sampled 43 PPIs from all PPIs among the 3152 *Sce* genes that have one-to-one *Kwa* orthologs, and repeated this process 1000 times.

2.8 Appendices

2.8.1 Probability that a reported *Sce* PPI is non-genuine provided that its corresponding *Kwa* PPI is true

Let A be the event that X and Y do not interact in *Sce*. Let B be the event that X and Y are found to interact in *Sce*, based on the existing database. Let C be the event that the orthologs of X and Y (X' and Y') interact in *Kwa*. On average, a protein interacts with n proteins. Let us assume that, in the existing database, an average protein is also found to interact with e proteins that are actually false positives. If we assume that the false positive rate in the existing database is $P(A|B)=50\%$ (Deane et al. 2002), $e=n$. Empirical data suggest that $n+e \approx 5$ (He and Zhang 2005). Thus, $e = n \approx 2.5$. Let N = 6000 be the total number of proteins in yeast and c be the probability that a true *Sce* PPI is conserved in *Kwa*.

We want to estimate $P(A \mid [B \& C])$, the probability that X and Y do not interact in *Sce*, given that they were reported to interact in *Sce* and that their orthologs are confirmed to interact in *Kwa*.

$P(A)$ is the probability that a random pair of proteins do not interact in *Sce*. We have

$$P(A) = \frac{N-n}{N}.$$

P([B & C] | A) is the probability that a random pair of *Sce* proteins (X and Y) are found to interact in the database and that their orthologs (X' and Y') interact in *Kwa*, given that X and Y actually do not interact in *Sce*. We have

$$P([B \& C] \mid A) = (\frac{n}{N}k)(\frac{e}{N-n}).$$

The first term on the right is the probability that X' and Y' interact in *Kwa*, given that X and Y do not interact in *Sce* (i.e., the probability of origination of the new PPI in *Kwa*). Because the probability that a random pair of proteins evolve a new PPI after a very long time is n/N, the probability that *Kwa* evolved a new PPI in a limited amount of time is $k(n/N)$, where $0<k<<1$.

The second term is the probability that X and Y are found to interact in *Sce* given that they actually do not interact in *Sce*.

$P(B \& C)$ is the probability that a random pair of *Sce* proteins (X and Y) are found to interact in the database and that their orthologs (X' and Y') interact in *Kwa*. We have

$$P(B \& C) = P(A)P([B \& C] \mid A) + P(\bar{A})P([B \& C] \mid \bar{A})$$
$$= (\frac{N-n}{N})(\frac{n}{N}k)(\frac{e}{N-n}) + \frac{n}{N}c = \frac{enk}{N^2} + \frac{cn}{N} \quad ,$$

where $\bar{A}$ means not A.

Using Bayes' theorem and above formulae, we have

$$P(A \mid [B \& C]) = \frac{P(A)P([B \& C] \mid A)}{P(B \& C)}$$

$$= \frac{(\frac{N-n}{N})(\frac{n}{N}k)(\frac{e}{N-n})}{\frac{enk}{N^2} + \frac{cn}{N}} = \frac{ek}{ek + Nc}$$

$$< \frac{ek}{Nc} < \frac{e}{Nc} = \frac{2.5}{6000c} = \frac{0.000417}{c}$$

Even if c is as low as 10%, P(A | [B & C]) < 0.005. In other words, the probability that X and Y do not interact in *Sce*, given that they were reported to interact in *Sce* and that their orthologs are confirmed to interact in *Kwa*, is very low.

2.8.2 Rate of PPI evolution

Let the rate of PPI evolution be r per PPI per year. The probability of no change of PPI after t years is $p = (1-r)^t$. Because all 43 *Sce* PPIs were found in *Kwa*, there has been no gain of PPI in *Sce* and no loss of PPI in *Kwa* since the species separation. This observation is equivalent to the lack of gain and loss of PPI in one lineage since the *Sce*-*Kwa* separation. The maximum likelihood estimate of r is thus 0. Assuming that the probability of observing what we have observed is at least 5%, we have $P = p^{43} = [(1-r)^t]^{43} > 0.05$. Using $t = 150 \times 10^6$ years (Wolfe and Shields 1997), we found that $r < 4.6 \times 10^{-10}$ per PPI per year. Thus, the 95% confidence interval of r is $(0, 4.6 \times 10^{-10})$.

Ravasi and colleagues examined the interactions between transcription factors (TFs) for 877 mouse and 762 human proteins, based on a high-throughput method (Ravasi et al. 2010). Using a gold-standard set of mouse TF-TF interactions, they estimated that the detection rate of their assay is 23/91=0.253. If all of the 23 mouse PPIs that they could detect in their assay are conserved in human, $23 \times 0.253 \approx 6$ are expected

to be detected in human. In other words, their assay actually examined the conservation of 6 mouse PPIs in human. The observed number of conserved PPIs was 6. Thus, based on the observation that 6 of 6 mouse PPIs are conserved in human, we estimated that the 95% confidence interval of r is $(0, 5.5\times10^{-9})$, using a divergence time of 90×10^{6} years for the species pair (Hedges et al. 2006). In the above estimation, we focused on the use of gold-standard PPIs to avoid false-positives.

Matthews and colleagues examined the conservation of *Sce* PPIs in *C. elegans* (Matthews et al. 2001). They observed that 35 of 216 *Sce* PPIs are conserved in *C. elegans*. However, most of the *Sce* PPIs were not confirmed. For the subset of 19 confirmed *Sce* PPIs, 6 were found to be conserved in *C. elegans*. But, the authors did not use one-to-one orthologs in their study. When we limited the analysis to one-to-one ortholgs in their data, 2 of 5 confirmed *Sce* PPIs are conserved in *C. elegans*. Based on this observation, the maximum likelihood estimate of r is 7.0×10^{-10}, which was estimated by solving $p = (1-r)^t = 0.4$, using $t = 1300\times10^{6}$ years (Hedges et al. 2006). To estimate the confidence interval of r, we assumed that both the probability of observing ≤ 2 conserved PPIs and the probability of observing ≥ 2 conserved PPIs exceed 5%. Let p be the probability that a PPI is conserved and $p = (1-r)^t$. We have

$\binom{5}{0}(1-p)^5 + \binom{5}{1}p(1-p)^4 + \binom{5}{2}p^2(1-p)^3 > 0.05$ and

$\binom{5}{2}p^2(1-p)^3 + \binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p) + \binom{5}{5}p^5 > 0.05$. By solving the two equations, we found that $0.0764 < p < 0.8107$, and thus the 95% confidence interval of r is $(1.6\times10^{-10}, 2.0\times10^{-9})$.

Combining the three datasets, we used a likelihood method to estimate r. The likelihood is $L = \{[(1-r)^{t_1}]^{43}\}\{[(1-r)^{t_2}]^{6}\}\{\binom{5}{2}[(1-r)^{t_3}]^2[1-(1-r)^{t_3}]^3\}$, where $t_1 = 150\times10^{6}$ years, $t_2 = 90\times10^{6}$ years, and $t_3 = 1300\times10^{6}$ years. We used the curvature method to estimate the standard error of the likelihood estimate.

Given the rate of PPI evolution r, the probability that a human PPI is present in a related species was calculated by $(1-r)^t$, where t is the time since the separation of human from the species concerned.

2.8.3 Rate of protein sequence evolution

Using ClustalW, we aligned the one-to-one orthologous proteins for those proteins involved in the PPIs used for estimating the rate of PPI evolution. We counted the number of amino acid positions with identical amino acids between the two species (n) and counted the total aligned amino acid positions (N). The p-distance was calculated for each orthologous pair by $p = 1 - \frac{n}{N}$. The Poisson-corrected distance was calculated by $d = -\ln(1-p)$ (Nei and Kumar 2000). The rate of protein sequence evolution per residue was calculated by $d/(2t)$ and the rate of protein sequence evolution per protein was $(Nd)/(2t)$, where t is divergence time between the two species concerned. We then calculated the overall rate of protein sequence evolution of a set of proteins by the arithmetic mean rate of these proteins.

2.9 Acknowledgements

Xionglei He, Edwin Chan, and Huailiang Xu helped with experiments. I also thank Meg Bakewell, Calum Maclean, Jian-Rong Yang for valuable comments.

Table 2.1 Yeast two-hybrid (Y2H) assay results in *S. cerevisiae* and *K. waltii*.

K. waltii						*S. cerevisiae*					
Gene cloned in pGADT7	Gene cloned in pGBKT7	Y2H signal from pGADT7-gene1 +pGBKT7-gene2	Y2H signal from pGADT7 +pGBKT7-gene2	Y2H signal from pGADT7-gene1 +pGBKT7	PPI?	Gene to be cloned in pGADT7	Gene to be cloned in pGBKT7	Y2H signal from pGADT7-gene1 +pGBKT7-gene2	Y2H signal from pGADT7 +pGBKT7-gene2	Y2H signal from pGADT7-gene1 +pGBKT7	PPI?
Kwa_10129	Kwa_57	-	-	-	No	YIL162W	YBR283C	-	-	-	No
Kwa_10785	Kwa_16074	-	-	-	No	YER107C	YEL024W	-	-	-	No
Kwa_12536	Kwa_16027	-	-	-	No	YBR088C	YJR043C	-	-	-	No
Kwa_12559	Kwa_5419	-	-	-	No	YPR054W	YPR191W	-	-	-	No
Kwa_13638	Kwa_24100	-	-	-	No	YCR035C	YPL211W	-	-	-	No
Kwa_15314	Kwa_1116	-	-	-	No	YGR172C	YFL038C	-	-	-	No
Kwa_15314	Kwa_18303	-	-	-	No	YGR172C	YDR468C	-	-	-	No
Kwa_15314	Kwa_22440	-	-	-	No	YGR172C	YML001W	-	-	-	No
Kwa_16027	Kwa_15321	-	-	-	No	YJR043C	YJR006W				
Kwa_16145	Kwa_22169	-	-	-	No	YEL015W	YOL149W	-	-	-	No
Kwa_18083	Kwa_18622	-	-	-	No	YER044C	YML008C	-	-	-	No
Kwa_1973	Kwa_10342	-	-	-	No	YER148W	YMR270C	+	-	+	NA
Kwa_20310	Kwa_21767	-	-	-	No	YOR294W	YKR081C	-	-	-	No
Kwa_21557	Kwa_23528	-	-	-	No	YOR036W	YDR100W	-	-	-	No
Kwa_21557	Kwa_6051	-	-	-	No	YOR036W	YMR197C	-	-	-	No
Kwa_21557	Kwa_8776	-	-	-	No	YOR036W	YPL151C	-	-	-	No
Kwa_23317	Kwa_8279	-	-	-	No	YLR285W	YKL056C	-	-	-	No

Kwa_23344	Kwa_22629	-	-	-	No	YLR288C	YOL093W	-	-	-	No
Kwa_23344	Kwa_3047	-	-	-	No	YLR288C	YIL139C	-	-	-	No
Kwa_4081	Kwa_22900	-	-	-	No	YKL183W	YGR074W	-	-	-	No
Kwa_4672	Kwa_4689	-	-	-	No	YDL216C	YJR084W	-	-	-	No
Kwa_5148	Kwa_8925	-	-	-	No	YFR004W	YOR261C				
Kwa_5419	Kwa_19854	-	-	-	No	YPR191W	YBL045C				
Kwa_7350	Kwa_22375	-	-	-	No	YJL097W	YIR038C	-	-	-	No
Kwa_8414	Kwa_19079	-	-	-	No	YOR106W	YGL212W	-	-	-	No
Kwa_8628	Kwa_18044	-	-	-	No	YOR212W	YFL026W	-	-	-	No
Kwa_8628	Kwa_969	-	-	-	No	YOR212W	YHR005C	-	-	-	No
Kwa_8727	Kwa_24100	-	-	-	No	YPL146C	YPL211W	+	+	-	NA
Kwa_8727	Kwa_18184	-	-	-	No	YPL146C	YHR034C				
Kwa_9345	Kwa_10129	-	-	-	No	YLR378C	YIL162W	-	-	-	No
Kwa_9345	Kwa_15180	-	-	-	No	YLR378C	YJL002C	-	-	-	No
Kwa_9345	Kwa_15803	-	-	-	No	YLR378C	YMR149W	-	-	-	No
Kwa_9345	Kwa_20888	-	-	-	No	YLR378C	YEL002C	-	-	-	No
Kwa_11062	Kwa_14964	+	-	-	Yes	YFR008W	YMR052W	+	-	-	Yes
Kwa_11410	Kwa_23895	+	-	-	Yes	YGL153W	YDR142C				
Kwa_12079	Kwa_20496	+	-	-	Yes	YBR193C	YGR104C	+	-	-	Yes
Kwa_12218	Kwa_9492	+	-	-	Yes	YPL101W	YMR312W	+	-	-	Yes
Kwa_12518	Kwa_9942	+	-	-	Yes	YBR087W	YJR068W	+	-	-	Yes
Kwa_13638	Kwa_18184	+	-	-	Yes	YCR035C	YHR034C				
Kwa_13638	Kwa_3969	+	-	-	Yes	YCR035C	YGR095C				
Kwa_15314	Kwa_5058	+	-	-	Yes	YGR172C	YFL005W	+	-	-	Yes
Kwa_16145	Kwa_16983	+	-	-	Yes	YEL015W	YML064C				
Kwa_18470	Kwa_24776	+	-	-	Yes	YDR448W	YGR252W				
Kwa_19615	Kwa_23895	+	-	-	Yes	YIL160C	YDR142C				

Kwa_2064	Kwa_5058	+	-	-	Yes	YER136W	YFL005W				
Kwa_2079	Kwa_10342	+	-	-	Yes	YBL025W	YMR270C				
Kwa_2079	Kwa_16706	+	-	-	Yes	YBL025W	YHR200W				
Kwa_20972	Kwa_17326	+	-	-	Yes	YIR005W	YGL174W				
Kwa_21273	Kwa_5012	+	-	-	Yes	YOR136W	YNL037C				
Kwa_21570	Kwa_21709	+	-	-	Yes	YOR039W	YOR061W	+	-	-	Yes
Kwa_21653	Kwa_20450	+	-	-	Yes	YKR068C	YBR254C				
Kwa_21691	Kwa_20211	+	-	-	Yes	YOR057W	YDR328C				
Kwa_22058	Kwa_14356	+	-	-	Yes	YOL135C	YDR308C	+	-	-	Yes
Kwa_22685	Kwa_22905	+	-	-	Yes	YDR489W	YDR013W	+	-	-	Yes
Kwa_23242	Kwa_14356	+	-	-	Yes	YOR174W	YDR308C	+	-	-	Yes
Kwa_23242	Kwa_22058	+	-	-	Yes	YOR174W	YOL135C	+	-	-	Yes
Kwa_23344	Kwa_6791	+	-	-	Yes	YLR288C	YOR368W				
Kwa_23407	Kwa_17650	+	-	-	Yes	YDR088C	YGR006W	+	-	-	Yes
Kwa_3341	Kwa_4689	+	-	-	Yes	YDR179C	YJR084W	+	-	-	Yes
Kwa_3741	Kwa_5124	+	-	-	Yes	YPR066W	YPL003W	+	-	-	Yes
Kwa_4224	Kwa_20206	+	-	-	Yes	YDL065C	YDR329C				
Kwa_4224	Kwa_22162	+	-	-	Yes	YDL065C	YOL147C	+	-	-	Yes
Kwa_4882	Kwa_11481	+	-	-	Yes	YNL056W	YNL032W				
Kwa_5148	Kwa_20444	+	-	-	Yes	YFR004W	YIR011C	+	-	-	Yes
Kwa_5266	Kwa_22433	+	-	-	Yes	YOL111C	YOR007C				
Kwa_5338	Kwa_7510	+	-	-	Yes	YPR173C	YMR077C				
Kwa_5338	Kwa_7535	+	-	-	Yes	YPR173C	YLR025W	+	-	-	Yes
Kwa_5338	Kwa_13257	+	-	-	Yes	YPR173C	YLR181C	+	-	-	Yes
Kwa_5351	Kwa_18249	+	-	-	Yes	YPR178W	YDR473C	+	-	-	Yes
Kwa_5638	Kwa_15453	+	-	-	Yes	YDR280W	YGR195W	+	-	-	Yes
Kwa_6435	Kwa_8932	+	-	-	Yes	YCR086W	YOR264W				

Kwa_7031	Kwa_8776	+	-	-	Yes	YAL032C	YPL151C				
Kwa_8084	Kwa_4106	+	-	-	Yes	YHL011C	YKL181W				
Kwa_8394	Kwa_4882	+	-	-	Yes	YNL099C	YNL056W				
Kwa_8628	Kwa_13175	+	-	-	Yes	YOR212W	YCL032W	+	-	-	Yes
Kwa_9149	Kwa_6435	+	-	-	Yes	YOR281C	YCR086W				
Kwa_10450	Kwa_7946	+	-	+	NA	YBR077C	YKR007W				
Kwa_10880	Kwa_18748	+	+	+	NA	YER094C	YOR157C				
Kwa_11416	Kwa_10167	+	-	+	NA	YGL154C	YGL254W				
Kwa_12239	Kwa_17572	+	+	+	NA	YBR175W	YAR003W				
Kwa_14654	Kwa_324	+	+	-	NA	YNL151C	YNR003C				
Kwa_15294	Kwa_5649	+	-	+	NA	YLR007W	YDR288W				
Kwa_17994	Kwa_23894	+	+	-	NA	YLR051C	YLR119W				
Kwa_1811	Kwa_6081	+	+	-	NA	YBR217W	YNR007C				
Kwa_18795	Kwa_13991	+	+	-	NA	YDR404C	YJL140W				
Kwa_6452	Kwa_23884	+	+	-	NA	YNR046W	YDR140W				
Kwa_7730	Kwa_20211	+	+	-	NA	YMR094W	YDR328C				

Figure 2.1 Flowchart describing the selection of candidate protein-protein interactions (PPIs) for yeast-two-hybrid (Y2H) assays and the experimental results of PPI conservation between *S. cerevisiae* (*Sce*) and *K. waltii* (*Kwa*).

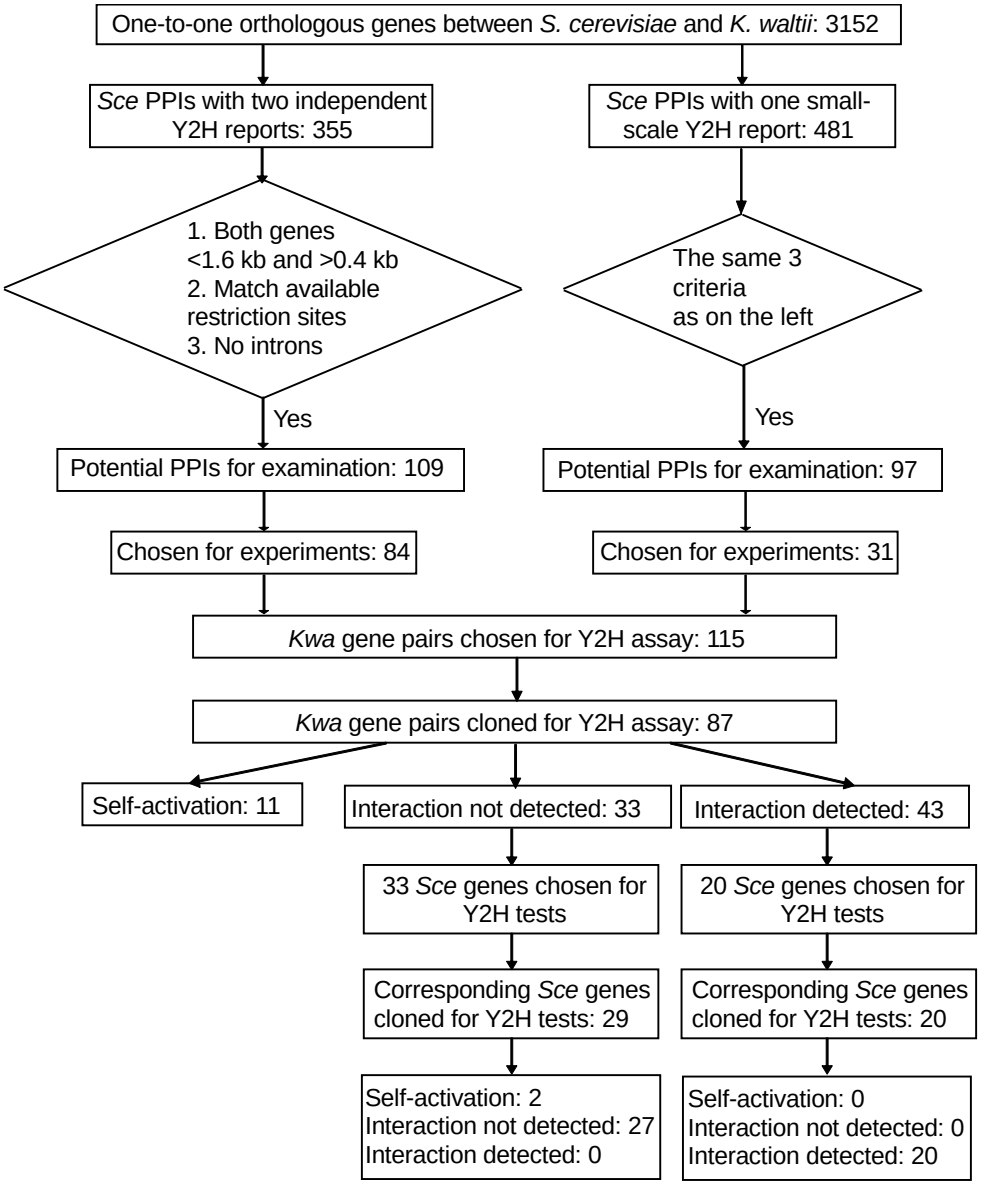


Figure 2.2 Examples of yeast two-hybrid (Y2H) experimental results for *S. cerevisiae* (*Sce*) and *K. waltii* (*Kwa*) genes.

Yeast cells for the Y2H assays were placed to the test plates (adenine, histidine, leucine and tryptophan dropout synthetic media with X-α-gal) in regions marked with black dots. Plasmids in the yeast cells are indicated. pGADT7-gene1 and pGBKT7-gene2 are the Y2H plasmids with gene 1 and gene 2 inserted, respectively. pGADT7-null and pGBKT7-null are the empty plasmids without gene inserts. Row III, column A is the negative control (pGADT7-null + pGBKT7-null). Row II, column B shows an example of self-activation of *Kwa_23884*. Columns C and D show an example of positive *Kwa* PPI whose corresponding *Sce* PPI is also confirmed. Column E shows an example of negative *Kwa* PPI, whereas column F shows that the corresponding *Sce* PPI is also negative. Gene names starting with *Kwa* are *Kwa* genes; otherwise, they are *Sce* genes. One-to-one orthologous genes have the same color in gene name and are connected by lines.

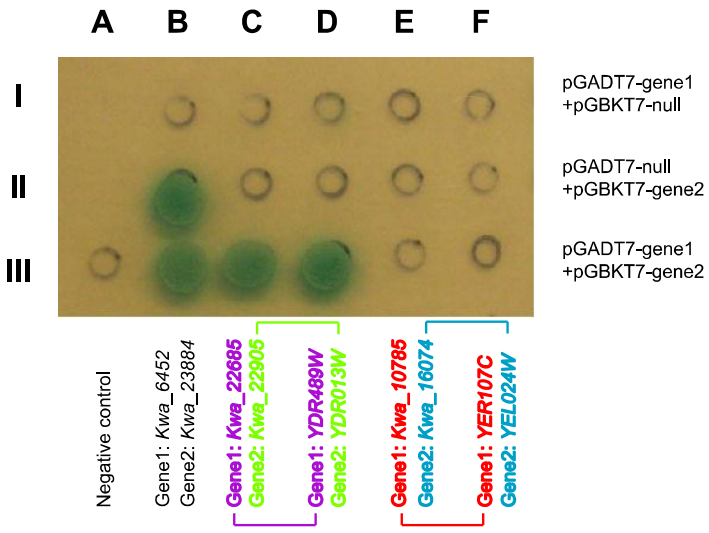


Figure 2.3 Maximum likelihood estimates (dots) and 95% confidence intervals (error bars) of the evolutionary rate of protein-protein interaction (PPI).

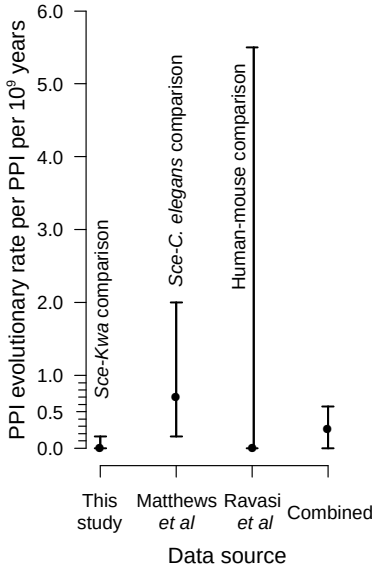


Figure 2.4 Comparisons in various gene properties between all 3152 *S. cerevisiae* genes that have one-to-one *K. waltii* orthologs (grey bars) and the 74 *S. cerevisiae* genes involved in the 43 PPIs measured for evolutionary conservation (black bars).

(A) A comparison in PPI degree (i.e., the number of PPIs that a gene has with the rest of the 3152 genes). (B) Frequency distribution of median PPI degrees of genes involved in 43 randomly sampled PPIs, derived from 1000 simulations. (C) A comparison in gene importance, measured by the fitness reduction caused by gene deletion. (D) Frequency distribution of median importance of the genes involved in 43 randomly sampled PPIs, derived from 1000 simulations. (E) A comparison in protein sequence identity between *S. cerevisiae* and *K. waltii*. (F) A comparison in the nonsynonymous/synonymous substitution rate ratio measured by comparing *S. cerevisiae* and *K. waltii* sequences. (G) A comparison in mRNA expression levels. (H) A comparison in protein expression levels. All *P* values are from Mann-Whitney tests, except those in panels B and D, which are from simulation tests.

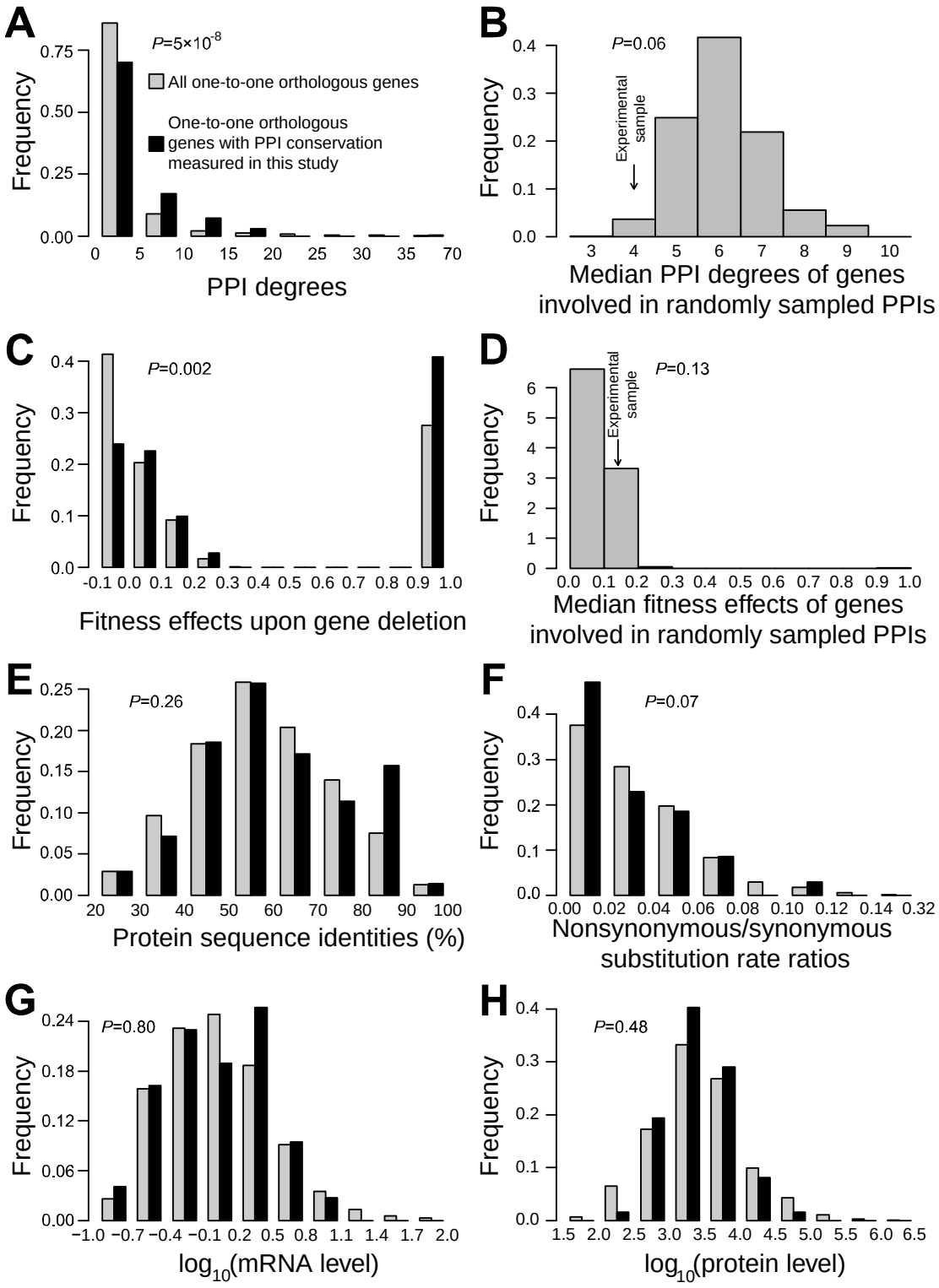

A
P=5×10-8
All one-to-one orthologous genes
One-to-one orthologous genes with PPI conservation measured in this study
Frequency
PPI degrees
B
P=0.06
Experimental sample
Frequency
Median PPI degrees of genes involved in randomly sampled PPIs
C
P=0.002
Frequency
Fitness effects upon gene deletion
D
P=0.13
Experimental sample
Frequency
Median fitness effects of genes involved in randomly sampled PPIs
E
P=0.26
Frequency
Protein sequence identities (%)
F
P=0.07
Frequency
Nonsynonymous/synonymous substitution rate ratios
G
P=0.80
Frequency
log10(mRNA level)
H
P=0.48
Frequency
log10(protein level)

Figure 2.5 Comparisons in Gene Ontology (GO) slim distributions between all 3152 *S. cerevisiae* genes that have one-to-one *K. waltii* orthologs (grey bars) and the 74 *S. cerevisiae* genes involved in the 43 PPIs measured for evolutionary conservation (black bars).

Bins with significant discrepancies at the false discovery rate of 5% are marked by stars.

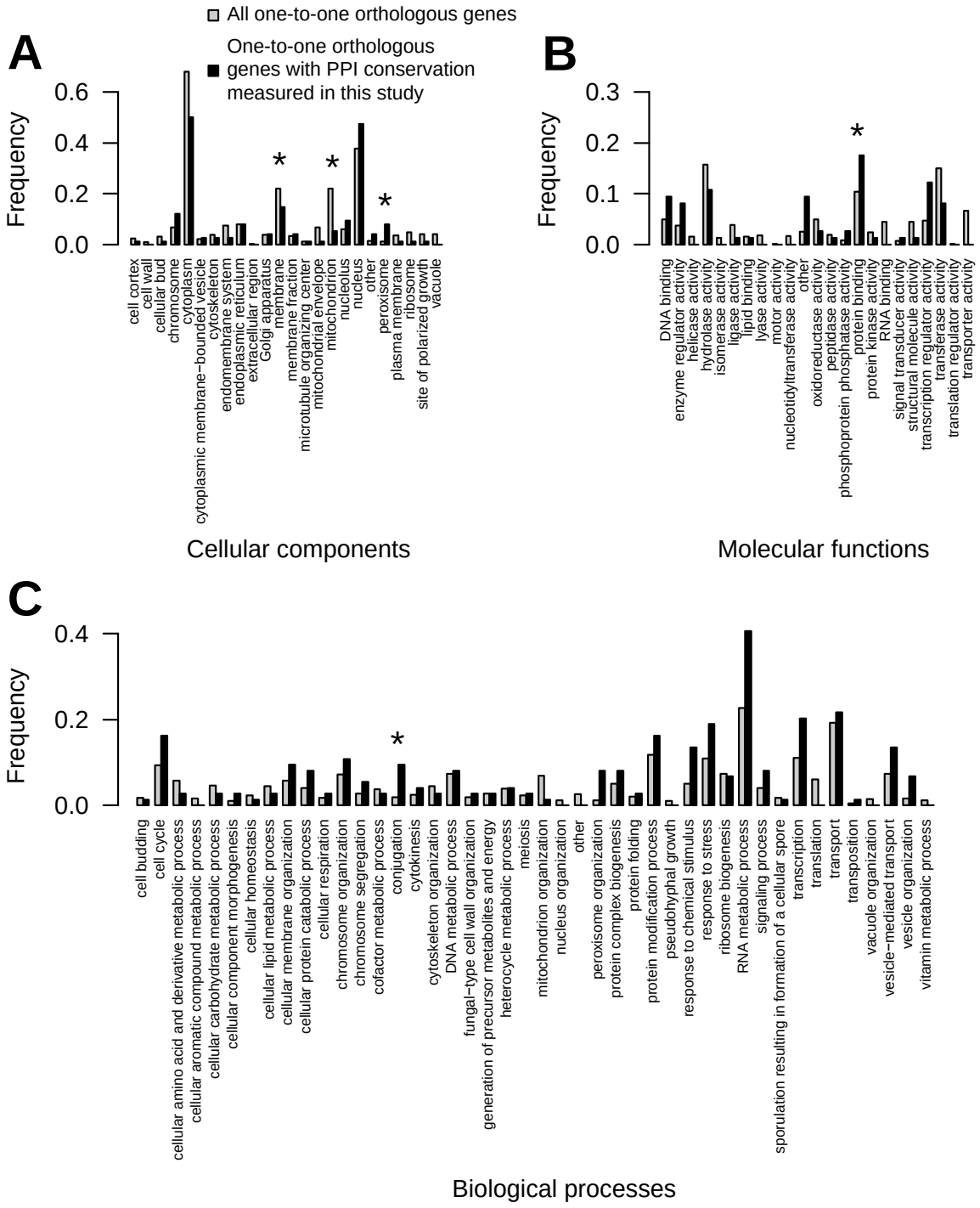
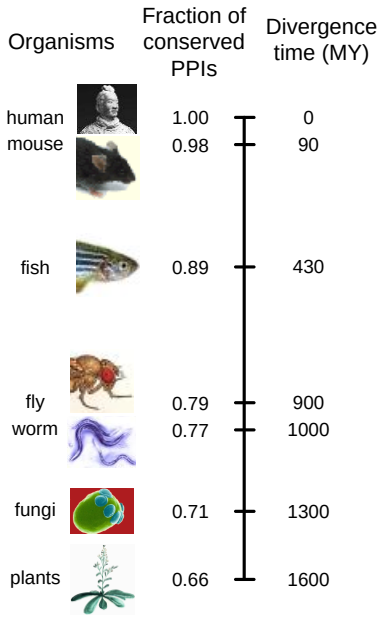


Figure 2.6 Fraction of human PPIs expected to be conserved in various widely used model organisms, based on previously estimated divergence times (Hedges et al. 2006) and our estimated rate of PPI evolution.



2.10 References

Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J. 2007. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res* **35**(Web Server issue): W91-96.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-29.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* **104**(18): 7489-7494.

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS et al. 2009. An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods* **6**(1): 91-97.

Chen YC, Rajagopala SV, Stellberger T, Uetz P. 2010. Exhaustive benchmarking of the yeast two-hybrid system. *Nat Methods* **7**(9): 667-668; author reply 668.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**(5652): 1960-1963.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**(12): 938-950.

Deane CM, Salwinski L, Xenarios I, Eisenberg D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* **1**(5): 349-356.

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**(4): 1915-1925.

Doolittle RF. 2010. The roots of bioinformatics in protein evolution. *PLoS Comput Biol* **6**(7): e1000875.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**(2): 341-352.

Fields S, Song O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* **340**(6230): 245-246.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* **296**(5568): 750-752.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* **425**(6959): 737-741.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E et al. 2003. A protein interaction map of Drosophila melanogaster. *Science* **302**(5651): 1727-1736.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.

He X, Zhang J. 2006. Why do hubs tend to be essential in protein networks? *PLoS Genet* **2**(6): e88.

He X, Zhang J. 2009. On the growth of scientific knowledge: yeast biology as a case study. *PLoS Comput Biol* **5**(3): e1000320.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971-2972.
Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* **411**(6841): 1046-1049.
Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**(5): 717-728.
Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* **411**(6833): 41-42.
Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12**(6): 962-968.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**(6983): 617-624.
Kim DU, Hayles J, Kim D, Wood V, Park HO, Won M, Yoo HS, Duhig T, Nam M, Palmer G et al. 2010. Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* **28**(6): 617-623.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**(5129): 624-626.
Kimura M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.
King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science* **164**(881): 788-798.
Koonin E, Galperin M. 2003. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* Kluwer Academic Publishers, Boston.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21): 2947-2948.
Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T et al. 2004. A map of the interactome network of the metazoan C. elegans. *Science* **303**(5657): 540-543.
Li W. 1997. *Molecular Evolution*. Sinauer, Sunderland, Mass.
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**(11): 2072-2080.
Liao BY, Weng MP, Zhang J. 2010a. Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci U S A* **107**(16): 7353-7358.
Liao BY, Weng MP, Zhang J. 2010b. Impact of extracellularity on the evolutionary rate of Mammalian proteins. *Genome Biol Evol* **2010**: 39-43.
Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.
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**(12): 2120-2126.

McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A* **107**(14): 6544-6549.

Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**(6): e170.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A* **106**(16): 6700-6705.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158**(2): 927-931.

Qian W, Zhang J. 2009. Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes. *Genome Biol Evol* **1**: 198-204.

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**(5): 744-752.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**(7062): 1173-1178.

Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* **25**(5): 210-216.

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**(1): 198-208.

Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**(7): 1283-1292.

Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet* **5**(1): e1000329.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A* **106**(18): 7273-7280.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634): 708-713.

Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* **6**: 421.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-1591.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N et al. 2008. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**(5898): 104-110.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Eco Evol* **18**(6): 292-298.

Zhang J. 2010. Evolutionary genetics: Progress and challenges. . In *Evolution Since Darwin: The First 150 Years*, (ed. MA Bell, DJ Futuyma, WF Eanes, JS Levinton), pp. 87-118. Sinauer, Sunderland, Mass.

Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* **22**(4): 1147-1155.

Zukerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, (ed. V Bryson, HJ Vogel), pp. 97-166. Academic Press, New York.

CHAPTER 3
ON THE BIOLOGICAL BASIS OF EPISTASIS

3.1 Abstract

Epistasis is a common genetic phenomenon with widespread and profound biological implications. Although high-throughput epistasis data from model organisms are being generated and used to construct genetic networks(Boone et al. 2007; Roguev et al. 2008; Tischler et al. 2008), to what extent epistasis reflects functional intimacy of involved genes is unclear(Cordell 2002; Moore and Williams 2005; Phillips 2008). We address this question using metabolic networks, where both epistasis and functional relationships of biochemical reactions can be evaluated through systemic analysis. We found that negative epistasis in fitness occurs mainly between nonessential reactions with overlapping functions, whereas positive epistasis usually involves essential reactions, is highly abundant, and surprisingly, often occurs between reactions without overlapping functions. We subsequently validated these results experimentally and identified their mechanistic basis. Our findings necessitate the distinction of the concept of genetic interaction from non-independent gene effects and call for reevaluation of evolutionary theories that depend on prevalent negative epistasis.

3.2 Introduction

Epistasis, a term coined by Bateson 100 years ago(Bateson 1909; Phillips 2008), refers to the phenomenon that the effect of a gene on a trait is masked or enhanced by one or more other genes. Fisher and other population and quantitative geneticists extended the concept to mean non-independent or non-multiplicative effects of genes(Fisher 1918; Phillips 2008). The direction, magnitude, and prevalence of epistasis is important for understanding gene function and interaction(Hartman et al. 2001; Boone et al. 2007; Phillips 2008), speciation(Coyne 1992), evolution of sex and recombination(Kondrashov 1988; Barton and Charlesworth 1998), evolution of ploidy(Kondrashov and Crow 1991),

mutation load(Crow and Kimura 1979), genetic buffering(Jasnos and Korona 2007), human disease(Cordell 2002; Moore and Williams 2005), and drug-drug interaction(Yeh et al. 2006).

Epistasis in fitness between two mutations is commonly defined by $\varepsilon = W_{XY} - W_X W_Y$, where W_X and W_Y represent the fitness values of two single mutants relative to the wild-type and W_{XY} represents the fitness of the corresponding double mutant. Epistasis is said to be positive when $\varepsilon > 0$, and negative when $\varepsilon < 0$. When deleterious mutations are concerned, positive epistasis lessens the fitness reduction predicted from individual mutational effects, whereas negative epistasis enhances it. The magnitude of epistasis between different pairs of mutations may be compared using scaled epistasis $\tilde{\varepsilon}$ (Segre et al. 2005), which is transformed from and has the same sign as ε, but is normally bounded between -1 and 1.

In this study, We apply the flux balance analysis (FBA) of metabolic networks(Price et al. 2004), a computational systems biology tool, to explore the functional association between biochemical reactions that are epistatic to each other. Assuming a steady state in metabolism, FBA maximizes the rate of biomass production under the stoichiometric matrix of all reactions and a set of flux constraints. The maximized rate in a mutant strain relative to that in the wild-type strain can be regarded as the Darwinian fitness of the mutant relative to the wild-type(Segre et al. 2005). FBA can be used to investigate the performance of the metabolic network and the fitness of the cell under various environmental and genetic perturbations(Ibarra et al. 2002; Papp et al. 2004) and has been used to generate the epistasis map of yeast metabolic genes(Segre et al. 2005; Deutscher et al. 2006; Harrison et al. 2007).

3.3 Quantifying epistasis by FBA

We first study the bacterium *Escherichia coli*, because its reconstructed metabolic network is of high quality and its FBA predictions have been empirically verified (Edwards et al. 2001; Ibarra et al. 2002). Using FBA, we identified from the *E. coli* metabolic network 270 reactions whose removal reduces the fitness under the glucose minimal medium. Removing any of the remaining 661 reactions has no such effect, primarily because the reaction has zero flux under this medium, or occasionally because

the network has another reaction that can fully compensate its loss. Among the 270 reactions, 212 are essential, meaning that deleting any one of them results in zero fitness. We considered a genetic perturbation in each reaction that constrains its flux to ≤50% of its wild-type optimal value and then computed the fitness of the mutant by FBA. We similarly computed the fitness values of all possible double mutants and obtained ε and $\tilde{\varepsilon}$ for all pairs of the 270 reactions, which reveal the global epistasis pattern within the metabolic network. Constraining the flux to ≤50% instead of zero(Segre et al. 2005; Deutscher et al. 2006; Harrison et al. 2007) allows the investigation of essential reactions. Consequently, the number of pairwise epistasis values obtained here exceeds 25 times that previously obtained(Segre et al. 2005). Constraining the flux to other non-zero levels does not alter our results qualitatively.

3.4 Establishing functional relation among reactions

To examine whether metabolic reactions with epistatic relationships are functionally associated, we need to identify the function of each reaction in generating the *E. coli* biomass, which is composed of 49 constituents. If a reaction is important for producing a set of biomass constituents, the removal of these constituents from the biomass function will recover the biomass reduction caused by the deletion of the reaction. Based on this idea, we designed a removal-recovery method to determine the functions of 255 of the 270 important reactions in generating biomass constituents (Figure 3.1A). For the remaining 15 reactions, the functions cannot be unambiguously determined and thus they are excluded from our analysis. The majority of the 255 reactions each contribute to only one biomass constituent, whereas a small number of reactions affect many or even all 49 constituents (Figure 3.1B). Note that the glucose minimal medium is again used in determining the function of each reaction, because some reactions have variable functions in different media. Functional assignment by our method is generally consistent with the conventional functional annotation(Reed et al. 2003), but our assignment is expected to be more objective and precise.

3.5 Distribution and underlying functional mechanisms of epistasis

We found 26 (0.08%) reaction pairs that show apparent negative epistasis ($\tilde{\varepsilon} \leq -0.01$). Among them, 25 pairs each share functions in producing at least one biomass constituent (Table 3.1; Figure 3.2 A & B). The remaining pair is between reactions MALS (catalyzed by malate synthase) and PPC (phosphoenolpyruvate carboxylase), anaplerotic reactions feeding the Krebs cycle. The lack of shared biomass constituents between them is due to the incomplete identification of MALS and PPC functions caused by their mutual functional compensation. A common interpretation of negative epistasis between two genes is that the two genes can individually perform a common function and thus each of them is able to compensate the loss of the other. Our observation that virtually every pair of reactions with negative epistasis share at least one function strongly support the above interpretation (Figure 3.2B). Although one might think that negative epistasis should occur between two nonessential reactions, this rule is not absolute. For example, two essential reactions (or one essential reaction and one nonessential reaction) may share a nonessential function in producing a biomass constituent and show negative epistasis by this common function (Table 3.1).

In contrast to the rare occurrence of negative epistasis, >97% of reaction pairs exhibit apparent positive epistasis ($\tilde{\varepsilon} \geq 0.01$) (Figure 3.2A). However, only ~26% of them occur between reactions that share at least one biomass constituent (Table 3.1; Figure 3.2C). There is also no significant difference in ε or $\tilde{\varepsilon}$ between functionally overlapping and non-overlapping reaction pairs with positive epistasis. It is often observed that a reaction is positively epistatic with a large number of apparently unrelated reactions. Use of ε instead of $\tilde{\varepsilon}$ in measuring epistasis does not change this pattern. The lack of functional overlap between most positively epistatic reaction pairs challenges the general interpretation of epistasis as functional association between involved parties(Szathmary 1993; Hartman et al. 2001; Boone et al. 2007).

Why does positive epistasis occur so frequently between functionally unrelated reactions? Figure 3.2A shows that virtually every essential reaction exhibits strong positive epistasis ($\tilde{\varepsilon} \sim 1$) with any other reaction regardless of its function and essentiality. This enigmatic observation has a simple explanation. A reaction is essential because, when its flux is constrained, at least one biomass constituent will have reduced

production that cannot be compensated by flux redistribution in the network. As a result, productions of all other biomass constituents must be reduced to maintain the composition stoichiometry of the biomass. In other words, when an essential reaction is constrained, almost all other reactions in the network do not work in their full capacity (Figure 3.3 A & B). Consequently, a genetic perturbation in a second reaction that reduces its capacity will have a negligible additional effect, causing positive epistasis. This is analogous to a barrel constructed of many wooden staves of equal height. The volume of liquid that can be stored in the barrel is reduced when a stave is shortened. Subsequent shortening of any other stave will not further reduce the volume as long as this stave is not shorter than the first shortened stave. The "barrel effect" explains why an essential reaction is positively epistatic with virtually every reaction, regardless of its function and essentiality (Figure 3.2A). Note that positive epistasis sometimes occurs between nonessential genes and in these cases ~80% (288/361) show functional overlaps (Figure 3.2B).

Why is there no barrel effect between nonessential reactions? There are three requirements for a metabolic reaction to be important yet nonessential. First, it must function in producing one or more biomass constituents. Second, there must be alternative reactions that can also make its product. Third, compared with the alternative reactions, it must be more efficient in producing at least one constituent. When the flux of a nonessential reaction is constrained, its less efficient alternative reaction will be turned on (Figure 3.3C). Due to the lower efficiency of the alternative reaction, nutrients that previously went through other reactions for making other biomass constituents can be redistributed in such a way that the biomass reduction by the flux constraint is minimized (Figure 3.3C). In the barrel analogy, when a stave is cut short, one can now take out a small piece of wood (representing redistributed nutrient flux) from each of the many other staves and attach it to the shortened stave to make all staves of equal height. Apparently, this way, there will be virtually no epistatic effect on the volume of the barrel between two cuts. It can be shown mathematically that when the number of reactions in the network is large, perturbations of two functionally unrelated nonessential reactions will have a nearly multiplicative effect on biomass production and cause negligibly weak positive epistasis(Segre et al. 2005; Jasnos and Korona 2007).

3.6 Analysis in yeast

Yeast (*Saccharomyces cerevisiae*) is another species whose reconstructed high-quality metabolic networks have been extensively validated experimentally(Papp et al. 2004; Harrison et al. 2007). We repeated the above analysis in yeast and obtained similar results (Table 3.1; Figure 3.4). Specifically, only 0.2% of reaction pairs show $\tilde{\varepsilon} \leq -0.01$, 83% of which have functional overlaps. By contrast, >95% of reaction pairs show $\tilde{\varepsilon} \geq 0.01$, but only 20% of which have overlapping functions.

3.7 Experimental validation in yeast

We further validated these results experimentally, with a focus on yeast, due to the difficulty in conducting partial gene deletion in *E. coli*. We deleted one allele per gene from a diploid cell to achieve partial disruption of a gene. Haploinsufficient genes were used to ensure that partial gene disruption affects fitness. Only non-metabolic genes were examined, because metabolic genes are rarely haploinsufficient(Deutschbauer et al. 2005). Non-metabolic genes should also be subject to the barrel effect(Kishony and Leibler 2003), as long as the final product is composed of multiple constituents with a fixed or preferred composition stoichiometry. We constructed a reference strain by inserting a yellow fluorescent protein gene into the genome of a wild-type strain. We then measured the fitness of each strain (including the wild-type) by competing it with the reference strain followed by cell counting using fluorescence activated cell sorting (FACS), a method that detects a fitness differential as small as 0.5%. We then calculated the fitness values of all single-deletion strains and all pairwise double-deletion strains relative to the wild-type, which allowed the estimation of epistasis between genes (Figure 3.5A; Table 3.2; Table 3.3). Six essential and two nonessential genes from seven functional categories were examined. Among the 27 gene pairs that involve at least one essential gene, 23 (85%) have significantly positive ε ($P < 0.05$, *t* test), two have significantly negative ε, and the remaining two do not show significant epistasis. The mean $\tilde{\varepsilon}$ among the 23 positively epistatic pairs is 0.78, and 11 of them have $\tilde{\varepsilon}$ not significantly smaller than 1. The epistasis between the two nonessential genes is not statistically significant. These results strongly support our computational findings.

Because the above experiment could not examine haplosufficient genes, we employed the newly developed DAmP method(Breslow et al. 2008) to mimic partial gene deletion, in which a marker gene is inserted into the 3' untranslated region of a gene such that its protein expression may be reduced to <50%. We studied 9 haplosufficient genes belonging to 8 functional categories, including 4 essential genes that are knocked down by DAmP and 5 nonessential genes that are knocked out (Table 3.2). We were able to measure the epistasis of 33 of the 36 gene pairs in haploid cells (Figure 3.5B; Table 3.4). Of the 23 gene pairs that have epistasis estimates and involve at least one essential gene, 20 (87%) show significantly positive ε ($P < 0.05$, *t* test), two show significantly negative ε, and the remaining one does not show significant epistasis (Figure 3.5B). These results further support our computational result of abundant positive epistasis involving essential genes, even among functionally unrelated ones.

3.8 Why previous studies did not find prevalent positive epistasis

Our computational and experimental results appear contradictory to previous studies of various model organisms, which did not find a gigantic number of positive epistasis cases(Elena and Lenski 1997; Segre et al. 2005; Boone et al. 2007; Jasnos and Korona 2007; Roguev et al. 2008; Tischler et al. 2008). The primary reason is that most of these studies used gene-deletion mutations and only estimated epistasis between nonessential genes, because essential genes are not amenable to such analysis(Segre et al. 2005; Jasnos and Korona 2007; Roguev et al. 2008; Tischler et al. 2008). The second reason is related to the sensitivity in detecting epistasis. Positive ε, having a maximum of $1-W_X W_Y$, is expected to be small despite the fact that $\tilde{\varepsilon}$ often approaches 1. Positive ε between natural mutations would be even smaller because most spontaneous mutations are only slightly deleterious. Such small positive ε is usually difficult to detect experimentally, unless the sensitivity of fitness measurement is high(Yeh et al. 2006; Breslow et al. 2008). Probably due to this difficulty, most previous studies were designed to identify negative epistasis only, even when essential genes were studied(Davierwala et al. 2005). Even in a study intended to detect both negative and positive epistasis involving essential genes, the low sensitivity did not allow reliable estimation of positive epistasis(Collins et al. 2007). While one study did observe more

positive epistasis than negative epistasis (with a ratio of 1.8 among statistically significant epistasis cases), the magnitude of the bias is not comparable to what we found here and the cause of the bias must be different because only nonessential genes were studied(Jasnos and Korona 2007). Recently, abundant positive epistasis between essential genes was detected upon improvement of the fitness measure, but the study was limited to functionally related genes(Breslow et al. 2008).

3.9 Conclusion

In summary, our systemic analysis of the *E. coli* and yeast metabolic networks and the subsequent experimental validations in yeast reveal a complex biological basis of epistasis. While negative epistasis accurately reflects functional intimacy of the involved reactions or genes, positive epistasis does not. Our results suggest a high prevalence of positive epistasis in living systems due to the barrel effect, an intrinsic property of essential components. The proportion of essential genes in the genome is ~7% in *E. coli*, 17% in yeast, and 55% in mouse(Liao and Zhang 2007). Thus, positive epistasis between functionally unrelated genes is likely to be even more prevalent in higher eukaryotes than is discovered here. These findings require the distinction of the concept of genetic interaction from non-multiplicative (or non-additive) gene effects and caution against the use of positive epistasis to infer genetic pathways and gene-gene interactions. While one may argue that, because all metabolic genes share the function in supporting cell growth, their epistasis is not surprising, we suggest that, if epistasis corresponds to such crude functional relationship, it provides little biological insight and is useless. Although our results are presented primarily with $\tilde{\varepsilon}$, it is clear that positive epistasis is highly abundant and much more prevalent than negative epistasis even when ε is used. This is also the case when the majority of mutations are only slightly deleterious. These observations and their theoretical basis call for reevaluation of evolutionary theories that depend on overall negative epistasis, such as the mutational deterministic hypothesis of the evolution of sexual reproduction(Kondrashov 1988) and the hypothesis of reduction in mutational load by truncation selection against deleterious mutations(Crow and Kimura 1979).

3.10 Materials and methods

3.10.1 Flux balance analysis (FBA) and minimization of metabolic adjustment (MOMA) were used to measure epistasis and identifying the functions of each reaction.

Details of FBA and MOMA have been described previously(Segre et al. 2002; Price et al. 2004). We used the optimization package CPLEX on the MATLAB platform to solve the programming problems. The *E. coli* metabolic model of iJR904 (Reed et al. 2003) and *S. cerevisiae* model of iMM904 (Mo et al. 2009) were used.

We first describe the analysis in *E. coli*. To delete reaction *i*, we set both of its upper-bound and lower-bound flux constraints to zero in FBA. To perturb a reaction, we set its upper-bound flux as 50% of its wild-type optimal flux. Essential reactions constitute 212/270 = 78% of all important reactions. We also studied the effect of different degrees of perturbation. Eleven of the 270 reactions have variable optimal fluxes in the wild-type. For these reactions, we used the minimal optimal fluxes so that any constraint in flux would be deleterious, allowing us to measure epistatic effects of deleterious mutations. Note that constraining the flux of a reaction to ≤50% of its wild-type level is not equivalent to constraining the enzyme concentration to ≤50% of its wild-type level, due to the nonlinear relationship between enzyme concentration and flux(Kacser and Burns 1981).

All pairwise double perturbations of the 270 important reactions were conducted. The relative fitness of a mutant is defined as the maximal biomass production rate of the mutant, relative to that of the wild-type(Segre et al. 2005). Epistasis is measured by $\varepsilon = W_{XY} - W_X W_Y$, where W_X and W_Y represent the fitness values of two single mutants relative to the wild-type and W_{XY} represents the fitness of the corresponding double mutant(Segre et al. 2005). Scaled epistasis(Segre et al. 2005) is defined by $\tilde{\varepsilon} = (W_{XY} / W_X W_Y) - 1$ when ε <0, and $\tilde{\varepsilon} = (W_{XY} - W_X W_Y) / [\min(W_X, W_Y) - W_X W_Y]$ when ε >0. Thus, $\tilde{\varepsilon}$ is normally between -1 and 1, although it can be >1 if W_{XY} is greater than $\min(W_X, W_Y)$.

When the fluxes of two reactions are constrained simultaneously, if both reactions have variable optimal fluxes, it is possible that their minimal optimal fluxes cannot be

simultaneously realized in the wild-type. In such instances, we actually over-constrained one of the two reactions in measuring W_{XY}, rendering ε underestimated and our conclusion of prevalent positive epistasis conservative. However, among the 255 reactions presented in Fig. 2, only 4 reactions have variable optimal fluxes and the pairwise ε values among them are all non-negative. Note that our epistasis measurement is completely independent from the identification of the function of each reaction. In other words, the observed abundance of positive epistasis is not dependent on the assumptions made in identifying the functions of metabolic reactions. The function of each reaction was identified by a removal-recovery method.

The yeast metabolic network contains 1,412 biochemical reactions, including 538 dead-end reactions. Using FBA, we found that 244 reactions have fitness effects upon deletion under the glucose minimal medium. We were able to identify the functions for 212 (158 essential and 54 nonessential) of these 244 reactions unambiguously, using the removal-recovery method. Epistasis between reactions was measured following the method used for the *E. coli* network.

3.10.2 Yeast experiments.

Haploid *Saccharomyces cerevisiae* strain BY4742 (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) and diploid strain BY4743 (*MAT***a***/MATα his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 lys2Δ0/+ met15Δ0/+ ura3Δ0/ura3Δ0*), both derived from the laboratorial strain s288c (Brachmann et al. 1998), were used in this study. The strains were grown on rich YPD media (1% yeast extract, 2% peptone, 2% glucose) or minimal synthetic dextrose (SD) media (0.67% yeast nitrogen base without amino acids, 2% glucose) with appropriate dropout (DO) supplements (Clontech, Mountain View, CA). 5-fluoro-orotic acid (5-FOA) agar media were made by mixing -Ura DO supplement (Clontech, Mountain View, CA), uracil (Sigma, St. Louis, MO, final concentration 50mg/L), and 5-FOA (Research Products International, Mount Prospect, IL, final concentration 0.1%) into SD agar media.

Two strategies were used. In the first strategy, we used the diploid strain BY4743 (Brachmann et al. 1998) as the wild-type strain and used either *URA3* or *LEU2* to replace one allele of a chosen gene in BY4743. Because the YPD media supplies both uracil and leucine, the replacement of target genes with a functional *URA3* or *LEU2* gene is

expected to have minimal fitness effect (see below for details). Note that the *LEU2* marker used in this study excludes the flanking tRNA genes that are commonly included in the *LEU2* marker. *URA3* was amplified from pRS416 (Stratagene, La Jolla, CA) and *LEU2* from PRS305 (American Type Culture Collection, Rockville, MD), using gene-specific primers containing ~60 nucleotides matching the sequences upstream and downstream of the open reading frame (ORF) of the gene to be replaced. The resulting cassette was used to replace the target gene using a homologous recombination based method(Amberg et al. 2005). *URA3*-inserted strains were selected on a uracil dropout synthetic media (SD –Ura), while *LEU2*-inserted strains were selected on a leucine dropout synthetic media (SD –Leu). Each target gene was independently replaced with both *LEU2* and *URA3*. Thus, 16 single gene replacement strains were made. We confirmed the status of heterozygous replacement by polymerase-chain reaction (PCR). Because 6 of the 8 genes are essential for growth in YPD, mating-based methods(Tong et al. 2001) cannot be used to make double-replacement strains. Instead, all 28 double-replacement strains were made by sequential replacement of two target genes with the two marker genes in diploid cells.

In the second strategy, we made DAmP strains following the original design(Breslow et al. 2008), except that we used *URA3* or *LEU2* rather than the Kan^R cassette as markers. We inserted the marker gene exactly after the stop codon of each gene. For nonessential genes, we deleted the ORF (from start codon to stop codon) by either *URA3* or *LEU2*. We used haploid BY4742 as the wild-type strain in this approach.

The reference strain was marked with the Venus variant of yellow fluorescent protein (vYFP)(Nagai et al. 2002) for fluorescent-activated cell sorting (FACS). vYFP was amplified from pBS7 (Yeast Resource Center, University of Washington, Seattle, WA) and introduced into plasmid p426GPD(Mumberg et al. 1995) using EcoRI and BamHI sites. vYFP proteins are expressed from an extremely strong promoter *GPD* in yeast (the promoter of *TDH3*) and with the *CYC1* terminator. We first replaced the ORF of *MET15* in BY4742 with *URA3* by PCR-based gene replacement method(Amberg et al. 2005) and selected it on a uracil dropout synthetic media (SD –Ura). We then replaced *URA3* with the vYFP gene (together with the *GPD* promoter and *CYC1* terminator) and selected it on 5-FOA plates. Yellow fluorescence was confirmed by live cell

fluorescence microscopy. All gene replacement strains were confirmed by genomic DNA extraction and PCR.

We used a competition assay to measure the fitness of each strain and then estimated epistasis and its confidence interval (see 3.11.2).

The fitness (*W*) values of the single heterozygous deletion strains estimated from our competition assay differ from previous estimates(Steinmetz et al. 2002; Deutschbauer et al. 2005). For two reasons, our results are more reliable than previous results. First, we measured the fitness based on counting hundreds of thousands of cells by flow cytometry, which is much more reliable than microarray-based method that was used in the previous fitness estimation. Second, we compared the deletion strains to a wild-type strain, which did not exist in previous fitness estimation.

The overwhelming positive epistasis we observed can potentially be a result of epistasis between the two selectable markers used in gene replacement (*LEU2* and *URA3*). To exclude this possibility, we used *LEU2* and *URA3* to each replace one allele of a non-functional gene (*HO*) in the diploid strain BY4743 and measured the epistasis between *LEU2* and *URA3*. *HO* encodes a site-specific endonuclease required for gene conversion at the *MAT* locus (homothallic switching). We selected *HO* for the following two reasons. First, if we simply insert the two marker genes into an intergenic region, it is possible that the marker genes destroy unknown functional elements in the region(Nagalakshmi et al. 2008) and cause unwanted fitness effects. Second, the *HO* gene in BY4743 and its ancestor s288c has several severe mutations and is apparently non-functional(Meiron et al. 1995; Ekino et al. 1999). Thus, replacing it with our marker genes will not have any unwanted side effect. We found no significant epistasis between *LEU2* and *URA3*. The fitness of the *URA3* insertion strain is W_{URA3}=1.014, the fitness of the *LEU2* insertion strain is W_{LEU2}=1.003, and the fitness of the *URA3* and *LEU2* double-insertion strain is $W_{URA3\text{-}LEU2}$=1.016. The epistasis between *URA3* and *LEU2* is ε=-0.001, $P > 0.9$, U test). Furthermore, the absolute value of ε between the two marker genes is small, compared with the majority of the epistasis values observed (Table 3.3).

3.11 Appendices

3.11.1 Selection of genes in the yeast experimental study

Two strategies are used to mimic partial gene disruption. The first strategy is to delete one allele of a gene from a diploid cell. Genes shown to be haploinsufficient in the YPD medium(Deutschbauer et al. 2005) were used to ensure that partial gene disruption affected fitness. The difference between $W_X W_Y$ and the smaller of W_X and W_Y is diminutive when either W_X or W_Y is close to 1, making positive epistasis difficult to detect. We chose genes that have relatively large fitness effects when one allele is deleted so that our assay would be sensitive enough when positive epistasis exists. Specifically, we required the heterozygous deletion strains to have fitness values no greater than 0.9 in each of the four previous measures(Steinmetz et al. 2002; Deutschbauer et al. 2005) (see Table 3.2). If two genes are located on the same chromosome, the fitness of the double heterozygous deletion strain may vary depending on whether the two null alleles are linked on the same chromosome. To avoid such potential complication, we selected genes located on different chromosomes. We chose 8 genes (Table 3.2) that meet all above criteria. They include 6 essential genes and 2 nonessential genes. The 8 genes belong to 7 different functional categories. Two ribosomal protein genes, one essential and one nonessential, share a functional category.

The second strategy to mimic partial gene deletion of essential genes is to use the *D*ecreased *A*bundance by *m*RNA *P*erturbation (DAmP) method(Breslow et al. 2008), in which a marker gene is inserted into the 3' untranslated region of a gene, causing 2.5 to >10-fold reduction in protein abundance(Breslow et al. 2008). This experiment was conducted in haploid cells. For nonessential genes, we simply deleted the ORF of the genes. We used an auxotrophic gene (*URA3* or *LEU2*) as selection markers to generate perturbation strains. We chose 8 essential genes with relatively strong fitness reduction found in previously constructed DAmP strains(Breslow et al. 2008) and chose 5 nonessential genes known to have fitness effects when deleted(Deutschbauer et al. 2005). Among these genes, 3 essential and 2 nonessential metabolic genes are included. We started with these 8 essential and 5 nonessential genes and measured the fitness of single perturbation strains. Four essential gene mutants showed fitness lower than 0.95 by *LEU2* and 2 showed fitness lower than 0.95 by *URA3*; none of these genes are metabolic

genes, probably due to the known insensitivity of fitness to enzyme dosage(Kacser and Burns 1981; Kondrashov and Koonin 2004). Deletion of each of the 5 nonessential genes had a fitness effect greater than 0.05. We thus examined epistasis among the above 9 genes (4 essential and 5 nonessential) in haploid cells by DAmP (for essential genes) and gene deletion (for nonessential genes) (Table 3.2). Except *GAA1* and *GAS1*, which are potentially related in function, other genes are not functionally related.

We caution that our experiments relied on the use of genes with relatively large fitness effects upon complete or partial deletion, rather than random genes.

3.11.2 Experimental estimation of epistasis and its confidence interval

We conducted competition experiments to measure the fitness of each yeast mutant strain relative to the vYFP-marked reference strain. Competitive growth of the two strains occurred in YPD at 30°C for 24 hours and the fitness for mitotic growth is measured. The detailed experimental procedure is as follows.

Yeast strains were first grown in YPD for 24 hours to acclimatize them in the competition environment. The strain of interest was then mixed with the vYFP-marked reference strain in YPD with a certain ratio. Different initial cell growth states may affect the fitness measurement. To avoid this problem, we measured the fraction of reference cells in the mixture after 6 hours of competition (P_0). We diluted the mixture by 50-fold at 18 hours to circumvent saturation and quantified the final fraction of the reference cells at 30 hours (P_1).

The fraction of cells in the culture that have yellow fluorescence was determined by cell counting using FACSCalibur (BD Biosciences, San Jose, CA) in the Flow Cytometry Core Facility at the University of Michigan. The data were collected from the FL1 detector, which has a filter with a 30-nm bandpass centered at 530nm. We first confirmed that our fluorescent strains and non-fluorescent strains can be separated successfully by FACS. For example, among 26,925 negative-control cells (BY4743), no cell showed positive signals, and among 38,775 positive-control cells (BY4742-vYFP), only 20 (or 0.0516%) were erroneously regarded as negative by FACS. For real samples, 100,000 gated events were counted per sample.

The number of cell divisions (or generations) during the competition assay was calculated by counting yeast cells on YPD plates. The competition mixture of the wild-type (BY4743) and reference (BY4742-vYFP) strains at 6 hours and 30 hours were diluted and plated on YPD agar media and colony numbers were counted (>300 colonies per mixture). Based on the ratio of the number of fluorescent and non-fluorescent cells, we calculated the generation numbers during the competition. From two replications, we found that the wild-type cells had 12.60 and 12.62 generations in 24 hours, respectively, and the reference cells had 11.54 and 11.61 generations, respectively. We used n = 11.58 generations in subsequent analysis of diploid strains, because the relative fitness of mutant strains were always directly measured against the reference strain. We note that the sign and magnitude of epistasis are insensitive to the n value. For example, use of n = 10.6 to 12.6 gave quantitatively similar results on epistasis. For measuring the fitness of haploid strains, the competition time was 20 hours and the generation number for the reference strain was n = 9.40.

A mutant strain's fitness (f'), relative to the reference strain, is calculated by

$$f' = 1 + \frac{\ln\left(\frac{P_0/(1-P_0)}{P_1/(1-P_1)}\right)}{n}.$$

The fitness (f_0) of the wild-type, relative to the reference strain, is similarly calculated. From two replications, we found that f_0 equals 1.039 and 1.036, respectively, for the diploid wild-type strain BY4743. The mean f_0 = 1.038 was used in subsequent analysis. For the haploid wild-type strain BY4742, f_0 = 1.010. A mutant strain's fitness relative to the wild-type is calculated by $W = f'/f_0$.

There are three potential sources of error that contribute to the estimation of the fitness of a strain. The first source is the sampling error in flow cytometry (i.e., the error in estimating P_0 and P_1). When counting k cells in flow cytometry, the standard deviation of P_1 is $\sqrt{P_1(1-P_1)/k}$ and the coefficient of variation of P_1 is $CV = \sqrt{(1-P_1)/(kP_1)}$. Because we counted at least k =100,000 cells, CV is approximately 0.003. For ~12 generations of growth, the fitness estimation error due to the sampling error of P_1 is approximately 0.003/12, or much lower than 0.001. Our experimental result is generally in agreement with this theoretical prediction. We measured cell ratios by FACS twice for

5 samples and found that the standard deviation of the estimated relative fitness contributed by the sampling error of P_1 is on average 0.0008 for the 5 samples. A similar level of error is expected in estimating P_0.

The second source of error is from mutations that occurred during the competition assay. At the end of the competition assay, the final cell number in the culture is about 3 ml $\times 2$ OD $\times (2.5 \times 10^7$ cells/ml/OD$) \approx 10^8$. That is, there are on the order of 10^8 cell divisions during the competition. Given the mutation rate of 10^{-10}/site/generation(Drake et al. 1998; Lang and Murray 2008) and the genome size of 10^7 nucleotides(Goffeau et al. 1996), the number of mutations during the competition is on the order of $10^8 \times 10^7 \times 10^{-10} = 10^5$. It is unclear how many of them are beneficial or deleterious and how large the fitness effects of these mutations are. We estimated the total effect of the first two sources of error by the following method. We measured the fitness of 9 groups of yeast cells from a wild-type BY4743 culture by separately competing them with the vYFP-marked reference strain. The standard deviation of the fitness of the wild-type strain from the 9 cultures is 0.0021. Thus, mutations during competition may have generated some noise for fitness measurement, but the first two sources of error amount to a relatively low standard deviation (~0.002) of fitness. We also confirmed that the initial relative cell numbers of the two strains in the mixture do not affect the estimated relative fitness values at any appreciable level.

The third source of error is mutations that occurred during the gene replacement procedure, in which the strain went through a single-cell bottleneck.

To estimate the total error from the three sources, we picked 13 different colonies from the same gene replacement experiment and measured the fitness of each of these colonies. We found that the standard deviation of fitness among the colonies is 0.0029. We further examined double-deletion strains and found the fitness variation among colonies of the same double-deletion strain to be 0.0016. Combining these data in a weighted manner, we estimated that the standard deviation of fitness among different colonies of the same strain to be 0.00256. Because this standard deviation includes all three sources of error, it best represents the degree of error in our fitness measurement and thus will be used in subsequent analysis.

The fitness values are stable between experiments at different times (6 months apart). We measured BY4742-vYFP's fitness relative to BY4743 (both strains were from reactivated glycerol stock strains stored in -80°C) twice and found that the fitness values were not significantly different.

The standard deviation of the estimated epistasis was calculated as follows. The variance (V) of each fitness measure is $V_f = V(f') = V(f_0) = (0.00256)^2 = 6.55 \times 10^{-6}$. The variance of the relative fitness of a mutant strain to the wild-type is

$$\begin{aligned} V(W) &= V(\frac{f'}{f_0}) = V(f')(\frac{1}{f_0})^2 + V(\frac{1}{f_0})(f')^2 + V(f')V(\frac{1}{f_0}) \\ &\approx V(f')(\frac{1}{f_0})^2 + V(\frac{1}{f_0})(f')^2 \\ &= V(f')(\frac{1}{f_0})^2 + (f')^2(\frac{1}{f_0})^4 V(f_0) \\ &\approx 2V_f \end{aligned} .$$

The approximations in the above formula are due to the fact that f_0 and f' are both close to 1 and $V(f') << (f')^2$. The variance of epistasis is then

$$\begin{aligned} V(\varepsilon) &= V(W_{XY} - W_X W_Y) = V(W_{XY}) + V(W_X W_Y) \\ &= V(W_{XY}) + V(W_X)W_Y^2 + V(W_Y)W_X^2 + V(W_X)V(W_Y) \end{aligned} .$$

Because W_X and W_Y are close to 1 and $V(W_X)$ and $V(W_Y)$ are small relative to W_X and W_Y, $V(\varepsilon) \approx 3V(W) \approx 6V_f$. Thus, the standard deviation of epistasis is $SD(\varepsilon) = \sqrt{V(\varepsilon)} = \sqrt{6V_f} = \sqrt{6} \times 0.00256 = 0.0063$. The 95% confidence interval of ε is $\varepsilon \pm 0.0123$.

3.12 Acknowledgements

Xionglei He and Zhi Wang performed FBA in this study and Ying Li helped with experiments. I thank Anuj Kumar for yeast strains and plasmids, Nike Bharucha, Gizem Kalay, Anuj Kumar, Jun Ma, and Barry Williams for advice and assistance in yeast experiments, Bernhard Palsson and his group for instruction on FBA, Ben-Yang Liao for drawing Figure 3.3, and Meg Bakewell, Soochin Cho, Wendy Grus, Ben-Yang Liao, and Calum Maclean for valuable comments. This work was partially supported by Block

grant from the Department of Ecology and Evolutionary Biology at University of Michigan.

Table 3.1 Numbers of reaction pairs that show epistatic relationships in glucose minimal medium.

		Epistasis in *E. coli***			Epistasis in yeast**		
Reaction pairs*	Functions	Negative	Zero	Positive	Negative	Zero	Positive
E-E	With overlap	0	9	4269	1	2	1780
	Without overlap	0	0	17667	0	3	10617
	Sum	0	9	21936	1	5	12397
E-N	With overlap	3	83	3704	10	67	2153
	Without overlap	0	34	5626	0	99	6203
	Sum	3	117	9330	10	166	8356
N-N	With overlap	22	267	288	24	137	402
	Without overlap	1	339	73	7	661	200
	Sum	23	606	361	31	798	602
All	With overlap	25	359	8261	35	206	4335
	Without overlap	1	373	23366	7	763	17020
	Sum	26	732	31627	42	969	21355

* Pairwise relationships among 255 important E. coli reactions and among 212 important yeast reactions. E, essential reaction; N, nonessential reaction.
** Scaled epistasis of ≥0.01 is considered positive, ≤0.01 is considered negative, and between 0.01 and 0.01 is considered zero.

Table 3.2 Genes used in the yeast experiments.

Genes used in the first set of yeast experiment							
ORF	Fitness 1*	Fitness 2*	Fitness 3*	Fitness 4*	Essential gene?	Gene name	Gene functions annotated in SGD (www.yeastgenome.org).
YAR002W	0.88	0.88	0.91	0.89	N	*NUP60*	Subunit of the nuclear pore complex (NPC), functions to anchor Nup2p to the NPC in a process controlled by the nucleoplasmic concentration of Gsp1p-GTP; potential Cdc28p substrate; involved in telomere maintenance
YDL193W	0.86	0.75	0.89	0.90	Y	*NUS1*	Prenyltransferase, required for cell viability; involved in protein trafficking
YFL039C	0.84	0.87	0.83	0.90	Y	*ACT1*	Actin, structural protein involved in cell polarization, endocytosis, and other cytoskeletal functions
YHR143W-A	0.82	0.84	0.83	0.86	Y	*RPC10*	RNA polymerase subunit, found in RNA polymerase complexes I, II, and III
YIL142W	0.86	0.88	0.89	0.89	Y	*CCT2*	Subunit beta of the cytosolic chaperonin Cct ring complex, related to Tcp1p, required for the assembly of actin and tubulins in vivo
YJR123W	0.82	0.77	0.86	0.84	Y	*RPS5*	Protein component of the small (40S) ribosomal subunit, the least basic of the non-acidic ribosomal proteins; phosphorylated in vivo; essential for viability; has similarity to E. coli S7 and rat S5 ribosomal proteins
YKL006W	0.82	0.84	0.89	0.88	N	*RPL14A*	N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl14Bp and has similarity to rat L14 ribosomal protein; rpl14a csh5 double null mutant exhibits synthetic slow growth
YPR181C	0.79	0.85	0.86	0.86	Y	*SEC23*	GTPase-activating protein; component of the Sec23p-Sec24p heterodimeric complex of the COPII vesicle coat, involved in ER to Golgi transport and autophagy; stimulates the GDP-bound form of Sar1p
Genes used in the second set of yeast experiment							

ORF	Fitness 1**	Fitness 2**	Metabolic gene?	Essential gene?	Gene name	Gene functions annotated in SGD (www.yeastgenome.org).
YBR243C	0.775	-	N	Y	*ALG7*	UDP-N-acetyl-glucosamine-1-P transferase, transfers Glc-Nac-P from UDP-GlcNac to Dol-P in the ER in the first step of the dolichol pathway of protein asparagine-linked glycosylation; inhibited by tunicamycin
YDR454C	0.759	-	N	Y	*GUK1*	Guanylate kinase, converts GMP to GDP; required for growth and mannose outer chain elongation of cell wall N-linked glycoproteins
YLR088W	0.710	-	N	Y	*GAA1*	Subunit of the GPI (glycosylphosphatidylinositol):protein transamidase complex, removes the GPI-anchoring signal and attaches GPI to proteins in the ER
YOL078W	0.734	-	N	Y	*AVO1*	Component of a membrane-bound complex containing the Tor2p kinase and other proteins, which may have a role in regulation of cell growth
YJL029C	0.568	0.773	N	N	*VPS53*	Component of the GARP (Golgi-associated retrograde protein) complex, Vps51p-Vps52p-Vps53p-Vps54p, which is required for the recycling of proteins from endosomes to the late Golgi; required for vacuolar protein sorting
YMR307W	0.706	0.747	N	N	*GAS1*	Beta-1,3-glucanosyltransferase, required for cell wall assembly; localizes to the cell surface via a glycosylphosphatidylinositol (GPI) anchor
YNL080C	0.662	0.762	N	N	*EOS1*	Protein involved in N-glycosylation; deletion mutation confers sensitivity to exidative stress and shows synthetic lethality with mutations in the spindle checkpoint genes BUB3 and MAD1; YNL080C is not an essential gene
YOL064C	0.787	0.680	Y	N	*MET22*	Bisphosphate-3'-nucleotidase, involved in salt tolerance and methionine biogenesis; dephosphorylates 3'-phosphoadenosine-5'-phosphate and 3'-phosphoadenosine-5'-phosphosulfate, intermediates of the sulfate assimilation pathway
YGR157W	0.523	0.854	Y	N	*CHO2*	Phosphatidylethanolamine methyltransferase (PEMT), catalyzes the first step in the conversion of phosphatidylethanolamine to phosphatidylcholine during the methylation pathway of phosphatidylcholine biosynthesis

* Fitness of heterozygous deletion strains in YPD were measured by two groups. Fitness 1 and Fitness 2 were measured by Steinmetz et al (2002), and Fitness 3 and Fitness 4 were measured by Deutschbauer et al (2005).

** Fitness of gene targeting strains in minimal media were measured by two groups. Fitness 1 were measured by Breslow et al (2008), and Fitness 2 were measured by Deutschbauer et al (2005).

Table 3.3 Fitness of double-deletion yeast strains and estimates of epistasis in the first set of experiment.

Double-deletion strains*	Fitness of double-deletion strains	Fitness of *LEU2* replacement strain	Fitness of *URA3* replacement strain	Expected multiplicative fitness	Epistasis ε**	Scaled epistasis $\tilde{\varepsilon}$
RPL14A & *NUP60*	0.886	0.898	0.991	0.889	-0.003	-0.004
RPL14A & *NUS1*	0.891	0.898	0.968	0.869	**0.023**	0.781
RPL14A & *ACT1*	0.867	0.898	0.891	0.800	**0.068**	0.742
RPL14A & *RPC10*	0.881	0.898	0.929	0.834	**0.047**	0.735
RPL14A & *CCT2*	0.890	0.898	0.918	0.824	**0.066**	0.894
RPL14A & *RPS5*	0.842	0.898	0.859	0.771	**0.071**	0.806
RPL14A & *SEC23*	0.874	0.898	0.921	0.827	**0.047**	0.663
NUP60 & *NUS1*	0.965	0.952	0.968	0.921	**0.044**	1.430
NUP60 & *ACT1*	0.872	0.952	0.891	0.848	**0.024**	0.570
NUP60 & *RPC10*	0.869	0.952	0.929	0.884	**-0.015**	-0.017
NUP60 & *CCT2*	0.924	0.952	0.918	0.874	**0.050**	1.140
NUP60 & *RPS5*	0.853	0.952	0.859	0.818	**0.035**	0.851
NUP60 & *SEC23*	0.904	0.952	0.921	0.877	**0.027**	0.604
NUS1 & *ACT1*	0.876	0.946	0.891	0.843	**0.034**	0.702
NUS1 & *RPC10*	0.867	0.946	0.929	0.879	-0.012	-0.013
NUS1 & *CCT2*	0.923	0.946	0.918	0.868	**0.055**	1.108
NUS1 & *RPS5*	0.852	0.946	0.859	0.813	**0.039**	0.837
NUS1 & *SEC23*	0.899	0.946	0.921	0.872	**0.027**	0.547
ACT1 & *RPC10*	0.856	0.906	0.929	0.842	**0.0143**	0.222
ACT1 & *CCT2*	0.866	0.906	0.918	0.832	**0.035**	0.464
ACT1 & *RPS5*	0.859	0.906	0.859	0.779	**0.080**	0.992
ACT1 & *SEC23*	0.865	0.906	0.921	0.835	**0.030**	0.420
RPC10 & *CCT2*	0.921	0.953	0.918	0.875	**0.046**	1.080
RPC10 & *RPS5*	0.848	0.953	0.859	0.819	**0.029**	0.716
RPC10 & *SEC23*	0.864	0.953	0.921	0.878	**-0.014**	-0.016
CCT2 & *RPS5*	0.850	0.938	0.859	0.806	**0.044**	0.829
CCT2 & *SEC23*	0.870	0.938	0.921	0.865	0.005	0.093
RPS5 & *SEC23*	0.867	0.888	0.921	0.818	**0.048**	0.694

* The first gene is replaced by LEU2 and second gene is replaced by URA3.

** Bold epistasis ε values are significantly different from 0 ($P<0.05$).

Table 3.4 Fitness of double-perturbation yeast strains and estimates of epistasis in the second set of experiment.

Double-deletion strains*	Fitness of Double perturbation strains	Fitness of *LEU2* replac./insert. strain	Fitness of *URA3* replac./insert. strain	Expected multiplicative fitness	Epistasis ε**	Scaled epistasis $\tilde{\varepsilon}$
ALG7 & *GAA1*	0.858	0.898	0.892	0.801	**0.058**	0.631
ALG7 & *VPS53*	0.814	0.898	0.861	0.773	**0.042**	0.472
ALG7 & *GAS1*	0.698	0.898	0.726	0.652	**0.046**	0.622
ALG7 & *EOS1*	0.883	0.898	0.879	0.789	**0.094**	1.048
ALG7 & *MET22*	0.852	0.898	0.885	0.794	**0.058**	0.643
ALG7 & *CHO2*	0.895	0.898	0.967	0.868	**0.027**	0.921
GUK1 & *ALG7*	0.852	0.903	0.932	0.841	0.011	0.179
GUK1 & *GAA1*	0.839	0.903	0.892	0.805	**0.034**	0.388
GUK1 & *VPS53*	0.000	0.903	0.861	0.777	**-0.777**	-1.000
GUK1 & *GAS1*	0.679	0.903	0.726	0.656	**0.023**	0.332
GUK1 & *EOS1*	0.806	0.903	0.879	0.793	**0.012**	0.144
GUK1 & *MET22*	0.681	0.903	0.885	0.798	**-0.117**	-0.147
GUK1 & *CHO2*	0.894	0.903	0.967	0.873	**0.021**	0.709
GAA1 & *VPS53*	0.000	0.831	0.861	0.715	**-0.715**	-1.000
GAA1 & *GAS1*	0.695	0.831	0.726	0.603	**0.092**	0.748
GAA1 & *EOS1*	0.800	0.831	0.879	0.730	**0.069**	0.690
GAA1 & *MET22*	0.783	0.831	0.885	0.735	**0.048**	0.502
GAA1 & *CHO2*	0.825	0.831	0.967	0.804	**0.021**	0.779
AVO1 & *ALG7*	0.925	0.929	0.932	0.865	**0.060**	0.940
AVO1 & *GAA1*	0.887	0.929	0.892	0.829	**0.059**	0.922
AVO1 & *VPS53*	0.814	0.929	0.861	0.799	**0.014**	0.231
AVO1 & *GAS1*	0.721	0.929	0.726	0.675	**0.047**	0.898
AVO1 & *EOS1*	0.000	0.929	0.879	0.816	**-0.816**	-1.000
AVO1 & *MET22*	0.000	0.929	0.885	0.822	**-0.822**	-1.000
AVO1 & *CHO2*	0.769	0.929	0.967	0.898	**-0.130**	-0.145
VPS53 & *GAS1*	0.646	0.901	0.726	0.654	-0.009	-0.013
VPS53 & *EOS1*	0.656	0.901	0.879	0.792	**-0.136**	-0.172
VPS53 & *MET22*	0.872	0.901	0.885	0.797	**0.075**	0.853
VPS53 & *CHO2*	0.812	0.901	0.967	0.872	**-0.060**	-0.068
GAS1 & *EOS1*	0.601	0.688	0.879	0.604	-0.004	-0.006
GAS1 & *MET22*	0.618	0.688	0.885	0.608	0.010	0.121
GAS1 & *CHO2*	0.638	0.688	0.967	0.665	**-0.027**	-0.041
EOS1 & *MET22*	0.699	0.798	0.885	0.706	-0.007	-0.010
EOS1 & *CHO2*	0.719	0.798	0.967	0.772	**-0.053**	-0.069
MET22 & *CHO2*	0.858	0.865	0.967	0.837	**0.021**	0.736

* The first gene is replaced/inserted with *LEU2* and second gene is replaced/inserted with *URA3*.

** Bold epistasis ε values are significantly different from 0 (P<0.05).

Figure 3.1 Functions of *E. coli* metabolic reactions under the glucose minimal medium.

(A) Functions of 255 important reactions in producing 49 biomass constituents. Columns represent reactions and rows represent biomass constituents. (B) Distribution of the number of biomass constituents affected by a reaction.

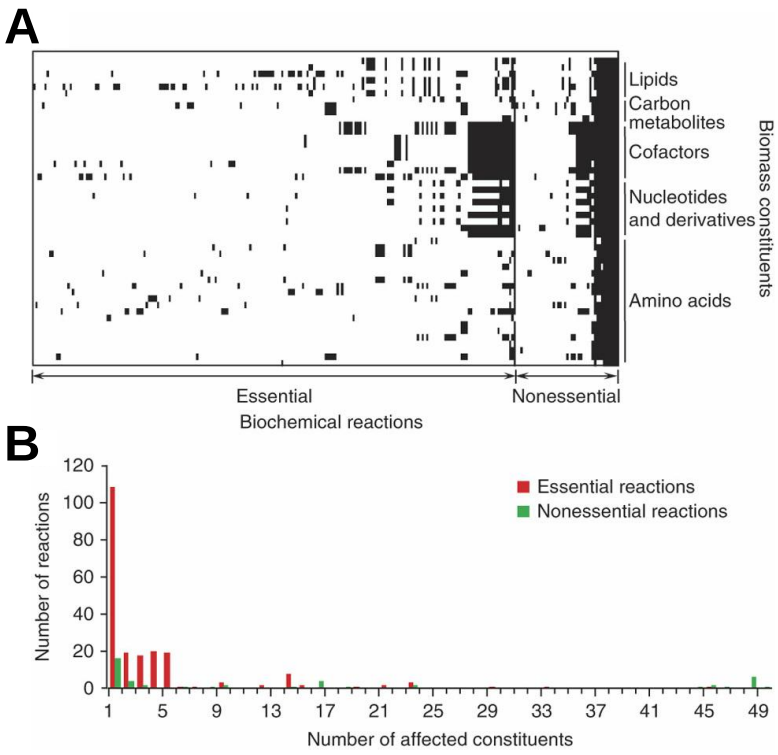


Figure 3.2 Pairwise epistasis and functional association among 255 important reactions in *E. coli*.

(A) An overview of epistasis and functional association among reactions. Both rows and columns represent reactions. Scaled epistasis between reactions is shown in the lower-left triangle by the heat map. Functional association between reactions is presented in the upper-right triangle, where a grey dot is shown when two reactions have overlapping functions. Epistasis and reaction functions are both determined in the glucose minimal medium. (B) Frequency distribution of scaled epistasis between nonessential reactions. (C) Frequency distribution of scaled epistasis between two reactions that include at least one essential reaction. E, essential; N, nonessential. Note the difference in Y-scale between panel b and c.

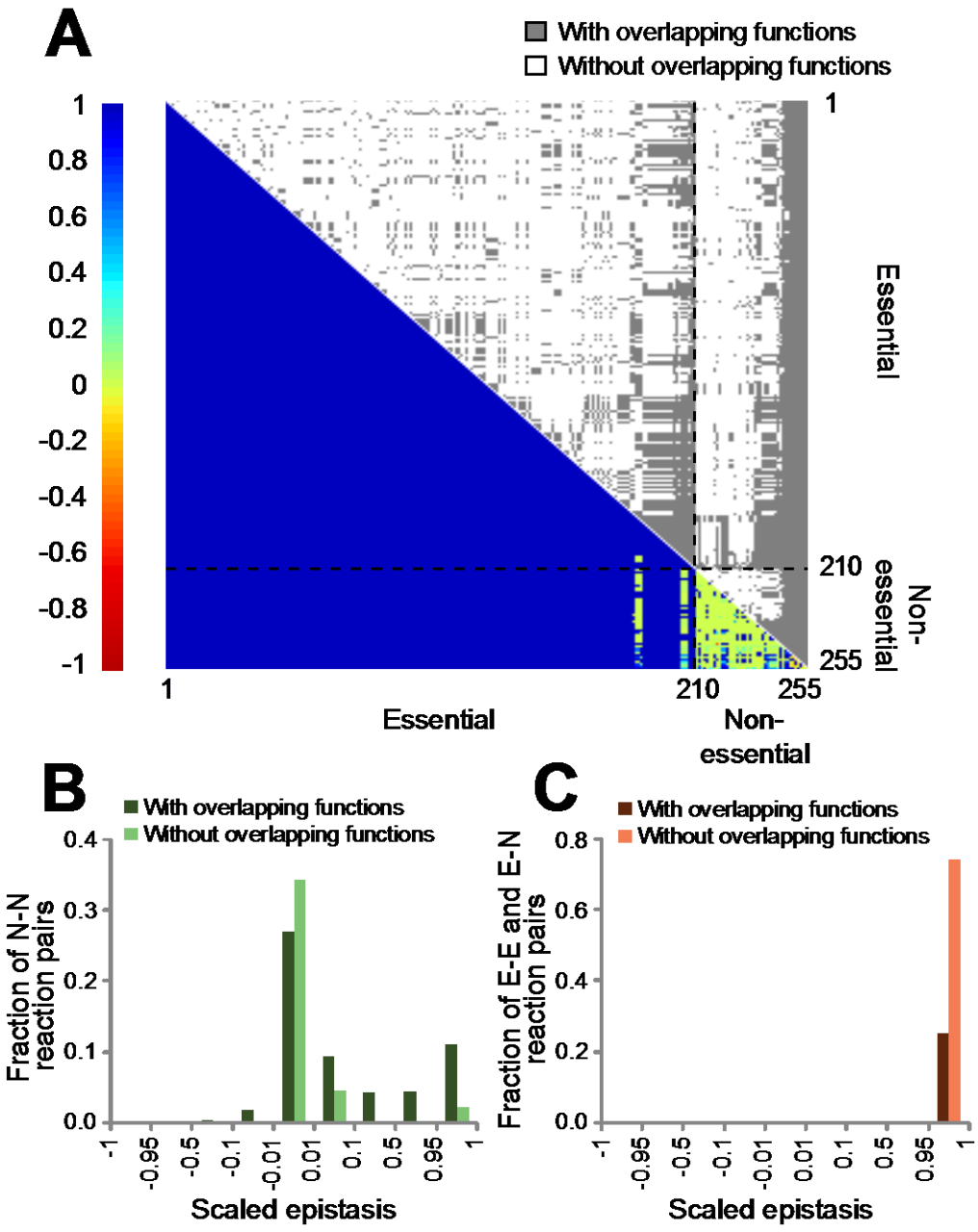

A
With overlapping functions
Without overlapping functions
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1
210
255
Essential
Non-essential
B
With overlapping functions
Without overlapping functions
Fraction of N-N reaction pairs
0.4
0.3
0.2
0.1
0.0
-1
-0.95
-0.5
-0.1
-0.01
0.01
0.1
0.5
0.95
1
Scaled epistasis
C
With overlapping functions
Without overlapping functions
Fraction of E-E and E-N reaction pairs
0.8
0.6
0.4
0.2
0.0
Scaled epistasis

Figure 3.3 A simple metabolic network that illustrates the mechanism underlying the different consequences on biomass production between constraining an essential reaction and constraining a nonessential reaction.

(A) Metabolic fluxes in the wild-type cell. Here the biomass is composed of three constituents with a 1:1:1 composition stoichiometry. Reactions #1-#5 contribute to the biomass, whereas #6 produces a non-biomass compound. Reactions #1 and #3 are nonessential, due to the presence of alternative reactions #2 and #4, respectively. Reaction #5 is essential. Reactions #2 and #4 are less efficient than #1 and #3, respectively, owing to the generation of byproducts that are not biomass constituents. In the wild-type, reactions #2 and #4 are silenced because they are less efficient than #1 and #3, respectively. The rate of biomass production is 10 units per unit time. (B) Metabolic fluxes in a mutant where the flux of the essential reaction #5 is constrained to ≤50% of its wild-type flux. Composition stoichiometry of the biomass renders the production of all biomass constituents halved, causing the fluxes of #1 and #3 also halved. The biomass production rate becomes 5 units per unit time (i.e., the relative fitness = 0.5). If the nutrient uptake rate is the same as in the wild-type, the extra nutrient absorbed is used to produce non-biomass compounds (black diamonds). In this mutant, a second mutation that constrains the flux of another reaction (e.g., #1) to ≤50% of its wild-type flux will not cause an additional fitness reduction, generating strong positive epistasis. This "barrel effect" may also occur in non-metabolic systems(Kishony and Leibler 2003). (C) Metabolic fluxes in a mutant where the flux of the nonessential reaction #1 is constrained to ≤50% of its wild-type flux. Reaction #2 is now turned on because of the constraint of #1. Byproducts (black circle) are produced from #2. Fluxes can be redistributed in such a way that the reduction in biomass production caused by the constraint on #1 is minimized. Consequently, the total flux of #1 and #2 exceeds that in the wild-type. It can be shown mathematically that the fitness reduction is tiny when the number of biomass constituents is large. In this mutant, a second mutation that constrains the flux of another nonessential reaction (e.g., #3) to ≤50% of its wild-type flux will cause an additional tiny fitness reduction, generating virtually no epistasis.

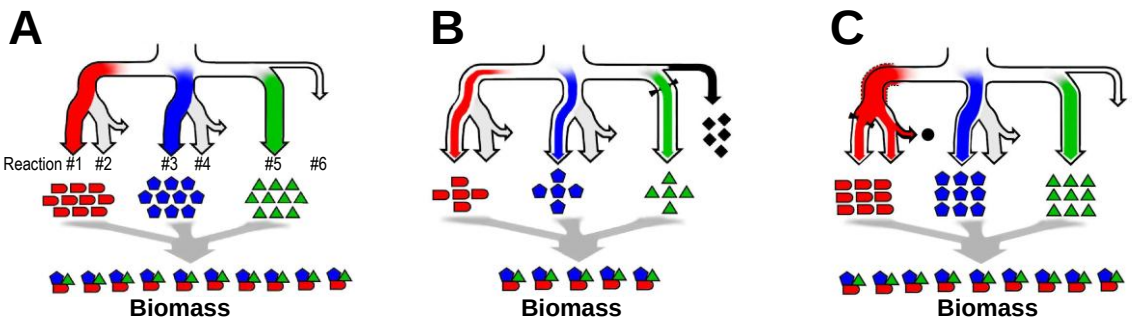


Figure 3.4 Pairwise epistasis and functional association among 212 important reactions in yeast.

(A) Frequency distribution of scaled epistasis between nonessential reactions. (B) Frequency distribution of scaled epistasis between two reactions that include at least one essential reaction. E, essential; N, nonessential. Note the difference in Y-scale between panel A and B.

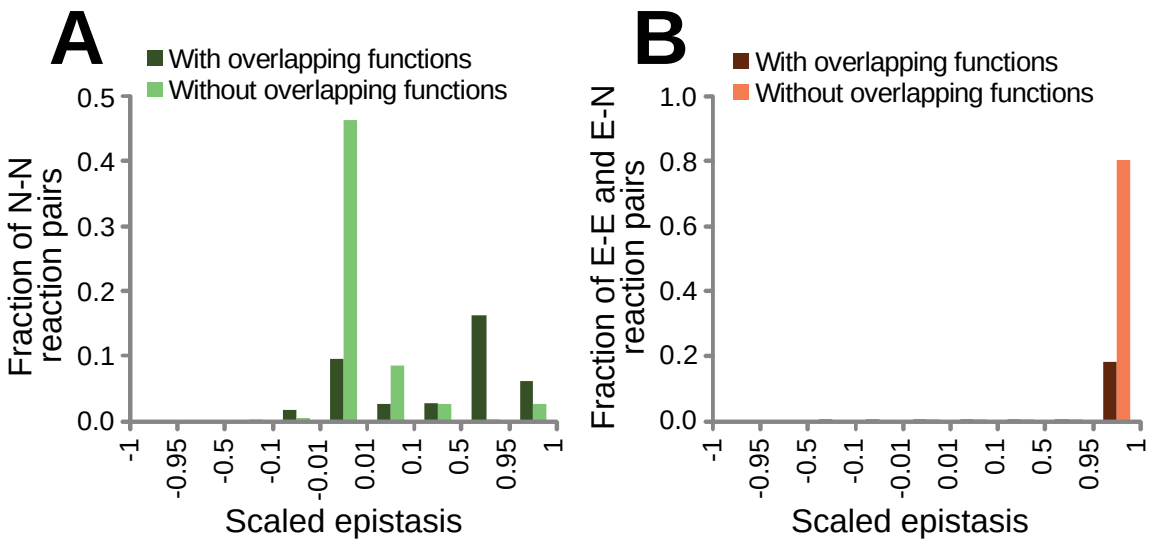
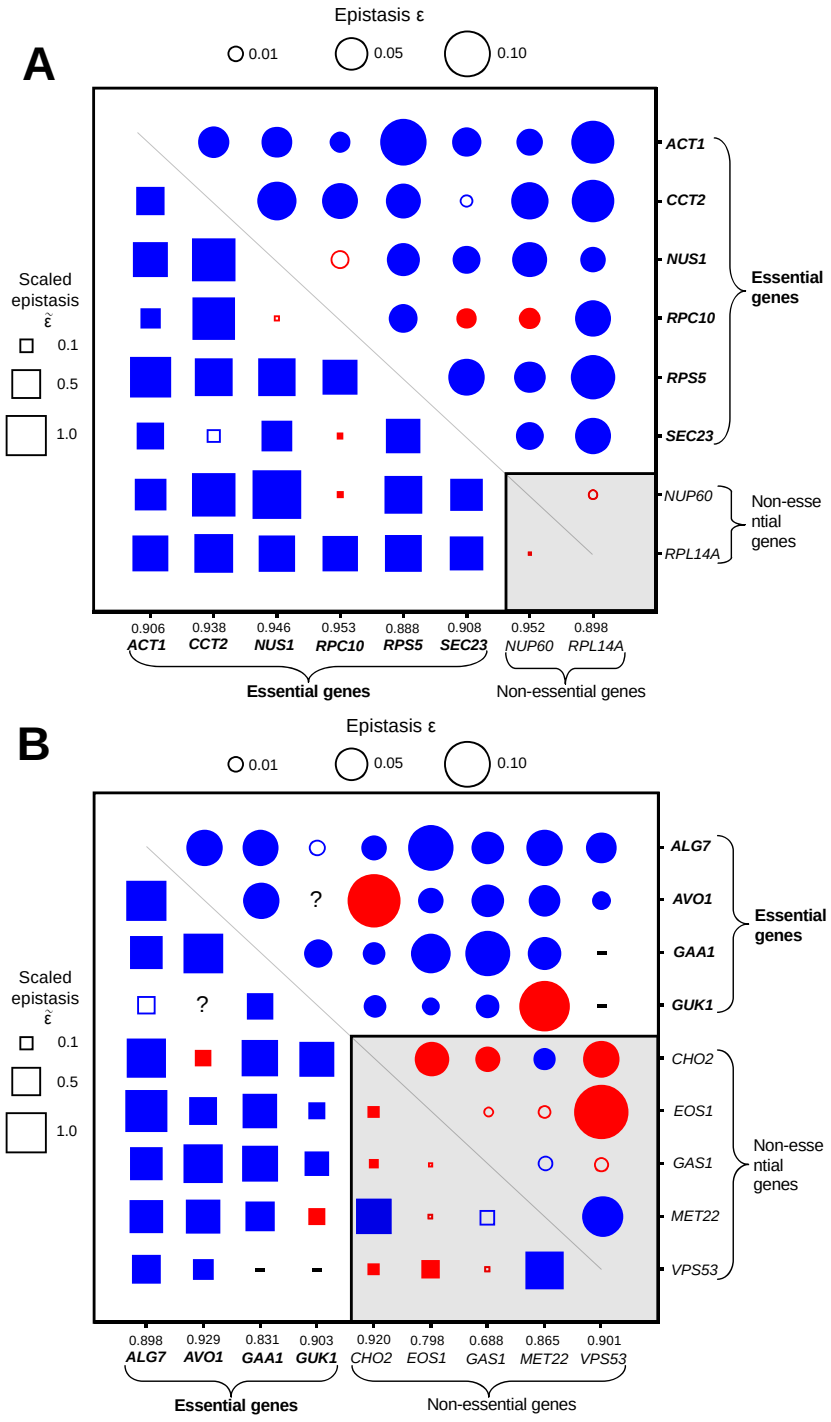


Figure 3.5 Epistasis (ε) and scaled epistasis ($\tilde{\varepsilon}$) among 17 yeast genes tested.

Circles in the upper-right half of the figure show ε, whereas squares in the lower-left half show $\tilde{\varepsilon}$. Blue and red colors indicate positive and negative epistasis, respectively, whereas the areas of the circles and squares are proportional to the absolute values of ε and $\tilde{\varepsilon}$, respectively, with the scales given on the top and left sides of each panel. Solid symbols indicate statistically significant epistasis ($P < 0.05$), whereas open symbols indicate insignificant epistasis. The shaded area in the lower-right corner shows relationships between nonessential genes. Fitness values of strains with genes replaced/inserted by *LEU2*, relative to the wild-types, are presented on the X-axis. (A) Epistasis among 8 haploinsufficient genes, measured in diploid cells after deletion of one allele per gene. All genes belong to different functional categories with the exception of *RPS5* and *RPL14A*, both of which encode ribosomal proteins. (B) Epistasis among 9 haplosufficient genes, measured in haploid cells after reduction of protein expression of essential genes and deletion of nonessential genes. All genes belong to different functional categories with the exception of *GAA1* and *GAS1*; the former encodes a subunit of the GPI (glycosylphosphatidylinositol):protein transamidase complex, whereas the latter requires a GPI anchor for protein subcellular localization. *MET22* and *CHO2* are metabolic genes, with FBA-predicted scaled epistasis equal to 1. "-", double-perturbation cells could not be obtained, likely due to unsuccessful experiments or synthetic lethality. "?", epistasis could not be measured due to the lack of fitness effect of single perturbations. Negative epistasis appears more abundant than expected between nonessential genes, concentrated on *CHO2* and *VPS53*. An examination of all experimentally determined negative epistasis cases to date shows that these two genes have more negative epistatic partners than typical yeast genes have.


A
Epistasis ε
0.01
0.05
0.10
Scaled epistasis ε̃
0.1
0.5
1.0
ACT1
CCT2
NUS1
RPC10
RPS5
SEC23
NUP60
RPL14A
Essential genes
Non-essential genes
0.906
0.938
0.946
0.953
0.888
0.908
0.952
0.898
B
Epistasis ε
0.01
0.05
0.10
Scaled epistasis ε̃
0.1
0.5
1.0
ALG7
AVO1
GAA1
GUK1
CHO2
EOS1
GAS1
MET22
VPS53
Essential genes
Non-essential genes
0.898
0.929
0.831
0.903
0.920
0.798
0.688
0.865
0.901

3.13 References

Amberg DC, Burke DJ, Strathern JN. 2005. *Methods in yeast genetics, a cold spring harbor laboratory course manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Barton NH, Charlesworth B. 1998. Why sex and recombination? *Science* **281**(5385): 1986-1990.

Bateson W. 1909. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.

Boone C, Bussey H, Andrews BJ. 2007. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**(6): 437-449.

Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD. 1998. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**(2): 115-132.

Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-Ornstein J, Newman HW, Braun S, Madhani HD, Krogan NJ, Weissman JS. 2008. A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5**(8): 711-718.

Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M et al. 2007. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**(7137): 806-810.

Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**(20): 2463-2468.

Coyne JA. 1992. Genetics and speciation. *Nature* **355**(6360): 511-515.

Crow JF, Kimura M. 1979. Efficiency of truncation selection. *Proc Natl Acad Sci U S A* **76**(1): 396-399.

Davierwala AP, Haynes J, Li Z, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding H, Zhu H, Chen Y et al. 2005. The synthetic genetic interaction spectrum of essential genes. *Nat Genet* **37**(10): 1147-1152.

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**(4): 1915-1925.

Deutscher D, Meilijson I, Kupiec M, Ruppin E. 2006. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* **38**(9): 993-998.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* **148**(4): 1667-1686.

Edwards JS, Ibarra RU, Palsson BO. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19**(2): 125-130.

Ekino K, Kwon I, Goto M, Yoshino S, Furukawa K. 1999. Functional analysis of HO gene in delayed homothallism in Saccharomyces cerevisiae wy2. *Yeast* **15**(6): 451-458.

Elena SF, Lenski RE. 1997. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**(6658): 395-398.

Fisher RA. 1918. The correlations between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* **52**: 399-433.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al. 1996. Life with 6000 genes. *Science* **274**(5287): 546, 563-547.
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**(7): 2307-2312.
Hartman JL, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. *Science* **291**(5506): 1001-1004.
Ibarra RU, Edwards JS, Palsson BO. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**(6912): 186-189.
Jasnos L, Korona R. 2007. Epistatic buffering of fitness loss in yeast double deletion strains. *Nat Genet* **39**(4): 550-554.
Kacser H, Burns JA. 1981. The molecular basis of dominance. *Genetics* **97**(3-4): 639-666.
Kishony R, Leibler S. 2003. Environmental stresses can alleviate the average deleterious effect of mutations. *J Biol* **2**(2): 14.
Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**(6198): 435-440.
Kondrashov AS, Crow JF. 1991. Haploidy or diploidy: which is better? *Nature* **351**(6324): 314-315.
Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20**(7): 287-290.
Lang GI, Murray AW. 2008. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* **178**(1): 67-82.
Liao BY, Zhang J. 2007. Mouse duplicate genes are as essential as singletons. *Trends Genet* **23**(8): 378-381.
Meiron H, Nahon E, Raveh D. 1995. Identification of the heterothallic mutation in HO-endonuclease of S. cerevisiae using HO/ho chimeric genes. *Curr Genet* **28**(4): 367-373.
Mo ML, Palsson BO, Herrgard MJ. 2009. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* **3**: 37.
Moore JH, Williams SM. 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* **27**(6): 637-646.
Mumberg D, Muller R, Funk M. 1995. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**(1): 119-122.
Nagai T, Ibata K, Park ES, Kubota M, Mikoshiba K, Miyawaki A. 2002. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol* **20**(1): 87-90.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.
Papp B, Pal C, Hurst LD. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**(6992): 661-664.

Phillips PC. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**(11): 855-867.
Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**(11): 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**(9): R54.
Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park HO, Hayles J et al. 2008. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* **322**(5900): 405-410.
Segre D, Deluna A, Church GM, Kishony R. 2005. Modular epistasis in yeast metabolism. *Nat Genet* **37**(1): 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**(23): 15112-15117.
Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ et al. 2002. Systematic screen for human disease genes in yeast. *Nat Genet* **31**(4): 400-404.
Szathmary E. 1993. Do deleterious mutations act synergistically? Metabolic control theory provides a partial answer. *Genetics* **133**(1): 127-132.
Tischler J, Lehner B, Fraser AG. 2008. Evolutionary plasticity of genetic interaction networks. *Nat Genet* **40**(4): 390-391.
Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**(5550): 2364-2368.
Yeh P, Tschumi AI, Kishony R. 2006. Functional classification of drugs by properties of their pairwise interactions. *Nat Genet* **38**(4): 489-494.

CHAPTER 4

BALANCED CODON USAGE OPTIMIZES EUKARYOTIC TRANSLATIONAL EFFICIENCY

4.1 Abstract

Cellular efficiency in protein translation is an important fitness determinant in rapidly growing organisms. It is widely believed that synonymous codons are translated with unequal speeds and that translational efficiency is maximized by the exclusive use of rapidly translated codons. Here we estimate the *in vivo* translational speeds of all sense codons from the budding yeast *Saccharomyces cerevisiae*. Surprisingly, preferentially used codons are not translated faster than unpreferred ones. We hypothesize that this phenomenon is a result of codon usage in proportion to cognate tRNA concentrations, the optimal strategy in enhancing translational efficiency under tRNA shortage. Our predicted codon-tRNA balance is indeed observed from all model eukaryotes examined, and its impact on translational efficiency is further validated experimentally. Our study reveals a previously unsuspected mechanism by which unequal codon usage increases translational efficiency, demonstrates widespread natural selection for translational efficiency, and offers new strategies to improve synthetic biology.

4.2 Introduction

Eighteen of the 20 amino acids are each encoded by two or more synonymous codons in the standard genetic code, yet the synonymous codons are often used unequally in a genome. Such codon usage bias (CUB) has been extensively documented in all three domains of life (Ikemura 1985; Sharp et al. 1988; Hershberg and Petrov 2009). Within a genome, highly expressed genes tend to have stronger CUB than lowly expressed ones

(Ikemura 1981), and the codons preferentially used in highly expressed genes of a species are referred to as *preferred codons*.

Although codon usage is clearly determined by the joint actions of mutation, drift, and selection (Bulmer 1991; Hershberg and Petrov 2008), the fitness benefit of CUB is less clear. There are two prevailing, non-mutually exclusive, hypotheses on the selective utility of CUB: accuracy and efficiency of protein translation (Hershberg and Petrov 2008). The translational accuracy hypothesis asserts that different synonymous codons have different probabilities of mistranslation, and that the use of accurately translated codons is beneficial because mistranslation reduces the number of functional molecules, wastes energy, and/or induces cytotoxic protein misfolding. Unequivocal evidence for this hypothesis exists (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Zhou et al. 2009).

By contrast, the translational efficiency hypothesis lacks direct evidence. This hypothesis holds that different synonymous codons are translated at different speeds, and that faster translation is beneficial because it minimizes ribosome sequestering and so helps alleviate ribosome shortage (Bulmer 1991; Akashi 2001; Kudla et al. 2009). The relevance of ribosome shortage is evident from the findings that most ribosomes are actively engaged in translation during rapid cell growth (Forchhammer and Lindahl 1971; Boehlke and Friesen 1975) and that ribosome concentration increases with the rate of cell growth (Warner 1999). An important observation invoked to support the efficiency hypothesis is that cognate tRNAs of preferred codons tend to have higher cellular concentrations (or more gene copies) than those of unpreferred codons (Ikemura 1981; Ikemura 1982), which may allow faster translation of preferred codons than unpreferred codons. While results from several earlier studies are consistent with this hypothesis (Carlini and Stephan 2003; Kudla et al. 2009), these studies do not exclude the possibility that the observed differences in activity or fitness caused by synonymous mutations are entirely due to CUB's influence on translational accuracy (see4.8.2). Here we directly test the efficiency hypothesis and its presumed underlying mechanism.

4.3 Estimating *in vivo* translational speeds

The translational efficiency hypothesis assumes that synonymous codons have different translational speeds, caused by disparities in codon selection time (*CST*), the time needed for ribosomal A site to find the cognate ternary complex of aminoacylated tRNA + eEF-1α + GTP. To test this proposition, we took advantage of a genome-wide ribosome profiling study of *Saccharomyces cerevisiae* that surveyed ribosome-protected mRNA fragments at a nucleotide resolution in a cell population at a given moment by Illumina deep sequencing (Ingolia et al. 2009). Because the probability that a codon is docked at the A site is proportional to its *CST*, we estimated the relative *CST*s of all 61 sense codons (Figure 4.1A) by the ratio of the observed codon frequencies at the A site in the ribosome profiling data and the expected codon frequencies estimated from mRNA-Seq data generated under the same condition in the same experiment. The standard errors of the *CST* estimates, measured by bootstrapping genes from the original datasets, are on average 12% of the *CST* estimates (Figure 4.1A), indicating that our *CST* estimates are overall quite precise.

CUB is commonly measured by the *relative synonymous codon usage* (*RSCU*), defined by the frequency of a codon relative to the average frequency of all of its synonymous codons in a set of highly expressed genes (Sharp et al. 1986). To compare the usage of all 61 sense codons, we also use *RSCU'*, which is the proportion of use of a given codon among synonymous choices in a set of highly expressed genes. Another commonly used measure of CUB is the *codon adaptation index* (*CAI*) (Sharp and Li 1987), which is calculated for a gene, and measures its usage of high-*RSCU* codons. The greater the *CAI*, the more prevalent are preferred codons in the gene.

Contrary to the widely held presumption that preferred codons are translated faster than unpreferred codons, no significant negative correlation between *RSCU'* and *CST* was observed among the 61 sense codons (Figure 4.1B). It is also believed that codons with abundant cognate tRNAs tend to have low *CST*s. Because tRNA gene copy number and tRNA concentration are highly positively correlated (Percudani et al. 1997; Tuller et al. 2010a), the former is often used as a proxy of the latter. However, neither tRNA gene copy number (Figure 4.1C) nor tRNA concentration (Figure 4.1D) correlates negatively with *CST*. Because codons and tRNAs do not have one-to-one

correspondence, in the foregoing analysis, we considered the best-matching tRNA species for each codon. This codon-tRNA relationship has been shown to be more accurate than the wobble rule, at least in yeast (Percudani et al. 1997).

We also examined each amino acid separately. Among the 18 amino acids with at least two codons, 12 (Ala, Asn, Cys, Gln, Glu, Gly, Ile, Lys, Ser, Thr, Tyr, and Val) showed a negative correlation between *RSCU'* and *CST*, while 6 (Arg, Asp, His, Leu Phe, and Pro) showed a positive correlation, when statistical significance of the correlation was not required (Figure 4.1A). The number of negative correlations is not significantly more than the chance expectation of 9 ($P = 0.12$, one-tail sign test).

Using the standard errors of the *CST* estimates for the foregoing 18 amino acids (Figure 4.1A), we tested whether the *CSTs* are significantly different between the synonymous codon with the highest *RSCU'* and that with the lowest *RSCU'*. After the control for multiple testing by the Bonferroni correction, only two amino acids showed significant differences. The highest-*RSCU'* codon has a lower *CST* than the lowest-*RSCU'* codon for glycine (nominal $P = 0.002$), while the opposite is true for arginine (nominal $P < 0.001$). Our results are robust to different multiple-testing corrections, as no other amino acids show a nominal $P < 0.01$. Furthermore, when *RSCU'* is not considered, arginine is the only amino acid for which synonymous codons show significant heterogeneity in *CST* at the 5% significance level after the correction for multiple testing. Following an earlier study (Hershberg and Petrov 2009), we also tried defining preferred codons without using gene expression data, but the results are not different. The overall lack of a significant negative correlation between *CST* and synonymous codon usage is real rather than an artifact of imprecise *CST* estimation, because the standard errors of *CSTs* are quite small (Figure 4.1A) and *CST*s of several nonsynonymous codons differ significantly from one another (see below).

To validate the above findings, we also directly compared *RSCU'* values of individual codon positions of Illumina reads from the ribosome profiling data, without estimating *CST*s. If unpreferred codons are translated more slowly and therefore stay at the ribosomal A site longer than preferred codons, codons at the A site should have a lower *RSCU'* on average than its neighboring sites of the same read, after the correction of sequencing bias by mRNA-Seq data. However, we observed no dip in *RSCU'* at the A

site (Figure 4.1E). We further calculated, within each gene, the ratio between the frequency of preferred codons and that of unpreferred codons at the ribosome A site of Illumina reads from the ribosome profiling data, after correction by mRNA-Seq. This ratio is expected to be 1 if preferred and unpreferred codons are translated equally fast. Indeed, after combining the ratio for all amino acids and all genes using the Mantel-Haenszel procedure (Sokal and Rohlf 1995), we found the overall ratio to be 0.984, not significantly different from 1 ($P = 0.21$, two-tail χ^2 test).

4.4 Optimal codon usage under tRNA shortage

The above findings are puzzling, because the first step in the interaction between tRNA and mRNA is non-specific (Ogle and Ramakrishnan 2005) and the relative waiting time for the cognate tRNA to arrive at the ribosome A site is expected to be inversely proportional to the relative concentration of the cognate tRNA. It was also reported that *CST* is the rate-limiting step in translational elongation (Varenne et al. 1984). The only plausible explanation of similar *CST*s among synonymous codons is that, in wild-type yeast cells for which the ribosome profiling was conducted, available cognate tRNAs for translating synonymous codons have effectively the same concentration.

In rapidly growing yeast, ~80% of total RNA is rRNA and ~15% is tRNA(Warner 1999). The mean length of yeast tRNAs is ~72 nucleotides and the total length of rRNAs per ribosome is 5469 nucleotides (Warner 1999). Thus, the number of tRNA molecules per cell is approximately (15%/72)/(80%/5469) = 14.2 times the number of ribosomes per cell, substantially exceeding the expected ratio of two tRNAs per active ribosome (at A and P sites, respectively) if tRNA recharging and diffusion is instantaneous.

In reality, however, tRNA recycling takes time and thus cannot be ignored. Each tRNA, after completing its job of transferring an amino acid to the elongating peptide and then exiting the ribosomal E site, needs to be recharged with the cognate amino acid and then with eEF-1α + GTP to form a ternary complex before it can be reused in translation. It has been estimated that each ribosome translates ~32.6 codons per second in yeast (von der Haar 2008). This implies that on average a tRNA molecule needs to be used 32.6/14.2 = 2.3 times per second, or once every 0.44 second. It is possible that the time for ternary complexes to form and diffuse to ribosomal A site is a substantial fraction of

0.44 second, so that the local concentration of ternary complexes is much lower than the total tRNA concentration. A recent study reported that consecutive synonymous codons in an mRNA tend to use the same tRNA and proposed that this codon choice is beneficial because a tRNA does not diffuse far from the ribosome after exiting its E site and is reused for translating the next synonymous codon when the ternary complex is formed again (Cannarozzi et al. 2010). This observation and its explanation strongly implies that the local concentration of ternary complexes is low; otherwise, the addition of one cognate tRNA molecule among on average 20 tRNAs (because identical amino acids are expected to be on average 20 residues apart) cannot significantly increase the relative concentration of the cognate tRNA around the ribosome. Based on available information in *E. coli*, we calculated that the physiological concentration of ternary complexes is only ~4.3% of the total concentration of tRNAs and ~22% of the concentration of ribosomes. These observations strongly support our hypothesis that available tRNA is in shortage during translation. Consistent with our hypothesis, total tRNA concentrations increase with the rate of cell growth in *E. coli* (Dong et al. 1996) and tRNA gene copy number increases with the shortening of the minimal generation time across species (Rocha 2004).

Under tRNA shortage, the optimal usage of synonymous codons in minimizing the total *CST* (i.e., maximizing translational efficiency) is to use isoaccepting tRNAs in proportion to their concentrations. That is, $p_i = q_i$, where p_i is the relative usage of the *i*th synonymous codon of an amino acid ($\Sigma p_i = 1$) and q_i is the relative concentration of the corresponding tRNA ($\Sigma q_i = 1$). Under this codon usage, available cognate tRNAs of synonymous codons have equal concentrations and synonymous codon selection times become identical. We will refer to this theoretical optimal codon usage under tRNA shortage as the proportional rule. The proportional rule is not predicted by other models. For example, without tRNA shortage, two optimal solutions in minimizing the total *CST* exist. When codon usage is fixed, isoaccepting tRNA concentrations should follow $q_i^2 / q_j^2 = p_i / p_j$, which is referred to as the square rule (Liljenstrom et al. 1985; Bulmer 1987). When tRNA concentrations are fixed, only the codon corresponding to the most abundant tRNA species should be used (Bulmer 1987), which is referred to as the truncation rule.

To test if the actual codon usage of yeast follows the proportional rule, we examined the 12 amino acids that are each translated by at least two tRNA species in yeast. For each amino acid, the relative transcriptomic usage of a codon among synonymous codons (i.e., p_i) is quite close to the relative gene copy number of its cognate tRNA among isoaccepting tRNAs (i.e., q_i), as predicted by the proportional rule (Figure 4.2A). We measured the Euclidian (Figure 4.2B) and Manhattan (Figure 4.2C) distances in synonymous codon usage from the observed values to those predicted by the proportional rule, and found these distances significantly shorter than expected by chance (Figure 4.2 B-D). Not surprisingly, genomic codon usage fits the proportional rule less well than the transcriptomic codon usage (Figure 4.2A), reflected by greater distances from the predicted values (Figure 4.2 B & C).

The better fitting of the transcriptomic codon usage to the proportional rule than to the square rule and truncation rule can be seen from a comparison of the distances under these three models (Figure 4.2D). We also compared the likelihood of the three models, given the observed codon usage (Figure 4.2D). The proportional model has a much higher $\log_{10}$(likelihood) than the square model. Because the likelihood of the truncation model is 0, this model is much worse than the other two models. The same conclusions are reached for the transcriptomic codon usage of all other model eukaryotes we examined (Figure 4.2 A & D).

In the above analysis, we combined synonymous codons that are recognized by the same tRNA species (referred to as iso-synonymous codons). Because the relative usage of such iso-synonymous codons does not affect the relative usage of isoaccepting tRNAs, it presumably does not affect translational efficiency. Nonetheless, iso-synonymous codons are not used equally, and factors other than translational efficiency (e.g., translational accuracy) may be at work.

4.5 Codon-tRNA imbalance reduces translational efficiency

The observation of similar *CST*s among synonymous codons and the empirical validation of the proportional rule strongly support the following model that includes three elements: (1) available tRNAs are in shortage during translation, (2) translational efficiency is optimized in nature by balanced codon usage according to tRNA

concentrations, and (3) synonymous codons are translated with similar speeds under the codon-tRNA balance. Our model predicts reduced translational efficiency due to ribosome sequestering when the codon-tRNA balance is broken. It further predicts lower efficiency under exclusive use of preferred codons than balanced use of preferred and unpreferred codons.

We experimentally tested the above predictions by quantifying the cellular efficiency in translation, represented by the protein expression of a reporter gene, under different levels of codon-tRNA imbalance induced by the expression of another gene. Unlike previous studies (Carlini and Stephan 2003; Kudla et al. 2009), our separation of the inducer and reporter allows the distinction among several potential mechanisms of CUB's impact on protein expression. We inserted our reporter gene, the *Venus* yellow fluorescent protein (vYFP) gene controlled by the *GPD* promoter, into Chromosome XII of a haploid strain of *S. cerevisiae* (Figure 4.3A). We then designed four synonymous sequences encoding another fluorescent protein, mCherry, as our inducer. The four *mCherry* sequences, named *mCherry*-1, 2, 3, and 4, cover the entire range of *CAI* of native yeast genes (Figure 4.3B). We developed an index, distance to native codon usage (D_{ncu}), to measure the difference between the codon usage of a (heterologous) gene and the overall codon usage of the host cell, which is proportional to tRNA concentrations (see Materials and Methods). The four *mCherry* versions also span a large range of D_{ncu} (Figure 4.3C) and show different degrees of codon-tRNA imbalance for individual amino acids. Other than synonymous codon usage, the four *mCherry* versions are nearly identical: they encode the same protein sequence, have similar G+C content (42-44%), and have identical sequences in the first 56 nucleotides of the coding region, because this region may affect the level of protein expression (Kudla et al. 2009; Gu et al. 2010; Tuller et al. 2010b). Each *mCherry* gene is expressed from a constitutive and strong promoter on a high-copy-number plasmid. The four plasmids were separately transformed to yeast cells carrying the vYFP reporter gene (Figure 4.3A). Our model predicts that the higher the D_{ncu} of *mCherry*, the lower the vYFP expression.

The four yeast strains were grown in rich media to the log phase, and the expression levels of vYFP and mCherry proteins were inferred from their fluorescent signals, which were simultaneously measured for each cell by fluorescence-activated cell

scanning of at least 300,000 cells. We found mCherry expression levels to be significantly different among the four strains. Within each strain, expression levels of mCherry and vYFP are negatively correlated among cells. Hence, the expressions of vYFP cannot be directly compared among strains. Instead, we separated the cells of each strain into three bins on the basis of mCherry expression and then compared vYFP expressions among the four strains for cells with similar mCherry expressions (Figure 4.3D). We found that, across the range of mCherry expressions shared by the four strains, the higher the D_{ncu} of *mCherry*, the lower the expression of vYFP (Figure 4.3D). Furthermore, the vYFP expression-level difference among the strains increases with the mCherry expression level (Figure 4.3D). Of special interest is the comparison between *mCherry*-3 and *mCherry*-4, which clearly shows that it is a low D_{ncu} rather than a high *CAI* that enhances translational efficiency (Figure 4.3D). A multivariate regression analysis of all cells from the four strains further demonstrated that D_{ncu} is significantly more important than *CAI* in explaining the variation of the vYFP signal ($P < 0.001$).

The above results were not due to different random mutations fixed in the genomes of the four strains during our experiments, because the vYFP signals were not significantly different among the strains upon removal of the plasmids (Figure 4.3E). We also sequenced the entire plasmid DNA from each strain and found no mutation. Using quantitative polymerase chain reaction, we further verified that the *vYFP* mRNA abundance is not different among the four strains (Figure 4.3F). Thus, the among-strain variation in vYFP signal must be due to a variation in translation. We also confirmed our results by a finer control of mCherry expression and ruled out the possibility that our observation is a byproduct of potential differences in translational accuracy among different *mCherry* versions. Furthermore, because the accuracy hypothesis is based on *CAI* and thus predicts a higher vYFP expression in the strain carrying *mCherry*-4 than that carrying *mCherry*-3, our results (Figure 4.3D) are inexplicable by this hypothesis. Similarly, mechanisms resulting from translational errors, such as protein misfolding or aggregation, cannot explain our observation either.

In the experiment, we used *vYFP* to represent native genes in the yeast genome. However, because vYFP and mCherry have 71/220 = 32% of protein sequence identity, one might ask whether our observation can be generalized. Specifically, could the

negative influence of mCherry expression on vYFP expression be caused entirely by the similarity in codon usage between *mCherry* and *vYFP*? We measured the codon usage dissimilarity between a pair of genes by a Euclidian distance and examined the distribution of this distance between each *mCherry* version and all yeast genes. The distribution is approximately bell shaped and the distance between *mCherry* and *vYFP* falls in the central part of the bell, suggesting that *mCherry* is no more similar to *vYFP* in overall codon usage than to average yeast genes. Furthermore, our results cannot be explained by amino acid similarity between mCherry and vYFP, because all *mCherry* versions have the same amino acid sequence and should not differentially affect vYFP expression through amino acid usage. Thus, our observation from *vYFP* can be extrapolated to native genes in the yeast genome.

4.6 Why more highly expressed genes have stronger CUB

If translational efficiency is maximized when the cellular codon usage follows the proportional rule, why do highly expressed genes necessarily prefer codons with highly abundant cognate tRNAs and have stronger CUB than lowly expressed genes? We hypothesize that these phenomena are due to differential selective coefficients associated with synonymous mutations occurring in highly expressed and lowly expressed genes in the regain of the codon-tRNA balance upon a genetic perturbation. Let us imagine an amino acid with two synonymous codons (codon1 and codon2) that each uses a distinct tRNA species (tRNA1 and tRNA2) and assume that the present codon usage follows the proportional rule. Now, if the proportion of tRNA1 rises due to a mutation, natural selection will promote the fixations of synonymous mutations from codon2 to codon1 to reestablish the codon-tRNA balance. Such advantageous mutations occurring in highly expressed genes affect tRNA usage more than those occurring in lowly expressed genes and hence have a greater selective advantage and are fixed faster. This difference becomes even bigger when clonal interference (Gerrish and Lenski 1998) is considered. As a result, highly expressed genes use more codon1 and fewer codon2 than before and show stronger CUB. The contrasting scenario, in which the tRNA usage is rebalanced by frequent use of codon1 in lowly expressed genes, requires many synonymous substitutions in many lowly expressed genes, which will not happen because it takes

much longer than rebalancing the tRNA usage by increasing codon1 frequency in highly expressed genes. Indeed, in a computer simulation of codon usage evolution that starts from the equal usage of 4 synonymous codons whose cognate tRNAs have different concentrations, the final usage of the codons, after 500 generations of random mutation, genetic drift, and natural selection for translational efficiency, follows the proportional rule (Figure 4.4A). More importantly, the preferential use of high-concentration tRNA species and strong CUB in highly expressed genes are seen from both the average of 1000 simulation replications (Figure 4.4B) and any one replication (Figure 4.4**C**). The standard deviations presented in Figure 4.4B indicate an extremely low probability for CUB to be stronger or a preferred codon to be used more frequently in lowly expressed genes than highly expressed genes. As expected, the phenomena in Figure 4.4 disappear when the natural selection for translational efficiency is removed in the simulation. These observations support our model that the high *CAI* of highly expressed genes is a byproduct of natural selection for an overall cellular efficiency in translation, rather than the direct product of stronger selection for translation efficiency in more highly expressed genes (Hershberg and Petrov 2008).

4.7 Optimal amino acid usage under tRNA shortage

Analogous to synonymous codon usage, we predict that the optimal amino acid (or nonsynonymous codon) usage in speeding up translation is in proportion to the corresponding tRNA concentrations. Indeed, amino acid frequencies inferred from transcriptome data were reported to correlate positively with the corresponding tRNA gene copy numbers in yeast (Akashi 2003) and *C. elegans* (Duret 2000). More importantly, actual amino acid usage is significantly closer than random usage to our predicted optimal (i.e., the diagonal line in Figure 4.5**A;** $P < 10^{-6}$, simulation test). This phenomenon is also true in all other model eukaryotes examined, although the level of match between the observation and prediction varies among species (Figure 4.5A). Transcriptomic amino acid usages instead of proteomic amino acid usages are plotted here because the latter are unavailable for most species. Nevertheless, *S. cerevisiae* data showed an almost perfect correlation between transcriptomic and proteomic amino acid usages, indicating that the former is a good proxy for the latter. We also predict a

positive correlation between aminoacyl tRNA synthetase concentration and corresponding tRNA concentration to enhance the efficiency of amino acid charging. Such a correlation is indeed found in *S. cerevisiae* ($r = 0.45$, $P = 0.03$).

If amino acid frequencies are in perfect proportion to tRNA concentrations, the mean *CST* for an amino acid should not vary among amino acids. This uniformity, however, is not observed in yeast, suggesting that amino acid usage is only roughly proportional to tRNA concentrations (Figure 4.5A), which may be due to mutational bias (Gu et al. 1998) or antagonistic selective pressures from factors such as physiochemical properties (Zhang 2000) and synthetic costs (Akashi and Gojobori 2002) of various amino acids. Our model predicts that the average *CST* of an amino acid increases with the decrease of the relative availability of tRNAs for the amino acid. Indeed, a negative correlation exists between the tRNA availability and *CST* for the 20 amino acids (Pearson's $r = -0.40$, $P = 0.03$, permutation test; Figure 4.5B). This finding reconfirms tRNA shortage in translation, explains in part why *CST*s of nonsynonymous codons vary, and indicates compromised translational efficiency due to other fitness effects of amino acid usage.

4.8 Discussion

4.8.1 The translational efficiency hypothesis of CUB

Results from several earlier experiments are consistent with the role of CUB in enhancing translational efficiency or reducing ribosome sequestering (Carlini and Stephan 2003; Kudla et al. 2009). For example, when expressing many synonymous versions of a green fluorescent protein (GFP) gene in *E. coli*, Kudla and colleagues reported that strains harboring high-*CAI GFP* genes tend to grow faster than those harboring low-*CAI GFP* genes, despite the lack of a correlation between the *GFP* protein expression level and its *CAI* (Kudla et al. 2009). Although these authors found no correlation between *CAI* and protein misfolding, their experiment was unlikely to be sensitive enough for quantifying GFP misfolding (Kudla et al. 2009). Thus, it could not rule out the possibility that the observed variation in fitness was entirely caused by CUB's influence on translational accuracy. By contrast, we were able to demonstrate

CUB's impact on translational efficiency after excluding its impact on translational accuracy.

A recent study in *E. coli* showed that the ribosome shortage induced by over-expression of unneeded proteins can be alleviated by physiological adaptation in 30 to 40 generations, owing to the manufacture of additional ribosomes (Shachrai et al. 2010). This finding suggests that the disadvantage of suboptimal codon usage may also be mitigated by physiological adaptation. Nevertheless, physiological adaptation takes time. If the growth rate fluctuates rapidly due to frequent environmental changes, the fitness of the individual with suboptimal codon usage is expected to be much lower than the individual with balanced codon usage.

We hypothesized and demonstrated that translational efficiency is optimized by codon-tRNA balance. This new model of translational efficiency by unequal codon usage differs substantially from the prevailing model (Table 4.1). One critical piece of evidence for our model is similar *CST*s of synonymous codons in wild-type yeast. Our *CST* estimation is based on the assumption that the time a codon occupies the ribosomal A site equals the waiting time for the cognate tRNA. Our estimates of all *CST*s would be biased upward to a similar level if downstream "traffic jams" happen during translational elongation. However, a recent study suggested that downstream traffic jams are unlikely, due to slow "ramps" at the beginning of an mRNA (Tuller et al. 2010a). Furthermore, even if downstream traffic jams occur, it should affect synonymous codons as well as nonsynonymous codons and thus cannot explain why only synonymous codons but not nonsynonymous codons have similar *CST*s.

Over two decades ago, Curran and Yarus indirectly estimated relative *CST*s for 29 sense codons in *E. coli*, under the assumption that the probability of a frame shift in the translation of a codon is proportional to the *CST* of the codon (Curran and Yarus 1989). They reported that only codons of very low *CST*s tend to be preferentially used (Curran and Yarus 1989). However, because their fundamental assumption about the frame-shift rate is incorrect (Vimaladithan and Farabaugh 1994), their *CST* estimates are unlikely to be correct. It is also possible that prokaryotes and eukaryotes have some differences in using CUB to regulate translational efficiency (e.g., translational attenuation in prokaryotes). In another *E. coli* study, Sorensen and colleagues reported faster

translation of a multicopy-plasmid-borne *lacZ* gene when a segment of the gene comprises mainly preferred codons than when it comprises mainly unpreferred codons (Sorensen et al. 1989). This result cannot be used to infer relative *CST*s of synonymous codons in wild-type cells, because the extremely high expression of synonymous versions of the endogenous *lacZ* gene from plasmids potentially breaks the codon-tRNA balance and alters *CST*s. Nevertheless, their observation is fully compatible with our finding of different levels of translational efficiency induced by the expressions of different synonymous versions of *mCherry*. Several other studies reported similar findings (Robinson et al. 1984; Varenne et al. 1984). Recently, some authors calculated *CSTs* by assuming that the *CST* of a codon is determined by the relative concentrations of its cognate, nearly cognate, and non-cognate tRNAs without considering tRNA shortage or using ribosome profiling data (Siwiak and Zielenkiewicz 2010). Because of the violation of the fundamental assumption they made, their estimates are likely to be incorrect. Indeed, their estimated *CST*s would predict a slower translation of mCherry version 3 than 4, contradictory to our experimental result (Figure 4.3D). While the present work was under review, Ingolia and colleagues reported estimates of translational elongation speeds in mouse embryonic stem cells using a pulse-chase strategy that does not involve expressions of heterologous genes (Ingolia et al. 2011). Although their method is different from ours, their finding of similar elongation speeds among synonymous codons is highly consistent with our results from yeast.

Our discoveries require reinterpretation of several earlier observations. For example, higher prevalence of codons with abundant cognate tRNAs in genes with higher expressions is often interpreted as a result of a stronger demand for fast translation of more abundant proteins (Sharp et al. 1986; Sharp and Li 1987). This interpretation is not supported by our results. Rather, we suggested and demonstrated by simulation that, the selection coefficient for synonymous mutations that help achieve the codon-tRNA balance is greater in highly expressed genes than in lowly expressed genes, leading to quicker and more acquisitions of codons with abundant cognate tRNAs in the former than in the latter. In this regard, our results support that CUB serves as a global strategy to enhance the efficiency of the translation system (Andersson and Kurland 1990; Kudla et al. 2009).

Within an organism, the transcriptome can vary among cell cycle stages, developmental stages, and tissues. How do such variations affect the codon-tRNA balance? We found pairwise Pearson's correlations in transcriptomic usage of all 61 sense codons to be nearly 1 among different time points in the *S. cerevisiae* mitotic cell cycle (Figure 4.6). We further analyzed the transcriptomic usage of all 61 codons across tissues and/or developmental stages in the worm, fruit fly, and human. If multiple replications of the same cell type exist in a dataset, we randomly chose one replication in our analysis. Similarly high correlations were observed among different cell types within species (Figure 4.6). By contrast, the correlation is generally below 0.5 between any pair of the four species examined here. The high correlation in codon usage across cell cycle stages, developmental stages, and tissues of the same species is likely due to house-keeping genes, which are always highly expressed. Thus, within-organism gene expression variations have little impact on the maintenance of the codon-tRNA balance. Further, tRNA concentrations may covary with the transcriptomic codon usage to maintain the codon-tRNA balance across tissues (Dittmar et al. 2006).

A byproduct of our *CST* estimation is the translational initiation rate of each gene. We found that the translational initiation rate is significantly positively correlated with the mRNA concentration ($\rho = 0.34$, $P = 6 \times 10^{-81}$), suggesting a coordinated regulation of gene expression at the transcriptional and translational levels. We also observed a strong positive correlation between the translational initiation rate and *CAI* ($\rho = 0.51$, $P < 10^{-196}$), suggesting that *CAI* provides a moderate amount of information about the translational initiation rate. This may explain why the protein concentration correlates with the product of mRNA concentration and *CAI* better than with the mRNA concentration alone (Brockmann et al. 2007). Several studies revealed reduced mRNA stability near the translation initiation site, suggesting that the reduced stability may enhance the translational initiation rate (Kudla et al. 2009; Gu et al. 2010; Tuller et al. 2010b). Indeed, we found a weak but significant positive correlation between the reduction in mRNA stability (Gu et al. 2010) and our estimated translational initiation rate ($\rho = 0.08$, $P = 1 \times 10^{-5}$).

4.8.2 Translational efficiency and accuracy are two separable benefits of CUB

Given that CUB improves both translational efficiency and accuracy, one wonders whether one of these effects is a side-effect of the other. For instance, it was previously suggested that the variation in translational accuracy among synonymous codons may be a byproduct of the variation in translational efficiency, because (i) most translational errors are believed to occur during codon selection, (ii) codon selection has been assumed to be faster for preferred codons than unpreferred codons, and (iii) faster codon selection is thought to result in fewer errors (Powell and Moriyama 1997). Because our result invalidates assumption (ii) for wild-type cells, the above argument no longer holds. Thus, even though translational accuracy may be affected by relative concentrations of tRNAs in engineered yeast cells with grossly imbalanced codon-tRNA usage (Kramer et al. 2010), this impact is not expected in wild-type cells because our results strongly suggest that isoaccepting tRNA species have effectively the same concentrations in wild-type cells. In addition, the enrichment of preferred codons at evolutionarily conserved amino acid residues cannot be explained by the translational efficiency hypothesis (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Zhou et al. 2009). Furthermore, experimental data showed that translational accuracies of iso-synonymous codons vary (Precup and Parker 1987), suggesting that the variation in accuracy cannot be entirely caused by the variation in cognate tRNA concentration, because iso-synonymous codons use the same cognate tRNA. Rather, comparative genomic analyses strongly suggest that translational accuracy is likely to be intrinsically different among synonymous codons (Rocha and Danchin 2004; Hershberg and Petrov 2009). Further, we were able to establish CUB's impact on translational efficiency even after we controlled its impact on translational accuracy (Figure 4.3). In addition, because translational accuracy is not entirely determined by translational efficiency (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Zhou et al. 2009), the proportional rule, which is predicted from selection for efficiency, is not predicted from selection for accuracy, especially because translational errors at different residues have different fitness effects. Thus, the impact on efficiency cannot be a byproduct of the impact on accuracy. Taken together, we conclude that translational accuracy and efficiency are two separable benefits of CUB.

4.8.3 Evolutionary models of codon usage bias

Let us compare three evolutionary models of CUB that differ in the roles of translational accuracy and efficiency as the selecting agent. We also consider mutational bias and genetic drift, two known factors in the evolution of CUB, in these models. In model I, translational efficiency is the sole selecting force (Figure 4.7). This model predicts co-evolution of codon usage and cognate tRNA concentrations and a codon-tRNA balance at which the relative frequency of a synonymous codon (p_i) equals the relative abundance of its cognate tRNA (q_i). The expected values of $p_i = q_i$ are determined by the mutational bias, which directly affects codon usage and indirectly affects tRNA concentrations. However, this model cannot explain the observation that, although preferred codons of an amino acid vary among species, this variation decreases substantially (but does not disappear) after the control of genomic GC content (Hershberg and Petrov 2009). For example, GTT and GTA both code for valine and have the same GC content, but GTT is frequently used as the preferred codon when the genomic intergenic GC content is below 50% (Hershberg and Petrov 2009). When the GC content exceeds 50%, GTG rather than GTC is often used as the preferred codon for valine (Hershberg and Petrov 2009). This observation suggests that, in addition to translational efficiency, there is a separate selecting force with a relatively constant direction.

In model II, translational accuracy is the sole selecting agent on CUB (Figure 4.7). The demand for translational accuracy, coupled with the mutational bias, determines the expected CUB, whereas selection for translational efficiency determines tRNA concentrations based on codon frequencies. The phenomenon of stronger CUB in more highly expressed genes is explainable by the protein-misfolding-avoidance hypothesis which predicts that highly expressed genes are translated more accurately by using accurate codons more frequently (Drummond and Wilke 2008; Yang et al. 2010). Model II predicts that, after the control for the mutational bias, accurate codons are always the preferred codons in a species. If the translational accuracy of a codon is an intrinsic property of the codon and does not vary among species (Rocha 2004), we should observe no variation in the choice of preferred codons, after the control of mutational bias. This prediction, however, is incorrect, because preferred codons are not always the same in

different species with the same mutational bias (Rocha 2004; Hershberg and Petrov 2009). A more rigorous test of this model is to compare the accurate and preferred codons of each amino acid in a species, because model II predicts a complete match between them. For each codon, we calculated an odds ratio by the relative use of the codon over other synonymous codons at conserved amino acid positions divided by that at non-conserved amino acid positions; the synonymous codon with the highest odds ratio is regarded as the most accurate codon because it is most preferentially used at important amino acid positions (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Zhou et al. 2009). By comparing *S. cerevisiae* with its relative *S. bayanus*, we identified conserved and non-conserved amino acid positions. We calculated the odds ratio for each codon in each gene and then combined the odds ratios from all genes using the Mantel-Haenszel procedure (Sokal and Rohlf 1995). By definition, the preferred codon of an amino acid is the one with the highest *RSCU'*. We found that, in 6 (Ala, Asp, Gly, His, Thr, and Val) of the 18 amino acids that have at least two synonymous codons, the codon with the highest odds ratio is different from the codon with the highest *RSCU'* (Figure 4.8). Furthermore, for three amino acids (Asp, His, and Thr), the codon with the highest *RSCU'* has an odds ratio significantly lower than 1 (Figure 4.8). We also used the 10% most highly expressed genes to calculate odd ratios; 8 (Ala, Arg, Asp, Cys, Ile, Leu, Thr, and Val) of the 18 amino acids show mismatches between the codon with the highest *RSCU'* and the codon with the highest odds ratio (Figure 4.8). These results provide unambiguous evidence for the inadequacy of model II.

In model III, selections for translational accuracy and efficiency jointly determine CUB (Figure 4.7). Let us consider three types of synonymous mutations with regard to their impacts on translational accuracy and efficiency. First, a synonymous mutation is likely to be fixed when it enhances both translational accuracy and efficiency, but is likely to be lost when it decreases both. Second, a synonymous mutation may increase the accuracy but reduce the efficiency. One possible outcome is that selection for higher accuracy will gradually alter the codon usage, which is followed by tRNA concentration changes that recover the loss of efficiency. Eventually, accurate codons will be the preferred codons. Alternatively, selection for higher accuracy may not be able to alter the codon usage permanently if the loss of efficiency is either too large or cannot be

recovered by a corresponding tRNA change as quickly as the switch back of the codon usage. Consequently, accurate codons cannot become the preferred codons and the system is trapped in a local fitness peak that is the maximum for efficiency but not accuracy. For example, while codon CCA is more accurate than CCT for proline (Figure 4.8), there are still about a quarter of bacterial species with GC% < 40 that use CCT as their preferred proline codon (Hershberg and Petrov 2009), suggesting that it is not rare for codon usage to be trapped in a local fitness peak. Third, a synonymous mutation may increase the efficiency but reduce the accuracy when the system is at a codon-tRNA imbalance. Although the fate of this mutation is determined by the relative strengths of the two forces, changes of tRNA concentrations could resolve the conflict better because they can increase efficiency without reducing accuracy. So, the final codon usage pattern will also depend on the rate of mutations that alter tRNA concentrations. While the quantitative aspects of model III require further exploration, it is clear that the model is able to explain, at least qualitatively, both the matches and mismatches between the accurate and preferred codons (Figure 4.8). It is also able to explain the codon-tRNA balance and the phenomenon of stronger CUB in genes with higher expressions. Thus, model III is most compatible with and best supported by available data. In addition to translational accuracy and efficiency, synonymous codon usage of individual genes may also be shaped by other forces, for example, those related to RNA splicing and stability (Chamary et al. 2006). But these forces are gene-specific and do not create genomic patterns of CUB.

4.8.4 Implications for synthetic biology

Synthetic biology designs and constructs novel biological functions not found in nature. It has long been known that, in many but not all cases, increasing the *CAI* of a transgene boosts its protein expression (Gustafsson et al. 2004; Kudla et al. 2009; Welch et al. 2009). Different protein expression levels of synonymous transgenes are likely caused by *CST* differences created by various degrees of codon-tRNA imbalance induced by transgene expressions. Consistent with this idea, overexpression of rare tRNAs of *E. coli* (the bio-reactor) can rescue the tRNA depletion when heterologous human genes are expressed in *E. coli* (Gustafsson et al. 2004). When an artificially designed gene is added

to a host cell, the potential imbalance between the overall cellular codon usage and the tRNA pool also affects the expressions of native genes and hence the growth of the host cell. We showed that D_{ncu}, a newly devised index measuring the distance in codon usage between the transgene and the host cell, is an accurate indicator of the impact of per transgene protein molecule production on the expressions of native genes. We demonstrated that it is the D_{ncu} rather than *CAI* of the transgene that predicts its impact on the host protein expression. Therefore, D_{ncu} should be considered in synthetic biology when the impact of transgene expression on host gene expressions is a concern. Further, when genes from multiple species are assembled into a synthetic genome, designing tRNA gene numbers in proportion to the usage of their cognate codons will likely make protein expressions in the entire cell most efficient.

4.9 Materials and Methods

4.9.1 Genomic data

The yeast ribosome profiling data (Ingolia et al. 2009) were downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE13750. Gene expression and protein expression levels were from http://web.wi.mit.edu/young/expression/ (Holstege et al. 1998), http://www.imb-jena.de/tsb/yeast_proteome/ (Beyer et al. 2004), and the supplementary data of a previous study (Ghaemmaghami et al. 2003). Transcriptomic data for the yeast mitotic cell cycle were from a previous study (Cho et al. 1998). Gene sequences and reading frames were downloaded from *Saccharomyces* Genome Database (SGD, www.yeastgenome.org). Numbers of tRNA gene copies were retrieved from an earlier study (Percudani et al. 1997).

Gene expression levels in *A. thaliana*, *D. melanogaster*, *M. musculus*, and *H. sapiens* were downloaded from Gene Expression Omnibus (GDS416, GDS2784, GDS592 and GDS596, respectively). Gene expression levels in *S. pombe* and *C. elegans* were retrieved from two earlier studies (Wilhelm et al. 2008; Hillier et al. 2009), respectively. Peptide and cDNA sequences of *S. pombe*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens* were from Ensembl (www.ensembl.org/).

Numbers of tRNA gene copies in the above species were obtained from the genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb/).

4.9.2 Estimation of codon selection time (*CST*)

Using the *S. cerevisiae* ribosome profiling data (Ingolia et al. 2009), we identified codons docked at the ribosomal A site, from the Illumina Genome Analyzer sequencing reads. By comparing the observed codon frequencies in the ribosome profiling data with the expected codon frequencies estimated from mRNA-Seq data generated under the same condition in the same experiment, we calculated the relative *CST*s of all 61 sense codons. Although Illumina sequencing may be biased toward certain sequences or nucleotides (Dohm et al. 2008), this bias affects the mRNA-Seq and ribosome profiling data equally and thus will not affect our estimation of *CST*. For a sequencing read from the ribosome profiling data, nucleotide positions 16-18 were considered to be at the ribosomal A site where codon selection occurs (Ingolia et al. 2009). Only those reads with exactly 28 nucleotides and 0 ambiguous sites were used to ensure the accurate determination of positions 16-18. We calculated the fraction of in-frame codons by comparing the read sequences with annotated yeast coding sequences. Consistent with what was previously reported (Ingolia et al. 2009), the majority of codons at positions 16-18 were in-frame in the ribosome profiling data. In the mRNA-Seq data, the fraction of each phase was close to one third, as expected. All out-of-frame codons were excluded. The probability of incorrect codon assignment was low, because only codons misaligned by at least 3 nucleotides may be assigned incorrectly. Transposons and uncharacterized genes were removed. Our *CST* estimation procedure is as follows.

We first calculated f_i, the observed frequency of codon i, in the ribosome profiling data by

$$f_i = \frac{\sum_{j=1}^{N} c_{ij}}{\sum_{i=1}^{61}\sum_{j=1}^{N} c_{ij}},$$

where c_{ij} is the count of codon i in mRNA j positioned at the ribosomal A site measured by ribosome profiling and N is the number of genes with ribosome profiling

data (N > 3000 for both rich and starvation conditions). The expected ribosome footprint frequencies of codon i (F_i) when all codons have equal *CST* can be calculated based on the frequency of the codon in the mRNA-Seq data using

$$F_i = \frac{\sum_{j=1}^{N}(R_j C_{ij})}{\sum_{i=1}^{61}\sum_{j=1}^{N}(R_j C_{ij})},$$

where R_j is the translational initiation rate of mRNA j and C_{ij} is the count of codon i in mRNA j measured by mRNA-Seq. Then, the relative codon selection time for codon i is calculated by

$$CST_i = f_i / F_i .$$

We used an iterative approach to estimate the translational initiation rates that appear in Equation for F_i. We first used $R_j = 1$ for all j. After the *CST* is calculated for each codon, the elongation rate e_j of mRNA j (i.e., the number of codons translated per unit time) is calculated by

$$e_j = \frac{L_j}{\sum_{i=1}^{61}(D_{ij} CST_i)},$$

where L_j is the number of codons in each molecule of mRNA j and D_{ij} is the number of codon i in each molecule of mRNA j. The translational initiation rate R_j can be estimated from

$$R_j = e_j d_j ,$$

where d_j is the ribosome density on mRNA j (i.e., the number of ribosomes per codon) and can be estimated by

$$d_j = \frac{\sum_{i=1}^{61} c_{ij}}{\sum_{i=1}^{61} C_{ij}}.$$

We then used the newly estimated translational initiation rates to calculate *CST*s. After 10 iterations, *CST* estimates converge and are considered as our final estimates. Because our estimates of *CST*s are relative values, we rescaled them by setting the maximal observed value at 1.

CST estimates from different experimental replicates were highly correlated (r = 0.79, $P = 6 \times 10^{-14}$) and were thus pooled for the rest of the analysis. Three different sets of initial values of translational initiation rates (uniform, proportional to *CAI* of each gene, inversely proportional to *CAI*) were used in *CST* estimation and they resulted in identical estimates of *CST*s. Thus, *CST* estimation does not depend on the initial values of *R*. The standard errors of the *CST* estimates were estimated by bootstrapping genes present in the ribosomal profiling data 1000 times. The *CST* estimates from two different media (rich and starvation) are also very similar. To ensure no mistake in the estimation of *CST*, the first two authors of this paper independently derived the formulas, wrote the computer programs, and estimated the *CST*s, and their results were virtually identical.

4.9.3 Estimation of synonymous codon usage bias in yeast

There are two commonly used measures of synonymous codon usage bias. The first is the *relative synonymous codon usage* (*RSCU*), defined by the frequency of a codon relative to the average frequency of all of its synonymous codons in a set of highly expressed genes (Sharp et al. 1986). Codons with *RSCU* >1 are preferred and those with *RSCU* <1 are unpreferred. To compare the usage of all 61 sense codons, we also used *RSCU'* = *RSCU*/*n*, where *n* is the number of synonymous codons of an amino acid. *RSCU'* of a codon is the proportion of use of a given codon among synonymous choices in a set of highly expressed genes. The second commonly used measure of synonymous codon usage bias is the *codon adaptation index* (*CAI*), which is calculated for a gene, and measures its usage of high-*RSCU* codons (Sharp and Li 1987). Briefly, *CAI* of a gene is the geometric mean of *RSCU* divided by the highest possible geometric mean of *RSCU* given the same amino acid sequence. *CAI* is a positive number no greater than 1. The greater the *CAI*, the more prevalent are preferred codons in the gene.

We first selected 200 most highly expressed genes based on a previous study (Beyer et al. 2004). Sixteen of these genes did not have expression information in another study (Holstege et al. 1998) and 4 had expression levels lower than 4 times the genomic average, that is 2.7 mRNA/cell reported in an earlier study (Holstege et al. 1998). The remaining 180 highly expressed genes were used to calculate *RSCU* and *RSCU'* for each codon. Our *RSCU* estimates were highly correlated with those

previously reported (Sharp and Li 1987) ($r = 0.995$, $P < 0.001$, permutation test). *CAI* was calculated for each yeast gene and for each version of *mCherry* based on the *RSCU* values obtained above, following a previous study (Sharp and Li 1987).

We also estimated the effective number of codons (*Ncp*) for each gene, after controlling the GC content of the gene (Wright 1990; Novembre 2002). We separately estimated the frequency (*f*) of each of the 61 sense codons in each gene. We then estimated Spearman's rank correlation (ρ) between *Ncp* and *f* among all genes for each codon. Among synonymous codons, those with more negative ρ values are considered to be more preferred (Hershberg and Petrov 2009).

4.9.4 Empirical test of the proportional rule

We measured the Euclidian distance and Manhattan distance in synonymous codon usage from the observed values to the values predicted from the observed tRNA fractions using the proportional rule. To evaluate whether the observed distances are shorter than expected by chance, we conducted a computer simulation with 10^6 replications under random codon usage. That is, the frequency of a synonymous codon is uniformly distributed between 0 and 1 with the constraint of the total frequency of all synonymous codons being 1. We then obtained the distribution of the distance between a random codon usage and the codon usage predicted from the observed tRNA fractions. We also conducted a second simulation with 10^6 replications, in which tRNA factions vary randomly according to the above uniform distribution. We then obtained the distribution of the distance between the observed codon usage and that predicted from random tRNA fractions. This way, the potential confounding effect of genomic GC content on the assumed null distribution of codon usage becomes irrelevant to the test. We similarly tested the square rule and the truncation rule.

4.9.5 Distance to native codon usage

We developed an index, distance to native codon usage (D_{ncu}), to measure how different the codon usage of a (heterologous) gene is from the overall codon usage of the host cell, which is presumably balanced with tRNA concentrations. First, the Euclidean

distance in synonymous codon usage between the heterologous gene and the host is calculated for each of the 18 amino acids with at least two synonymous codons by

$$D_i = \sqrt{\sum_{j=1}^{n_i} (Y_{ij} - X_{ij})^2},$$

where Y_{ij} is the fraction of codon j among the synonymous codons of amino acid i for the heterologous gene and X_{ij} is the fraction of codon j among the synonymous codons in the host transcriptome, n_i is the number of synonymous codons for amino acid i. D_{ncu} of the gene is defined as the weighted geometric mean of D_i, or

$$D_{\text{ncu}} = (\prod_{i=1}^{k} D_i^{m_i})^{\frac{1}{l}},$$

where $k \leq 18$ is the number of amino acid types encoded by the gene excluding Met and Trp, which have no synonymous codons, m_i is the number of amino acid i found in the protein, and l is the protein length excluding Met and Trp residues. By definition, D_{ncu} is between 0 and 1.

4.9.6 Yeast experiments

The *mCherry* gene sequence was obtained from a previous study (Shaner et al. 2004). We designed four synonymous DNA sequences encoding the same mCherry peptide. The first 56 nucleotides were the same for all four sequences to avoid potential effects on the mRNA secondary structure, which affects protein translation (Kudla et al. 2009; Gu et al. 2010; Tuller et al. 2010b). The GC contents of the four sequences (42-44%) were also made similar to each other and to the average value in yeast coding sequences (40%). In all sequences, synonymous codons were randomized in order and thus were unlikely to cause differences in order-related effects (Cannarozzi et al. 2010). The different versions of *mCherry* DNA sequences were synthesized by Blue Heron Biotechnology. They were cloned into p426GPD (Mumberg et al. 1995) at SpeI and XhoI (New England Biolabs; Promega) and are under the control of the *GPD* promoter. The plasmids were subsequently transformed individually into a haploid yeast cell (BY4742) with *vYFP* (Nagai et al. 2002) inserted into Chr XII (He et al. 2010). The genotype of the cell is *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 hoΔ0::*P_{GPD}*-Venus*.

We measured the expressions of mCherry and vYFP in log growth phase in Yeast extract/Peptone/Dextrose (YPD) media by florescence-activated cell scanning (FACSCalibur, BD). Fluorescence of mCherry was measured from FL4 with a 670 nm pass filter and fluorescence of vYFP was measure from FL1 with a filter having a 30 nm bandpass centered on 530 nm. Yeast cells with mCherry fluorescence signals greater than the BY4742 negative control cells (i.e., mCherry fluorescence signals >10) were gated. We retrieved the forward scatter (FSC, which is proportional to cell size) and mCherry and vYFP fluorescence signals for all gated cells. The expression levels of fluorescent proteins were defined as their fluorescence signals divided by FSC. The mean mCherry expression level is 3.388±0.002, 6.468±0.007, 14.003±0.032, and 14.544±0.022 for the strains carrying *mCherry*-1, 2, 3, and 4, respectively. Expression levels of mCherry and vYFP were negatively correlated for each strain (*mCherry*-1: r = -0.22; *mCherry*-2: r = -0.57; *mCherry*-3: r = -0.60; *mCherry*-4: r = -0.62; $P < 2.2 \times 10^{-16}$ in all cases). All gated cells were then grouped into 3 (Figure 4.3D) bins with equal mCherry expression ranges. For each genotype, multiple independently transformed strains were examined on different days, but the results were highly similar. We thus combined all results obtained from different strains of the same genotype. The total numbers of cells measured were 456333, 648792, 352863, and 793832, respectively, for the strains carrying *mCherry*-1, 2, 3 and 4 (Figure 4.3B). To confirm that our results were not due to random secondary mutations, we removed the plasmids from each strain by using 5'-FOA media to select against the plasmids, and then measured the vYFP fluorescence intensities. We also sequenced the entire plasmid DNA from each of the four strains.

To compare the *vYFP* mRNA levels among strains, we extracted the total RNA (RiboPure-Yeast Kit, Ambion) from three independently transformed strains of each genotype. The total RNA was reversely transcribed into cDNA (Moloney Murine Leukemia Virus Reverse Transcriptase, Invitrogen) with random hexamer primers. The *vYFP* mRNA level was measured by quantitative polymerase chain reaction (7300 Real-Time PCR System, Applied Biosystems) with *ACT1* as an internal control. The primers for *vYFP* are 5' – CATGGCCAACACTTGTCACT– 3' and 5' –

TACATAACCTTCGGGCATGG– 3, while the primers for *ACT1* are 5' - CTGCCGGTATTGACCAAACT - 3' and 5' – CGGTGATTTCCTTTTGCATT – 3'.

4.9.7 Multivariate regression analysis

The software package RELAIMPO (http://cran.r-project.org/web/packages/relaimpo/) was used for a multivariate regression analysis of the yeast experimental data from all cells of the four strains. We compared the relative importance of D_{ncu} and *CAI* in explaining the among-cell variation in vYFP signal by the LMG method and used 1000 bootstrap replications to determine the statistical significance. Use of other methods (LAST, FIRST, and PRATT) implemented in RELAIMPO gave similar results.

4.9.8 Computer simulation of the evolution of synonymous codon usage bias

We simulated the evolution of synonymous codon usage in an asexual haploid unicellular digital organism. In this organism, we focused on a single amino acid with four synonymous codons (codon1 to codon4) that are respectively recognized by four distinct tRNA species (tRNA1 to tRNA4). We assume that the relative concentrations of the four tRNA species are 2^0, 2^1, 2^2, and 2^3, respectively. The digital organism has ten genes with relative (mRNA and protein) expression levels from 2^0 to 2^9, respectively. These genes each have 12 codons that are sampled from the four synonymous codons. We started the simulation with exactly the same usage of the four synonymous codons in each gene. Synonymous mutations among codons all have the same rates and the total mutation rate per genome is assumed to be one synonymous change per generation. The relative *CST* for a codon is assumed to equal the number of times the codon is used in translation divided by the number of corresponding tRNA molecules. The total time (*T*) required for translating all the proteins can be considered as the generation time. *T* can be calculated by summing up the *CSTs* of all codons in all transcripts if there is only one ribosome in the cell. If there are *m* ribosomes in the cell, the time required would simply be *m* times shorter. Thus, without loss of generality, we assume $m = 1$. A strain with a shorter generation has a higher fitness and will spread in the population. Genetic drift is simulated by random sampling of cells for the next generation. The population size is 10^4

individuals and the simulation lasts for 500 generations. We repeated the simulation 1000 times. Our results did not change when we simulated the evolution for more generations. By contrast, when we removed the natural selection for translational efficiency in simulation, the phenomena observed in Figure 4.4 disappeared.

Note that, in the simulation, we allow codon usage to evolve while fixing tRNA concentrations. If tRNA concentrations evolve while the codon usage is fixed, we also expect to observe the rebalance of codon-tRNA usage, but the correlation (or the lack of) between CUB and gene expression level will not change during this evolutionary process. In reality, tRNA concentrations and synonymous codon usage likely co-evolve to regain the balance. As long as codon usage is allowed to evolve, we expect stronger CUB to appear in more highly expressed genes, as demonstrated in Figure 4.4.

4.10 Appendices

4.10.1 Mathematical proof that proportional codon usage maximizes translational efficiency

Without loss of generality, we assume that an amino acid is encoded by synonymous codons 1 and 2, which are respectively recognized by isoaccepting tRNAs 1 and 2. Let the relative usage of the two codons be p_1 and $p_2 = 1\text{-}p_1$ and the relative concentrations of the two tRNAs be q_1 and $q_2 = 1\text{-}q_1$, respectively. Let the codon selection time for the two synonymous codons be t_1 and t_2, respectively. Thus, the expected codon selection time for the amino acid concerned is $t = p_1t_1 + p_2t_2$. When tRNAs are in shortage, the local concentrations of tRNA 1 and 2 are aq_1/p_1 and aq_2/p_2, where a is a constant. Because codon selection time is proportional to the inverse of the local tRNA concentration, we have $t = \frac{p_1 b}{aq_1 / p_1} + \frac{p_2 b}{aq_2 / p_2}$, where b is another constant. The above formula can be simplified to $t = b(p_1^2 / q_1 + p_2^2 / q_2) / a = (b / a)[1 + (p_1 - q_1)^2 / (q_1 q_2)]$. It is easy to find that t reaches its minimal value of b/a when $p_1 = q_1$ and $p_2 = q_2$. In other words, the expected codon selection time is minimized and thus translational efficiency is maximized when relative synonymous codon frequencies equal relative tRNA concentrations. Under this condition, codon selection time equals b/a for both codons and local tRNA concentration equals a

for both tRNAs. A full treatment considering tRNA cycle and kinetics gave the same result (Liljenstrom et al. 1985).

4.10.2 Concentrations of ternary complexes in *E. coli*

It has been reported that the physiological concentration of the ternary complex is ~200 nM for Phe tRNA and Lys tRNAs in *E. coli* (Uemura et al. 2010). Because the number of Phe tRNA and Lys tRNA molecules per cell is 1830 and 4300, respectively (Jakubowski and Goldman 1984), we calculated that the Phe tRNA concentration is $1830/(6.02\times10^{23})/(1.1\times10^{-15}) = 2.8\times10^{-6}$ M = 2800 nM, where 6.02×10^{23} is the number of molecules per mole and 1.1×10^{-15} liter is the average volume of an *E. coli* cell. Similarly, Lys tRNA concentration is estimated to be 6500 nM. Thus, about 200/[(2800+6500)/2] = 4.3% of tRNAs are in ternary complexes. Because there are $\sim1.2\times10^{4}$ ribosomes per *E. coli* cell (Jakubowski and Goldman 1984), ribosome concentration is ~18,000 nM. Thus, the ratio in the concentration of ternary complexes to that of ribosomes is expected to be $200\times20/18000 = 0.22$, if Lys and Phe can represent all 20 amino acids in ternary complex concentration.

4.10.3 Impact of potential errors in translation on our experiments

Proponents of the translational accuracy hypothesis might argue that, because different synonymous codons have different mistranslation rates (Precup and Parker 1987; Rodnina and Wintermeyer 2001) and preferred codons are considered to be more accurately translated than unpreferred codons (Drummond and Wilke 2008), the *mCherry* with a low *CAI* is expected to produce fewer functional protein molecules than the *mCherry* with a high *CAI* even when the same numbers of protein molecules are produced. In other words, using red florescent signals may have led to a more severe underestimation of protein expression for the *mCherry* with a low *CAI* than for that with a high *CAI*. The average mistranslation rate has been estimated to be $\sim5\times10^{-4}$ per codon, and unpreferred codons have been posited to undergo mistranslation five times as often as preferred codons (Drummond and Wilke 2008). Based on these numbers and the *CAI*s of the four *mCherry* versions (Figure 4.3B), we assume that the mistranslation rate is 10×10^{-4}, 8×10^{-4}, 5×10^{-4}, and 2×10^{-4} per codon for *mCherry*-1 to *mCherry*-4, respectively.

Let us further assume that no mistranslated protein fluoresces. Given the length of mCherry (236 amino acids), we expect that 11.8%, 9.44%, 5.9%, and 2.36% of mCherry-1 to mCherry-4 proteins respectively fail to fluoresce due to mistranslation. On this assumption, we corrected mCherry expression levels from the observed florescent signals. We also conducted a better control of mCherry expression among strains by dividing cells of each strain into 15 bins based on the above corrected mCherry expression. Again, we observed a lower vYFP expression when the D_{ncu} of the *mCherry* gene is higher, across the range of mCherry expressions shared by the three strains. This result is conservative, because only a minority of mistranslations are expected to prevent fluorescence, and it is likely that we have overcorrected the effect of mistranslation.

4.11 Acknowledgments

Jian-Rong Yang helped with computational analysis; Nathaniel M. Pearson and Calum Maclean helped with experiments. We thank Barry Williams for discussion and Hiroshi Akashi, Meg Bakewell, Chungoo Park, Josh Plotkin, and Claus Wilke for valuable comments.

Table 4.1 Comparison between the old and new models of translational efficiency by unequal codon usage.

Comparisons	Old model	New model
Ternary complexes of aminoacylated tRNA + eEF-1α + GTP	In excess.	In shortage.
Translational speeds of synonymous codons in wild-type cells	Faster for those with higher cognate tRNA concentrations.	Equal, because codon usage has been optimized to be proportional to cognate tRNA concentrations.
Translational speeds of synonymous codons in mutant cells	Faster for those with higher cognate tRNA concentrations.	Unequal when the codon-tRNA balance is broken. Slower for codons with higher ratios between the codon fraction and the cognate tRNA fraction.
Why is the codon usage bias stronger in more highly expressed genes?	Fast translation of highly expressed genes is favored over fast translation of lowly expressed genes.	Synonymous mutations in highly expressed genes have larger effects than those in lowly expressed genes in regaining the codon-tRNA balance, which increases the overall translational efficiency of the cell.
Why is the codon usage proportional to cognate tRNA concentration?	No explanation.	It maximizes the overall cellular translational efficiency when ternary complexes are in shortage.
How to reach the highest cellular translational efficiency in making a synthetic cell?	Exclusive use of preferred codons.	Codon usage in proportion to cognate tRNA concentrations.

Figure 4.1 Relative codon selection times (*CST*s) in wild-type yeast cells in rich media.

(A) *CST* (grey bars) and *RSCU'* (orange dots) of each sense codon. *CST*s are rescaled such that the maximal observed value is 1. Error bars show one standard error, estimated by the bootstrap method. No significant negative correlation between *CST* and (B) *RSCU'*, (C) tRNA gene copy number, or (D) tRNA concentration. Spearman's rank correlation coefficients (ρ) and associated *P* values are presented above each panel. The *P* value in (B) is calculated by a permutation test because of the non-independence among *RSCU'* values of synonymous codons. (E) No dip in *RSCU'* at the ribosomal A site, compared to P, E, and other neighboring sites. Geometric means of *RSCU'* is calculated at each codon position (as in the calculation of *CAI*) for ribosome profiling sequencing reads and mRNA sequencing reads, respectively; the ratio at each position is presented. Error bars show one standard error estimated by bootstrapping sequencing reads 1000 times.

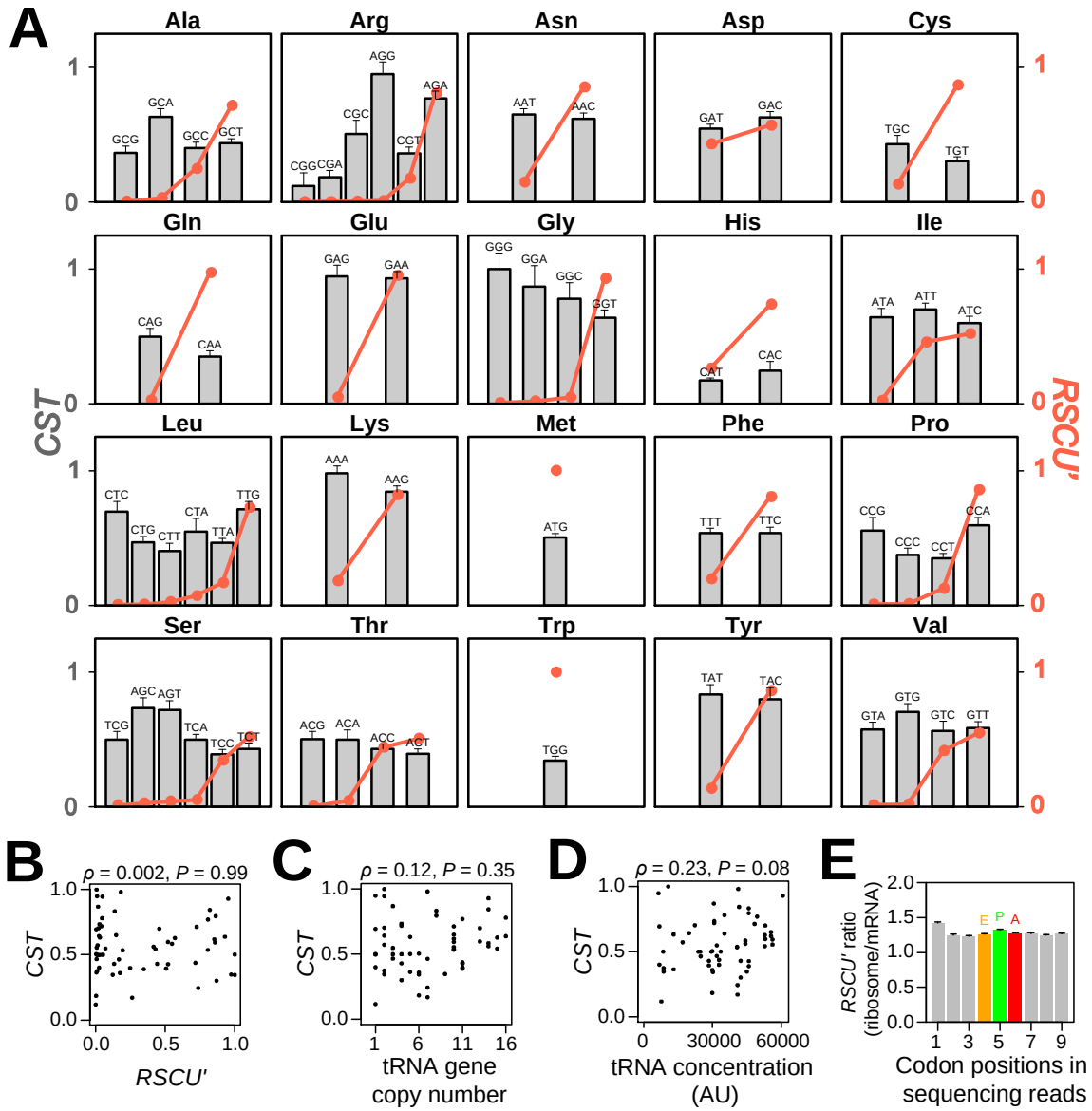
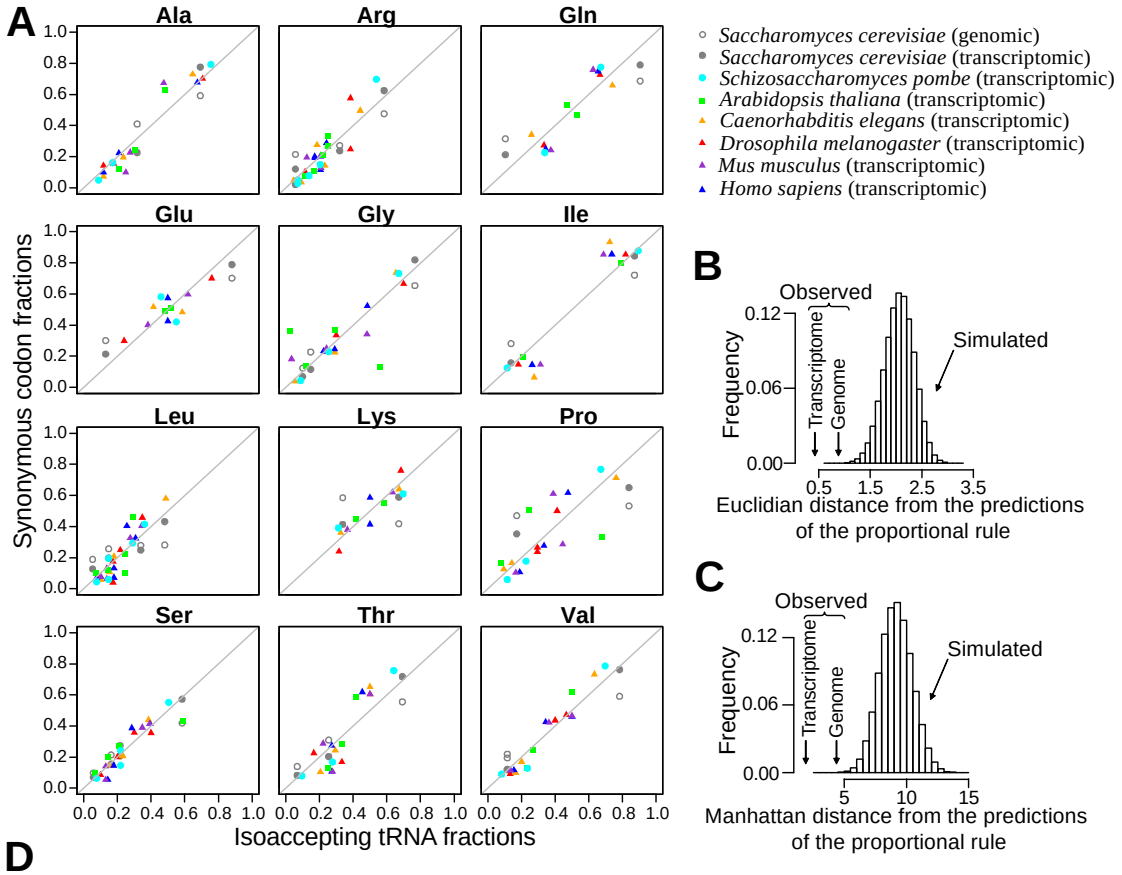


Figure 4.2 Synonymous codons are used in proportion to cognate tRNA concentrations.
(A) Relative uses of synonymous codons in the transcriptomes of seven model eukaryotes are compared to the relative concentrations of cognate tRNAs measured from gene copy numbers, for the 12 amino acids that have at least two isoaccepting tRNA species. For comparison, genomic synonymous codon usage in *S. cerevisiae* is also presented. The diagonal line shows the predicted proportional relationship between tRNA concentrations and cognate codon uses that maximizes translational efficiency under tRNA shortage. (B) Euclidian and (C) Manhattan distances between the observed synonymous codon usage in *S. cerevisiae* and the prediction by the proportional rule are significantly smaller than chance expectations. Euclidian and Manhattan distances are defined by $\sqrt{\sum_{i=1}^{k}(p_i - q_i)^2}$ and $\sum_{i=1}^{k}| p_i - q_i |$, respectively, where p_i and q_i are codon and cognate tRNA fractions, respectively, and k is the number of different tRNA species for the amino acid concerned. The chance expectations are shown by the frequency distributions of the distances under uniformly random codon usage, determined from 10^6 simulations. (D) Euclidian and Manhattan distances between the observed synonymous codon usage and the predictions under the proportional rule, square rule, and truncation rule, respectively. *P* values indicate the probability that a distance generated by random codon usage is smaller than the observed distance, determined by 10^6 simulations. Log_{10}(likelihood ratio) measures the likelihood of the proportional rule, relative to the square rule, given the actual codon usage.



D

Transcriptomes	Euclidian distance (*P* value)			Manhattan distance (*P* value)			Log_{10}(likelihood ratio)
	Proportional	Square	Truncation	Proportional	Square	Truncation	Proportional / Square
S. cerevisiae	0.42 ($<1\times10^{-6}$)	0.90 (1×10^{-6})	1.45 (3×10^{-5})	1.91 ($<1\times10^{-6}$)	4.49 (1×10^{-5})	7.29 (7×10^{-4})	3.46×10^{5}
S. pombe	0.44 ($<1\times10^{-6}$)	0.62 ($<1\times10^{-6}$)	1.56 ($<1\times10^{-6}$)	2.22 ($<1\times10^{-6}$)	3.06 ($<1\times10^{-6}$)	7.66 (1×10^{-5})	4.60×10^{6}
A. thaliana	0.84 (6×10^{-6})	1.18 (8×10^{-5})	2.23 (2×10^{-3})	3.55 (1×10^{-6})	4.68 (3×10^{-6})	11.78 (3×10^{-2})	3.58×10^{8}
C. elegans	0.49 ($<1\times10^{-6}$)	0.71 ($<1\times10^{-6}$)	1.65 ($<1\times10^{-6}$)	2.44 ($<1\times10^{-6}$)	3.47 ($<1\times10^{-6}$)	8.41 (6×10^{-5})	1.40×10^{5}
D. melanogaster	0.43 ($<1\times10^{-6}$)	0.66 ($<1\times10^{-6}$)	1.83 (8×10^{-6})	1.95 ($<1\times10^{-6}$)	3.25 ($<1\times10^{-6}$)	9.41 (3×10^{-4})	1.87×10^{7}
M. musculus	0.60 ($<1\times10^{-6}$)	0.67 ($<1\times10^{-6}$)	2.20 (5×10^{-5})	2.83 ($<1\times10^{-6}$)	3.12 ($<1\times10^{-6}$)	11.4 (3×10^{-3})	4.07×10^{5}
H. sapiens	0.49 ($<1\times10^{-6}$)	0.51 ($<1\times10^{-6}$)	2.03 (2×10^{-5})	2.42 ($<1\times10^{-6}$)	2.51 ($<1\times10^{-6}$)	10.27 (6×10^{-4})	7.36×10^{6}

Figure 4.3 Experimental evidence for the impact of codon usage imbalance on translational efficiency.

(A) Experimental design for examining the impact of mCherry expression on the expression of the reporter vYFP. An *mCherry* gene is constitutively expressed from a 2-micron plasmid in *S. cerevisiae*, whereas *vYFP* is constitutively expressed from Chromosome XII. Four different synonymous versions of *mCherry* are compared. (B) The codon adaptation indices (*CAI*s) of the four synonymous *mCherry* sequences (circled numbers), in comparison to *CAI*s of all *S. cerevisiae* genes. (C) Values of distance to native codon usage of yeast (D_{ncu}) for the four *mCherry* sequences, in comparison to that of all *S. cerevisiae* genes. (D) Relationship between vYFP expression and the *CAI* or D_{ncu} of *mCherry*, when the mCherry expression is controlled for. Error bars, which are barely seen, show one standard error. (E) vYFP expressions in the four strains after the removal of the plasmids that carry *mCherry*. Error bars show one standard error. (F) *vYFP* mRNA levels of the four strains relative to that of the wild-type strain, which does not carry *mCherry*. The mean expressions from three biological replications and the standard errors are presented.

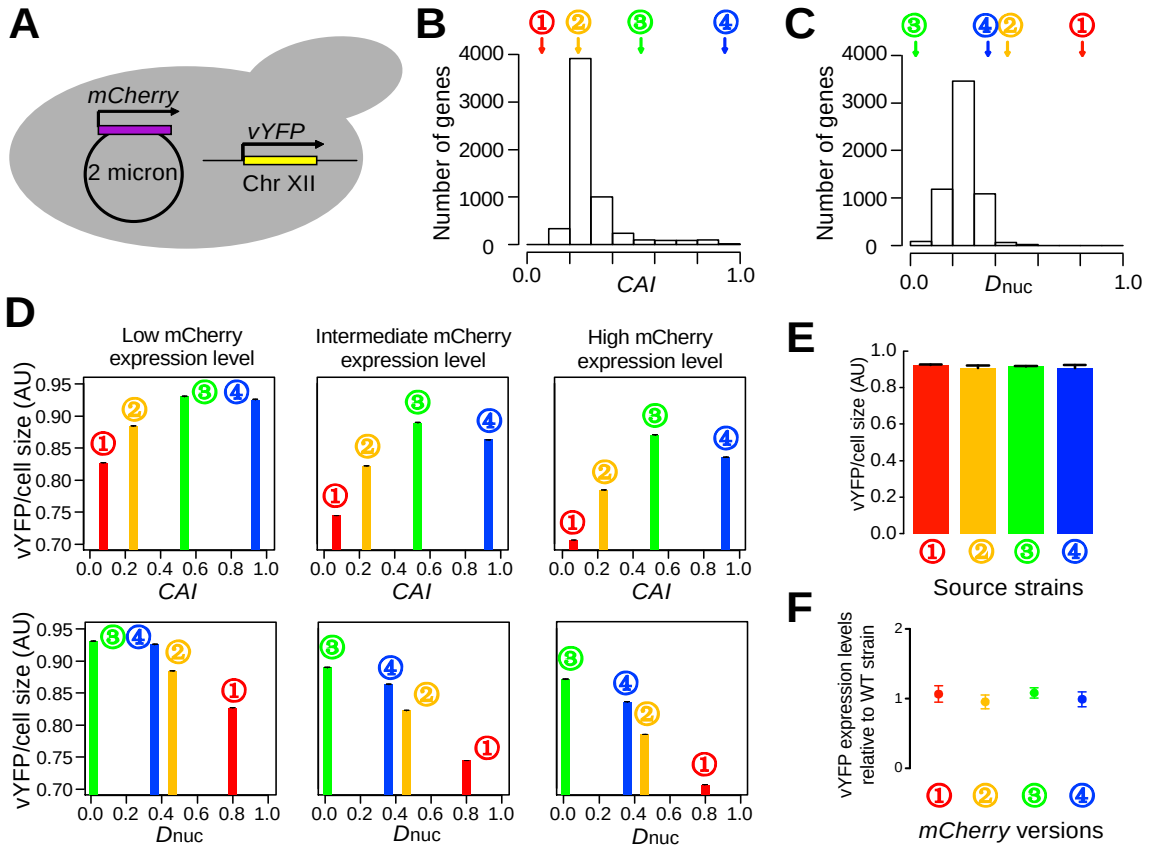


Figure 4.4 Computer simulation demonstrates that selection for translational efficiency results in the preferential use of codons with abundant cognate tRNAs in highly expressed genes.

Ten genes with different expression levels are considered for a haploid organism. Four synonymous codons of an amino acid are each recognized by its cognate tRNA. Concentrations of the four tRNAs differ, but the initial codon frequencies are equal. Synonymous mutations, genetic drift, and natural selection for translational efficiency are considered (see Materials and Methods). (A) Overall changes of transcriptomic codon usage averaged from 1000 simulation replications. Error bars show one standard deviation. (B) Highly expressed genes evolved stronger codon usage biases than lowly expressed genes. The averages from 1000 simulation replications are presented. Error bars show one standard deviation. (C) Evolutionary changes in the usage of codon4, the codon recognized by the most abundant tRNA, in a randomly chosen simulation replication.

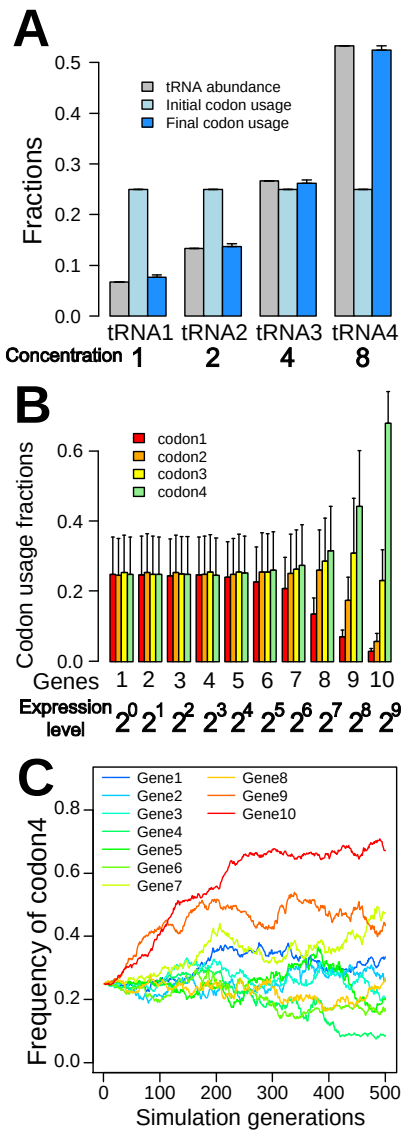


Figure 4.5 Amino acids are used approximately in proportion to cognate tRNA concentrations.

(A) Relative uses of amino acids estimated from the transcriptomic data of 7 model eukaryotes are compared to the relative concentrations of their cognate tRNAs measured from gene copy numbers. The diagonal line shows the predicted proportional relationship between tRNA concentrations and cognate amino acid frequencies that maximizes translational efficiency under tRNA shortage. P_E (or P_M) is the probability that the Euclidian (or Manhattan) distance between the amino acid usage randomly generated under a uniform distribution and that predicted by the proportional rule is smaller than the observed distance, and is estimated from 10^6 simulations. The distance definitions are the same as those in the legend of Fig. 2, except that *i* is an amino acid instead of a codon. (B) The average *CST* of an amino acid in *S. cerevisiae* is negatively correlated with the availability of its cognate tRNAs, which is measured by the fraction of cognate tRNA genes among all tRNA genes divided by the frequency of the amino acid estimated from the transcriptome. The *P*-value is determined from 1000 permutations of *CST*s.

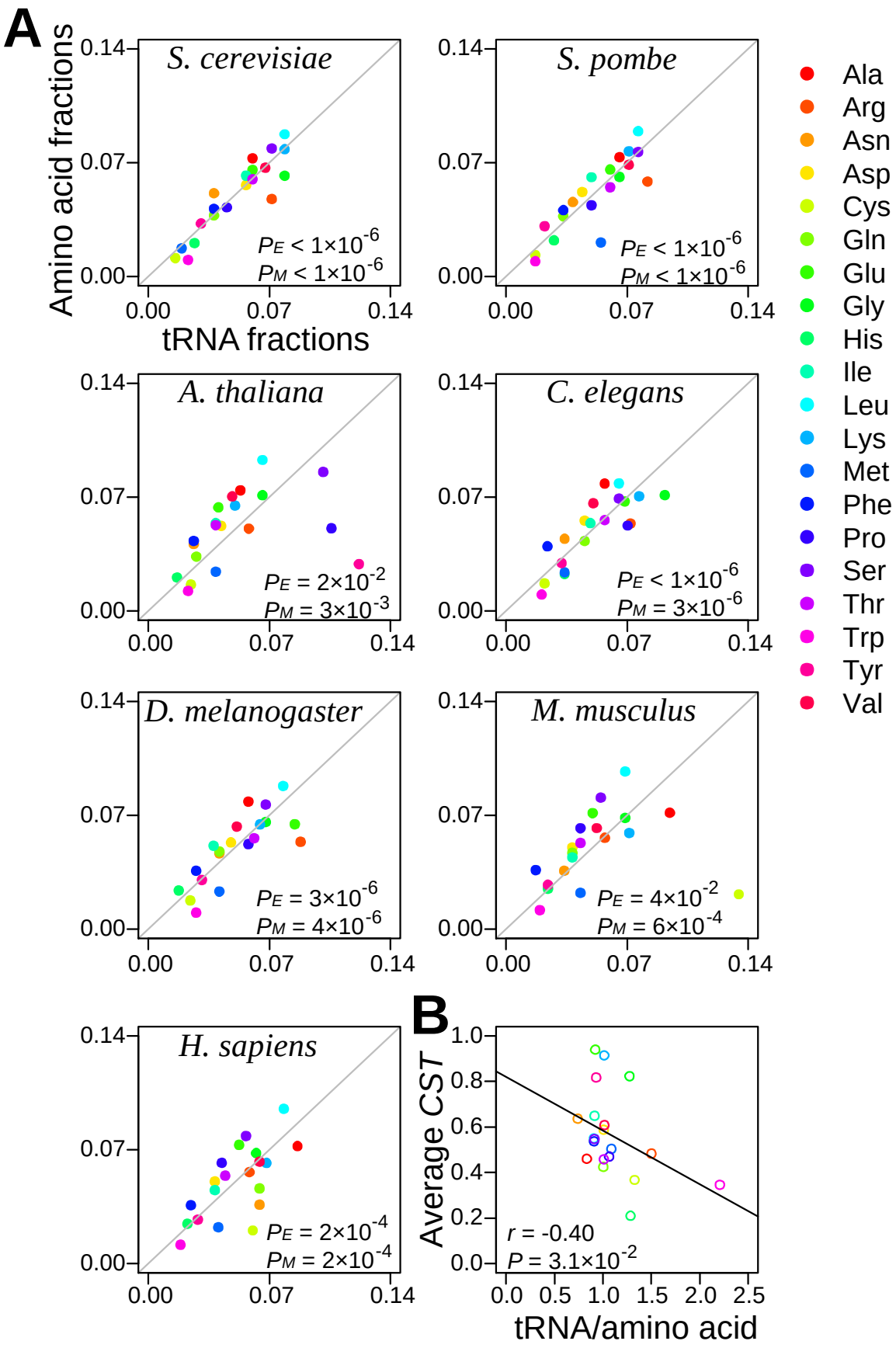

A
Amino acid fractions
tRNA fractions
S. cerevisiae
PE < 1×10-6
PM < 1×10-6
S. pombe
PE < 1×10-6
PM < 1×10-6
A. thaliana
PE = 2×10-2
PM = 3×10-3
C. elegans
PE < 1×10-6
PM = 3×10-6
D. melanogaster
PE = 3×10-6
PM = 4×10-6
M. musculus
PE = 4×10-2
PM = 6×10-4
H. sapiens
PE = 2×10-4
PM = 2×10-4
0.00
0.07
0.14
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
B
Average CST
tRNA/amino acid
r = -0.40
P = 3.1×10-2
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5

Figure 4.6 Similarity in transcriptomic codon usage across cell cycle stages, developmental stages, and tissues.

Distributions of pairwise Pearson's correlations of codon usage among (A) mitotic cell cycle stages in *S. cerevisiae*, (B) developmental stages in *C. elegans*, (C) tissues and developmental stages in *D. melanogaster*, and (D) among tissues in *H. sapiens*.

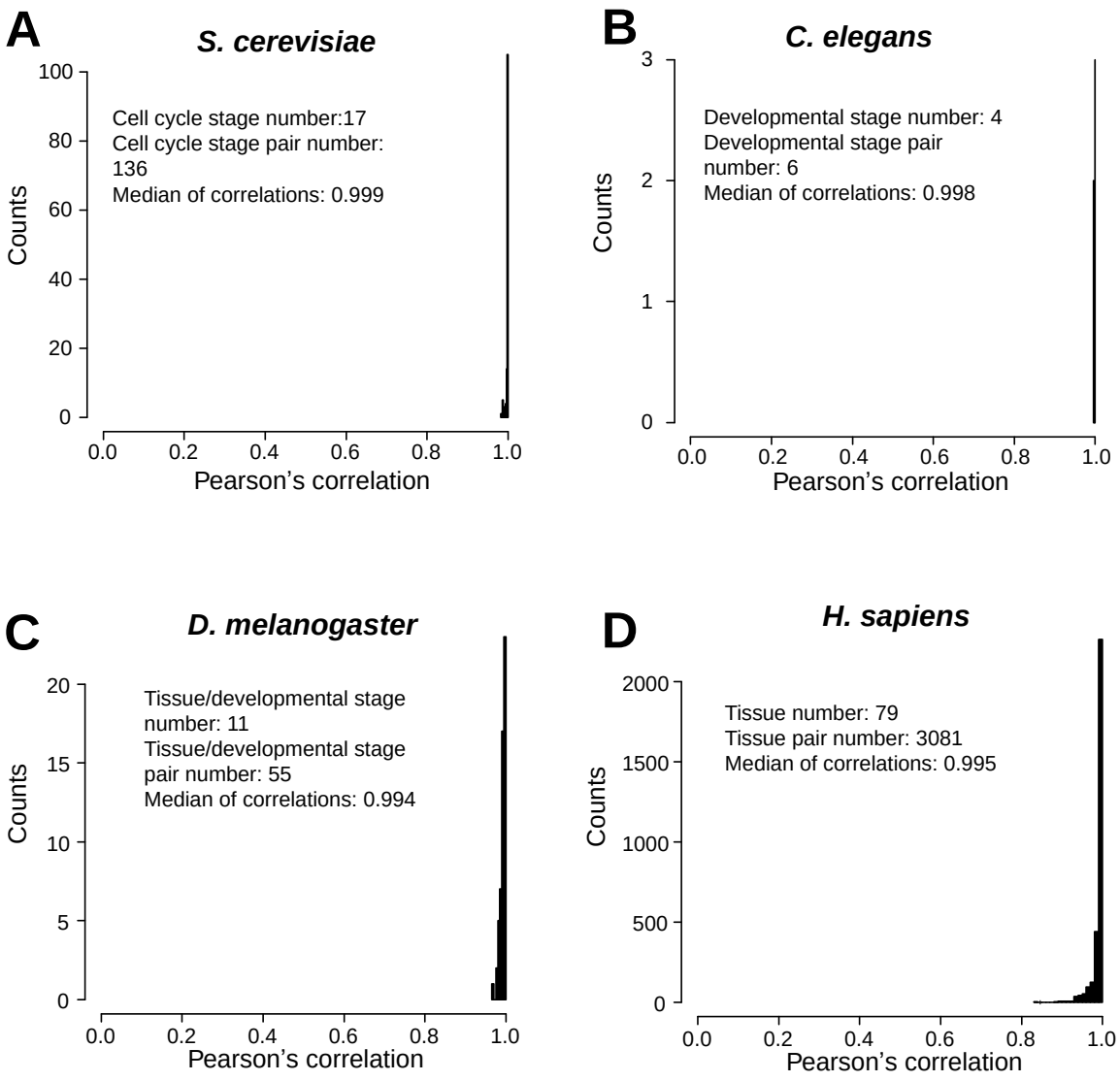


Figure 4.7 Evolutionary models of synonymous codon usage bias.

Three models that differ in the involvement of natural selection for translational accuracy and efficiency in the evolution of codon usage bias. Models I and II can be rejected by the existing data, whereas model III is supported by available data.

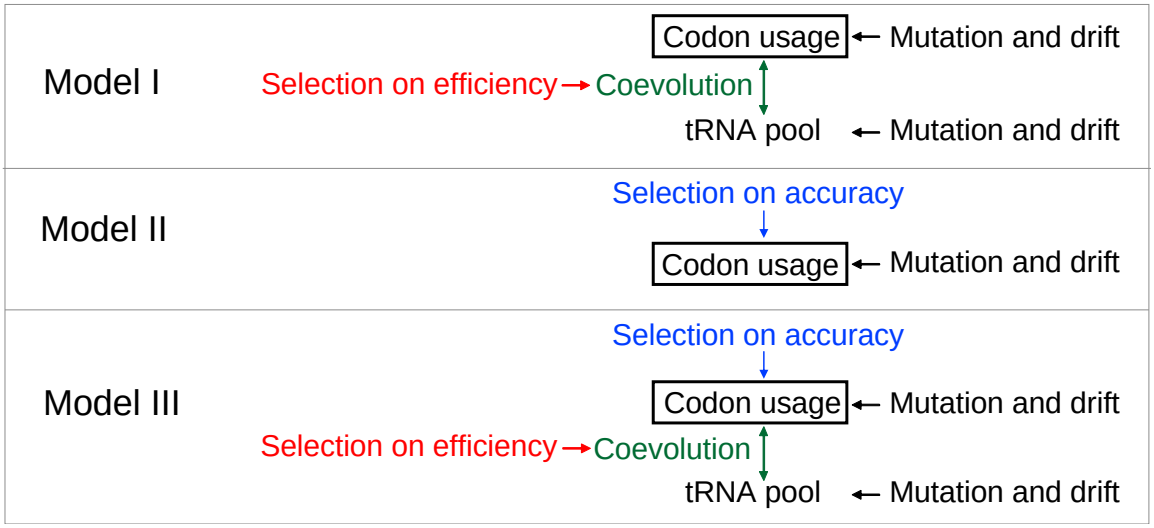
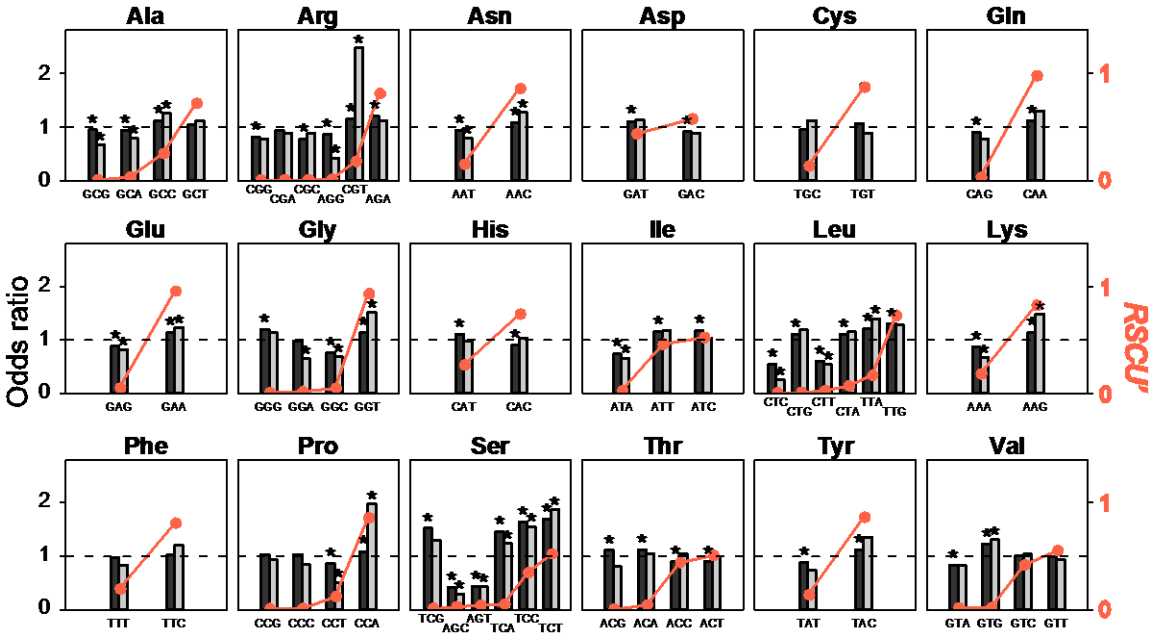


Figure 4.8 Matches and mismatches between preferred codons and accurate codons in *S. cerevisiae*.

Odds ratio (bars) measures the enrichment of a synonymous codon at evolutionarily conserved amino acid residues relative to that at non-conserved residues and is used as a proxy for translational accuracy. *RSCU'* (orange dots) measures the preference in codon usage. Odds ratios are estimated from either all genes (black) or the 10% most highly expressed genes (grey) of *S. cerevisiae*. Asterisks indicate 5% significance in the deviation of an odds ratio from 1 (uncorrected for multiple testing).



4.12 References

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**(3): 927-935.

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev* **11**(6): 660-666.

Akashi H. 2003. Translational selection and yeast proteome evolution. *Genetics* **164**(4): 1291-1303.

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* **99**(6): 3695-3700.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev* **54**(2): 198-210.

Beyer A, Hollunder J, Nasheuer HP, Wilhelm T. 2004. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics* **3**(11): 1083-1092.

Boehlke KW, Friesen JD. 1975. Cellular content of ribonucleic acid and protein in Saccharomyces cerevisiae as a function of exponential growth rate: calculation of the apparent peptide chain elongation rate. *J Bacteriol* **121**(2): 429-433.

Brockmann R, Beyer A, Heinisch JJ, Wilhelm T. 2007. Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput Biol* **3**(3): e57.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**(6106): 728-730.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**(3): 897-907.

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* **141**(2): 355-367.

Carlini DB, Stephan W. 2003. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics* **163**(1): 239-243.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**(2): 98-108.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1): 65-73.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol* **209**(1): 65-77.

Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. *PLoS Genet* **2**(12): e221.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**(16): e105.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**(5): 649-663.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**(2): 341-352.

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**(7): 287-289.
Forchhammer J, Lindahl L. 1971. Growth rate of polypeptide chains as a function of the cell growth rate in a mutant of Escherichia coli 15. *J Mol Biol* **55**(3): 563-568.
Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. *Genetica* **102-103**(1-6): 127-144.
Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* **425**(6959): 737-741.
Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6**(2): e1000664.
Gu X, Hewett-Emmett D, Li WH. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* **102-103**(1-6): 383-391.
Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol* **22**(7): 346-353.
He X, Qian W, Wang Z, Li Y, Zhang J. 2010. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**(3): 272-276.
Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet* **42**: 287-299.
Hershberg R, Petrov DA. 2009. General rules for optimal codon choice. *PLoS Genet* **5**(7): e1000556.
Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. 2009. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res* **19**(4): 657-666.
Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**(5): 717-728.
Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**(3): 389-409.
Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**(4): 573-597.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**(1): 13-34.
Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924): 218-223.
Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**(4): 789-802.

Jakubowski H, Goldman E. 1984. Quantities of individual aminoacyl-tRNA families and their turnover in Escherichia coli. *J Bacteriol* **158**(3): 769-776.
Kramer EB, Vallabhaneni H, Mayer LM, Farabaugh PJ. 2010. A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA* **16**(9): 1797-1808.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**(5924): 255-258.
Liljenstrom H, von Heijne G, Blomberg C, Johansson J. 1985. The tRNA cycle and its relation to the rate of protein synthesis. *Eur Biophys J* **12**(2): 115-119.
Mumberg D, Muller R, Funk M. 1995. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**(1): 119-122.
Nagai T, Ibata K, Park ES, Kubota M, Mikoshiba K, Miyawaki A. 2002. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol* **20**(1): 87-90.
Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19**(8): 1390-1394.
Ogle JM, Ramakrishnan V. 2005. Structural insights into translational fidelity. *Annu Rev Biochem* **74**: 129-177.
Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J Mol Biol* **268**(2): 322-330.
Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in Drosophila. *Proc Natl Acad Sci U S A* **94**(15): 7784-7790.
Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262**(23): 11351-11355.
Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. *Nucleic Acids Res* **12**(17): 6663-6671.
Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**(11): 2279-2286.
Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21**(1): 108-116.
Rodnina MV, Wintermeyer W. 2001. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu Rev Biochem* **70**: 415-435.
Shachrai I, Zaslaver A, Alon U, Dekel E. 2010. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Mol Cell* **38**(5): 758-767.
Shaner NC, Campbell RE, Steinbach PA, Giepmans BN, Palmer AE, Tsien RY. 2004. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotechnol* **22**(12): 1567-1572.
Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res* **16**(17): 8207-8211.

Sharp PM, Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**(3): 1281-1295.
Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**(13): 5125-5143.
Siwiak M, Zielenkiewicz P. 2010. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Comput Biol* **7**(7): e10000865.
Sokal RR, Rohlf FJ. 1995. *Biometry*. Freeman and Company, New York.
Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in Escherichia coli. *J Mol Biol* **207**(2): 365-377.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol* **24**(2): 374-381.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010a. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**(2): 344-354.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010b. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* **107**(8): 3645-3650.
Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD. 2010. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464**(7291): 1012-1017.
Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* **180**(3): 549-576.
Vimaladithan A, Farabaugh PJ. 1994. Special peptidyl-tRNA molecules can promote translational frameshifting without slippage. *Mol Cell Biol* **14**(12): 8107-8116.
von der Haar T. 2008. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* **2**: 87.
Warner JR. 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24**(11): 437-440.
Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009. Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One* **4**(9): e7002.
Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**(7199): 1239-1243.
Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87**(1): 23-29.
Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* **6**: 421.
Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J Mol Evol* **50**(1): 56-68.
Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* **26**(7): 1571-1580.

CHAPTER 5
THE GENOMIC LANDSCAPE OF ANTAGONISTIC PLEIOTROPY IN YEAST

5.1 Abstract

Antagonistic pleiotropy (AP) or genetic tradeoff is an important concept invoked frequently in theories of aging, cancer, genetic disease, and other common phenomena. But, it is unclear how prevalent AP is, which genes are subject to AP, and to what extent and how AP may be resolved. By measuring the fitness difference between the wild-type and null alleles of ~5000 nonessential genes in yeast, we find that, in any given environment, yeast expresses hundreds of genes that harm rather than benefit the organism, demonstrating widespread AP. Nonetheless, under sufficient selection, AP is often resolvable through regulatory evolution, primarily by trans-acting changes, although in one case we also detect a *cis*-acting change and localize its causal mutation. AP resolution, however, is slower in smaller populations, predicting more unresolved AP in multicellular organisms than in yeast. These findings provide the empirical foundation for AP-dependent theories and have broad biomedical and evolutionary implications.

5.2 Introduction

Antagonistic pleiotropy (AP) is a form of pleiotropy (Wagner and Zhang 2011) in which the relative advantage of two alleles of a gene is reversed in different components of fitness such as different sexes, developmental stages, and external environments. Note that, by definition, AP among environments is also a type of genotype by environment (G×E) interaction, which describes different phenotypic effects of a genetic change in different environments. The only distinction is that AP among environments is a special and strong type of G×E interaction where the effects of a genetic change in two environments are opposite.

AP is commonly invoked in explanations and models of senescence (Williams 1957), cancer (Rodier et al. 2007), genetic disease (Carter and Nguyen 2011), sexual conflict (Rice 1992; Innocenti and Morrow 2010), cooperation (Foster et al. 2004), evolutionary constraint (He and Zhang 2006), adaptation (Fisher 1930; Orr 2000; Wang et al. 2011), neofunctionalization (Hughes 1994), and speciation (Berlocher and Feder 2002). For instance, a prevailing theory of aging asserts that mutations accumulated during evolution due to their benefits to development and reproduction in early stages of life tend to be deleterious later in life and cause senescence (Williams 1957). AP dictates that a mutation is unlikely to be advantageous to multiple traits or in multiple environments, leading to compromises among adaptations of different traits or in different environments (Fisher 1930). This fundamental property limits the extent and rate of adaptation (Orr 2000) and guarantees that no species would outperform all others in all environments (Levins 1968).

In contrast to the prominent roles of AP in many theories, our empirical knowledge of AP is limited. Early artificial selection experiments showed that improving one trait often worsens another, suggesting that AP is not uncommon (Mather and Harrison 1949; Rice 1992; Cooper and Lenski 2000; Ostrowski et al. 2005). Consistently, a *Drosophila* study proposed over 1000 candidate genes that are subject to sexual antagonism, based on correlations between gene expression levels and organismal fitness across 15 genotypes and two sexes (Innocenti and Morrow 2010). However, because correlation does not imply causation, the actual AP genes remain elusive. As such, neither the prevalence of AP nor the identity of AP genes is known at the whole-genome scale, although individual cases of AP genes have been reported in recent years (Lang et al. 2009; Magwire et al. 2010; Wenger et al. 2011). It is also unclear to what extent AP may be resolved evolutionarily, which genetic mechanisms are mainly responsible for AP resolution, and what population genetic parameters are conducive to AP resolution. We here address these fundamental questions by a combination of genomics, genetics, and modeling, on the basis that, if a gene is subject to AP between two environments, deleting the gene would lower the organismal fitness in one of the environments but improve it in the other.

5.3 Identification of AP genes

To quantify AP at the genomic scale, we took advantage of the yeast gene deletion collection (Giaever et al. 2002), which was constructed by individually knocking out each of the 4642 nonessential genes and 11 pseudogenes from a laboratory strain of *Saccharomyces cerevisiae* and by placing in each deletion strain a 20-nucleotide unique barcode that can be amplified by universal primers. We grew all homozygous deletion strains together and quantified their relative frequencies at multiple time points by amplifying and then sequencing the barcodes using the Illumina-based Bar-seq method (Smith et al. 2009), which provides a large dynamic range and low background noise (Smith et al. 2009) (Figure 5.1A). Bar-seq digitally counts every strain, while the previous microarray-based method (Giaever et al. 2002) does not provide a signal that is linear with the frequency of a strain. Although amplification biases from polymerase-chain-reaction (PCR) may exist in library preparation for Illumina sequencing, the biases would not affect our fitness measurement, because fitness is estimated by comparing the frequency of a strain between two samples obtained at different time points and the PCR biases are cancelled out from between-sample comparisons. Similarly, Illumina sequencing biases (Dohm et al. 2008) do not affect our fitness measurement because the effects are cancelled out when two samples obtained from two time points are compared. It has been reported that Illumina sequencing has a relative high sequencing error (1%, http://www.illumina.com). However, the sequencing errors do not affect our results because any two barcodes differ from each other by at least 5 nucleotides (Shoemaker et al. 1996), beyond what sequencing errors can do. We discarded all sequencing reads that differ from the known barcodes by more than 1 nucleotide.

Fitness was measured in six distinct media that represent a subset of the diverse environments that wild, domesticated, and laboratory yeast strains have experienced, including the rich medium (YPD), glycerol medium (YPG), ethanol medium (YPE), synthetic complete medium (SC), synthetic oak exudate medium (OAK), and rich medium with 6% ethanol (ETH). We estimated the fitness of each deletion strain relative to the wild-type by using the 11 pseudogene deletion strains as 11 biological replicates of the wild-type. By contrast, previous high-throughput fitness quantifications lacked wild-type references and effectively used the weighted average strain in the whole population

as the reference (Giaever et al. 2002; Steinmetz et al. 2002; Deutschbauer et al. 2005; Dudley et al. 2005; Hillenmeyer et al. 2008), which would be problematic for identifying beneficial null alleles, for two reasons. First, because the frequencies of low-fitness strains decrease in competition, the fitness of the weighted average strain increases during competition, which makes fitness estimation unreliable. Second, because there are many low-fitness strains in the population, the average fitness of the population is lower than the fitness of the wild-type. Thus, a strain found to be fitter than the population average in these earlier studies may not be fitter than the wild-type. In our study, we used the 11 pseudogene deletion strains to estimate the standard deviation of our fitness measurement, which allowed us to estimate the probability (*P*-value) that the fitness of a deletion strain equals the wild-type and the corresponding *Q*-value after considering multiple testing.

Under YPD, 62.2% of the nonessential gene deletion strains are not significantly different from the wild-type in fitness ($Q > 0.01$), while 32.6% are significantly less fit ($Q < 0.01$) and 5.1% are significantly fitter ($Q < 0.01$) (Figure 5.1B). Qualitatively similar observations were made in each of the other five media. The number of deletion strains that are significantly fitter than the wild-type varies from 147 to 643 in the six media (Figure 5.1C), with decreasing numbers of strains that are fitter than the wide-type in more media (Figure 5.1C).

The reliability of our fitness estimation is reflected by the high Pearson's correlation between two biological replicates ($r = 0.94$, Figure 5.2A), low false negative rate (only one of 11 previously identified beneficial null alleles (Sliwa and Korona 2005) was not rediscovered here), and small fitness variation among the 11 pseudogene deletion strains in most media (Figure 5.2B). It is also important to estimate the false positive rate, because a number of secondary mutations are known in the yeast gene deletion collection (Hughes et al. 2000) and because beneficial secondary mutations are more likely than deleterious ones to be included in the collection. To gauge the false positive rate, we randomly chose 24 genes whose null alleles are fitter than the wild-type in Bar-seq, independently deleted these genes, and re-measured their fitness by a more accurate low-throughput method (He et al. 2010). We found that 46% of them can be confirmed (Figure 5.2C). Several lines of evidence suggest that most false positives arose from

secondary mutations accumulated in the gene deletion collection rather than Bar-seq errors.

Strictly speaking, AP is inferred when the null allele of a gene is (i) fitter than the wild-type in at least one condition and (ii) less fit than the wild-type in at least one condition. We dropped the second criterion here because it should have been met by all the genes examined; otherwise, the null allele would have been fixed in the species. Indeed, compared with null alleles having similar fitness as the wild-type ($Q > 0.01$) in a medium, those significantly fitter than the wide-type ($Q < 0.01$) in the medium tend to be less fit than the wide-type in other media (Figure 5.2D). Under the first criterion, 1249 AP genes were identified. After considering the false negative and false positive rates, we estimated that $1249\times(11/24)\times(11/10) = 630$ genes, or 13.6% of all nonessential genes examined, are subject to AP. For three reasons, this is likely to be a conservative estimate. First, while AP can potentially occur between any two alleles at a locus, only two specific alleles per locus are compared here. Second, because the number of identified AP genes increases with the number of media examined (Figure 5.2E) and because yeast experiences more than six environments in nature, the actual number of genes subject to AP should be much greater than estimated here. Third, although our fitness measure is more sensitive than all other high-throughput methods, its sensitivity (~0.01) is still lower than that of natural selection, which can detect a fitness differential as small as the inverse of the effective population size (N_e), which is $\sim 10^7$ in yeast (Wagner 2005). Thus, there are likely many more genes than detected here that are subject to milder AP.

5.4 Properties of AP genes

The identified AP genes differ in several aspects from "neutral genes", which have similar fitness between null and wild-type alleles ($Q > 0.01$) in all six media. First, AP genes are less likely to be lost than neutral genes when a diverse panel of 64 strains sampled from different environments are examined (Figure 5.3A), suggesting that overall AP genes are more important and less dispensable than neutral genes. Second, natural selection acting on the coding sequence of a gene can be quantified by the ratio of the number of nonsynonymous substitutions per nonsynonymous site (d_N) to the number of

synonymous substitutions per synonymous site (d_S). We found d_N/d_S to be lower for AP genes than neutral genes when *S. cerevisiae* is compared with its sister species *S. paradoxus* (Figure 5.3B), suggesting stronger purifying selection on the coding sequences of AP genes than those of neutral genes. Third, AP genes show lower expression divergences than neutral genes when the microarray gene expression data from several yeast species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae* , and *S. kudriavzevii*) (Tirosh et al. 2006) are compared (Figure 5.3C). This difference could be due to (i) smaller mutational target sizes and/or (ii) stronger purifying selection on the expressions of AP genes than neutral genes. We found that AP genes have lower expression divergences than neutral genes in a set of mutation accumulation lines (MA) of yeast (Landry et al. 2007) (Figure 5.3D). Because MA lines are subject to virtually no natural selection, the above finding indicates that AP genes have smaller mutational targets for expression changes than neutral genes. Stochastic expression variation among isogenic cells (expression noise) (Newman et al. 2006) reflects the strength of purifying selection (Batada and Hurst 2007; Lehner 2008; Wang and Zhang 2011) and is not influenced by mutation target size. We found that AP genes have smaller expression noise than neutral genes (Figure 5.3E), suggesting stronger purifying selection on expression level in AP genes than neutral genes. Therefore, both smaller mutation target sizes and stronger purifying selection contribute to the lower expression divergences of AP genes than neutral genes. Note that the above observations are valid not only for all AP genes as a whole (black bars in Figure 5.3), but also for AP genes identified in each environment (grey bras in Figure 5.3).

By definition, the expression of an AP gene reduces fitness in some environments. What are the underlying molecular mechanisms of these adverse effects? We found that AP genes are enriched or deprived in a number of Gene Ontology categories. For instance, compared to all the genes in the genome, genes with a null allele fitter than the wild-type allele under ETH are enriched in six GO categories, after controlling for multiple hypothesis-testing. These six GO categories can be further divided into three groups: phospholipid transport, ER-associated protein catabolic process, and heterochromatin (Figure 5.4), which appear to be related to the known cellular effects of ethanol. For example, ethanol influences cell membrane integrity (Ingram and Buttke

1984), and ethanol tolerance relies on the phospholipid composition of the cell membrane (Mishra and Prasad 1988). Phospholipid transporters enable directed movements of phospholipids and thus may be harmful under high ethanol concentrations. In addition, ethanol induces the production of endogenous DNA-damaging molecules (Brooks 1997) and interferes with chromatin condensation (Talebi et al. 2011). Thus, expressions of genes related to heterochromatin could be deleterious in the presence of ethanol. Because ethanol metabolism disrupts protein catabolism (Donohue 2009), expressions of genes involved in protein catabolism could be harmful in ETH. While the exact molecular mechanisms of specific AP await future detailed studies, the enriched and deprived GO categories offer insights for such studies. Complementary to most previous studies that provided lists of genes that are vital to specific traits or biological processes, our study provides lists of genes that are detrimental to these traits or processes. Such information is important for a complete understanding of the mechanisms underlying these traits or processes.

5.5 Evolutionary resolution of AP

In theory, AP between a functional allele and a null allele of a gene can be resolved by lowering the expression of the functional allele in the environment where it is harmful. Two hypotheses may explain the unresolved AP in the laboratory yeast we studied: (i) paucity of regulatory mutations that could resolve AP and (ii) paucity of selection for the fixation of such mutations if the environment concerned is rarely encountered. To distinguish between these hypotheses, we examined four yeast strains that are adapted to their respective ecological niches. The second hypothesis is supported if AP involving the native environment of a strain has been largely resolved; otherwise, the first hypothesis is supported.

We began by confirming our prior knowledge (Warringer et al. 2011) about the adaptations of the four strains to their respective environments (Figure 5.5A) through measuring their relative fitness in four media that approximate the four environments (Figure 5.5B). For instance, the sake strain K12 is expected to have and indeed has the highest fitness in the rich medium with 6% ethanol (ETH) among the four media tested. If AP is resolvable by sufficient natural selection, we can make three predictions about a

gene whose expression is beneficial in environment A but harmful in environment B. First, the expression level of the gene in B should be lower for a strain more adapted to B than for a strain less adapted to B. Second, for a strain adapted to both environments, the expression of the gene should be lower in B than in A. Third, a strain that has adapted to both A and B should have a greater expression difference between these two environments than a strain that is adapted to only one of the environments.

We tested these predictions by quantifying the expression levels of the validated AP genes in Figure 5.2C. For example, *PDR17* encodes a phosphatidylinositol transfer protein that participates in phospholipid synthesis and transport and is involved in resistance to multiple drugs. Its null allele is fitter than the functional allele in YPG but the opposite is true in ETH (Figure 5.5C). We measured the mRNA concentrations of *PDR17* from two strains (M22 and K12) in two media (YPG and ETH). We observed that (i) in YPG, *PDR17* expression is lower for the strain better adapted to YPG (M22) than for the strain less adapted to YPG (K12) (Figure 5.5D); (ii) for M22, *PDR17* expression is lower in YPG than in ETH (Figure 5.5E); and (iii) the expression difference between the two media is greater for the strain adapted to both environments (M22) than the strain adapted to only one environment (K12) (Figure 5.5F). In total, the three predictions are respectively supported by 31 of 35 (Figure 5.5G), 22 of 25 (Figure 5.5H) and 4 of 5 (Figure 5.5I) cases examined.

In addition to transcriptional regulation, we observed protein subcellular relocalization (Komeili and O'Shea 2000) in AP resolution. *MIG1* encodes a transcription factor that works exclusively inside the nucleus in glucose repression (Schuller 2003). Its functional allele is fitter than the null allele in YPD, but the opposite is true in OAK (Figure 5.5J). In the wild strain YPS1000 that is adapted to an environment mimicked by the OAK medium, MIG1 is localized to the nucleus under YPD. But under OAK where MIG1 would be deleterious, MIG1 is localized to the cytoplasm and hence imposes no harm (Figure 5.5K). Together, the findings of many AP-mitigating regulations at the transcriptional or posttranscriptional levels strongly suggest that the unresolved AP in the laboratory strain is largely attributable to a paucity of selection rather than a paucity of mutation, consistent with the recent report that the mutational target size for expression alterations of a gene is substantial (Gruber et al.

2012). Also consistent with this conclusion is the observation that, in the laboratory strain that is adapted to YPD (Figure 5.5A), relatively few null alleles are fitter than the wild-type allele under YPD, compared to other media (Figure 5.1C; 5.10.2).

5.6 Genetic mechanisms of AP resolution

To understand the genetic basis of environment-specific transcriptional regulation that mitigates AP, we investigated whether it occurs by *cis*-acting changes, which act through the same DNA molecule that encodes the focal gene, or *trans*-acting changes that operate via diffusible molecules. We crossed two parental diploid strains (M22 and K12) to make a hybrid strain (M22 ×K12) and used pyrosequencing to measure allele-specific expressions in the hybrid as well as in mixed parents. The expression difference between the two alleles in the hybrid is caused by *cis*-acting changes, while the difference in allele-specific expression ratio (M22/K12) between the hybrid and mixed parents is caused by *trans*-acting changes (Wittkopp et al. 2004).

We examined three AP genes with large environment-specific expression regulation. For *PDR17*, the M22/K12 expression ratio in the hybrid is not significantly different from 1, under either YPG or ETH (Figure 5.6A, $P = 0.50$ and $P = 0.70$, respectively, two-tailed *t* test), suggesting the lack of *cis*-acting differences between the two strains. Consistent with the results in Figure 5.5F, the M22/K12 expression ratio is significantly below 1 in mixed parents under YPG ($P = 0.01$) but not under ETH ($P = 0.44$, Figure 5.6A). Thus, the YPG-specific *PDR17* expression divergence between M22 and K12 is primarily caused by *trans*-acting changes. A similar conclusion can be made for the second examined gene, *APQ12* (Figure 5.6B).

The third gene studied (*STP4*) showed a different mechanism. *STP4* encodes a transcription factor involved in multiple cellular processes and drug resistance. The null allele is fitter than the functional allele in YPG, but this relation is reversed in ETH (Figure 5.6C). We found the M22/K12 expression ratio of *STP4* in YPG to be lower than 1 by a similar amount in mixed parents and the hybrid ($P = 0.97$, two-tailed *t* test, Figure 5.6D), indicating that the *STP4* expression divergence between M22 and K12 in YPG is mainly caused by *cis*-acting changes. We suspected that a 250-nucleotide promoter region of *STP4* that harbors four single nucleotide differences between the two strains is

responsible for the expression divergence between them in YPD. To test this hypothesis, we swapped this region between the two strains in haploid cells. Indeed, *STP4* expression in K12 is reduced to the M22 level when its promoter is replaced with that of M22 (Figure 5.6E), suggesting that one or more of the four nucleotide mutations caused the expression difference between M22 and K12. Interestingly, *STP4* expression in M22 is not enhanced by using the K12 promoter (P = 0.84, Figure 5.6E), demonstrating a genetic background-specific effect of these regulatory changes.

The above regulatory mutations that are beneficial to M22 under YPG may be harmful under ETH, because *cis*-acting changes tend to affect gene expression in multiple conditions (Smith and Kruglyak 2008). Indeed, the M22 allele has a lower expression compared to the K12 allele in the hybrid under ETH (P = 0.001, Figure 5.6D), and replacing the native promoter with the M22 promoter in K12 lowers *STP4* expression in ETH as in YPG (P = 0.05, Figure 5.6E). Nevertheless, this deleterious *cis* effect in ETH is compensated by *trans*-acting changes, evident from the comparison of the M22/K12 expression ratio in mixed parents and the hybrid (P = 0.0003, Figure 5.6D). Together, *trans*-acting changes were found in all three examined cases of AP resolution, while only one case involves an additional *cis*-acting change.

5.7 Population genetics of AP resolution

Our observation that most AP is resolvable at least partially yet AP is still present in many genes in the laboratory strain prompts us to investigate the population genetic parameters conducive to AP resolution. Specifically, we formulated the expected waiting time for an AP-alleviating mutation destined for fixation to appear in a population (i.e., time to mutation T_m) and the expected time from the appearance to the fixation of this mutation (i.e., time to fixation T_f). The expected total waiting time for the appearance and fixation of the first AP-alleviating mutation is $T = T_m + T_f$. We assume that, relative to the wild-type, the mutant has a selective advantage of s in environment B but 0 is environment A and that the population spends a fraction (f) of its time in B and the rest of time in A. It can be shown that the equivalent selection coefficient $s_e = sf$ (5.10.5). We considered two additional parameters: N_e and the equivalent number of nucleotide sites at which all point mutations alleviate AP (i.e., mutation target size L). Mutation rate per

site per generation (u) is relatively constant among cellular organisms and the estimate from yeast (3×10^{-10}) is used here (Lynch et al. 2008). A larger u has the same effect as a larger L, as uL is what matters. For yeast, T_m/T_f <1 when L >2.5 (Figure 5.7A), indicating that, when the mutation target size is not very small, time to AP resolution is primarily determined by the time to fixation rather than time to mutation. But in species with smaller N_e, the situation is easily reversed (Figure 5.7A). While T_m/T_f is independent of s_e, T decreases with rising s_e (Figure 5.7B). For yeast, depending on its generation time g in nature, time to AP resolution (gT) varies from 1 to 10,000 years (Figure 5.7C). For example, when sf = 0.001, L = 4 nucleotides, and g =16 hours, gT is ~100 years. It is possible that non-repetitive environmental changes are so frequent that a yeast population has yet to fix an AP-resolving allele before the specific environment vanishes. AP would be hard to resolve under this scenario.

5.8 Discussion

By measuring the fitness effects of null mutations in almost all yeast nonessential genes under six different environments, we offered the first whole-genome-scale quantification of AP in any species. Although our AP quantification was performed in a laboratory strain of yeast, we believe that the conclusion extends to wild strains, because the laboratory strain, similar to many wild strains, experiences multiple different environments and because the six media under which AP was surveyed are routine laboratorial media. This view is supported by the observation that, even in YPD, the most frequently used medium for culturing the laboratory strain, there are over 200 AP genes (114 after correcting for false negatives and false positives), and this number is likely a gross underestimate, as mentioned earlier. Although the specific genes subject to AP may vary among strains due to the different environments that different strains encounter, AP is probably more frequent in wild strains than in laboratory strains, because the number of environmental variables in the wild is likely greater than that commonly applied in the laboratory. Our finding that AP is often resolvable in strains well adapted to certain environmental factors (e.g., a high ethanol concentration) only means that AP related to this factor is resolved in these strains. But, they can and will have unresolved AP related to other environmental factors to which the strains are not

well adapted (e.g., ambient temperature that varies both deterministically and stochastically). As long as the environment is not constant, AP is expected to exist. Our finding that at any condition yeast expresses hundreds of genes that are harmful rather than advantageous to the organism demonstrates the prevalence of AP and the importance of considering AP in understanding biology.

For two reasons, AP is expected to be even more abundant in multicellular organisms than in yeast. First, while our yeast study focuses exclusively on AP among different external environments, multicellular organisms are subject to additional types of AP. For example, some alleles advantageous to one sex are known to be harmful to the other sex in *Drosophila* (Innocenti and Morrow 2010). In humans, mutations causing Huntington's disease, a neurodegenerative disorder in which symptoms typically manifest after the reproductive age, is known to i30-45ncrease fecundity (Carter and Nguyen 2011). The existence of sexes, tissues, and life stages in complex multicellular organisms creates a greater potential for AP. Second, our population genetic analysis showed that it takes longer to resolve AP when the effective population size is smaller or when the generation time is longer. Because multicellular organisms have much smaller effective population sizes and much longer generation times than yeast (Lynch 2007), the fraction of AP that is unresolved is expected to be much greater in the former than the latter.

Because AP is invoked in current explanations and models of senescence (Williams 1957), cancer (Rodier et al. 2007), genetic disease (Carter and Nguyen 2011), sexual conflict (Rice 1992; Innocenti and Morrow 2010), cooperation (Foster et al. 2004), evolutionary constraint (He and Zhang 2006), adaptation (Fisher 1930; Orr 2000; Wang et al. 2011), neofunctionalization (Hughes 1994), and speciation (Berlocher and Feder 2002), our finding of prevent AP provides an empirical foundation for these theories and has profound implications for many areas of biology. In particular, if many disease-causing mutations are kept in the population because of their unexpected benefits in other aspects of life (e.g., development, fecundity, and host defense), as has been suggested in Huntington's disease, cystic fibrosis, sickle-cell anemia, glucose-6-phosphate dehydrogenase deficiency, cancer, and many others (Carter and Nguyen 2011), special cautions would be needed in treating these diseases, because a treatment may lead to

adverse effects in other aspects of life. On the other hand, discerning the underlying mechanisms of AP in such diseases could lead to improvement of certain traits such as host defense. This so-called positive biology (Farrelly 2012) complements the common practice of focusing exclusively on diseases in biomedical research. The identified natural solutions to AP may also guide designs of synthetic genomes and organisms (Gibson et al. 2010) that need to perform well in multiple environments. That is, when introducing a gene into a host genome, one should examine the effect of the introduction in multiple environments, sexes, tissues, and life stages, because a gene beneficial in one condition can be deleterious in another. To optimize the function of the synthetic organism, a well-designed expression regulation network is required to suppress the expression when it is harmful and to activate the expression when it is advantageous. We hope that our first genome-scale quantification of AP will stimulate further studies in this area of universally recognized importance that has thus far been largely untouched by systematic empirical analysis.

5.9 Experimental procedures

5.9.1 Fitness measurement by Bar-seq

The yeast single gene deletion collection (Giaever et al. 2002) was purchased from Invitrogen (Cat. No. 95401.H1Pool). The strains are diploid, with a homozygous deletion of a nonessential gene per strain. In the process of gene deletion, a unique 20-nucleotide DNA sequence (barcode) was inserted into each strain. The yeast strains were competed in six media, including YPD (1% yeast extract, 2% peptone, and 2% glucose), YPG (1% yeast extract, 2% peptone, and 5% glycerol), YPE (1% yeast extract, 2% peptone, and 2% ethanol), SC (0.17% yeast nitrogen base without amino acid and ammonia sulfate, 0.5% ammonia sulfate, 2% glucose, and 0.2% dropout mix), OAK (0.1% yeast extract, 0.15% peptone, 0.5% glucose, 0.5% fructose, and 1% sucrose) (Murphy et al. 2006) and ETH (1% yeast extract, 2% peptone, 2% glucose, and 6% ethanol).

The mixed strains from the deletion collection were first cultured for two generations. The resultant sample was named generation 0. We used a fraction of the sample from generation 0 to seed a 25 ml competition culture, which was grown at 30℃ with shaking at the speed of 250 revolutions per minute. To reduce the impact of genetic

drift, we maintained a relatively large population. The competition culture was diluted 10 fold with fresh medium every 3-4 generations. The frequency of each strain was measured at generation 0, generation 3, and generation 26 by the barcode sequencing (Bar-seq) method (Smith et al. 2009). To perform Bar-seq, we extracted the genomic DNAs from each yeast population, amplified the barcodes by polymerase chain reaction (PCR), and PCR-added sequences recognizable by Illumina sequencing primers. By design, each deletion strain has an upstream barcode and a downstream barcode (Giaever et al. 2002). We used only the upstream barcode, because the downstream barcode is known to be missing in some strains (Deutschbauer et al. 2005). We sequenced 40 nucleotides from one end of each PCR amplicon using one lane on an Illumina Genome Analyzer IIx at the University of Michigan DNA Sequencing Core. The Illumina Pipeline software version 1.5 was used for base calling from the image data. Because all the sequences started with the same 18 base pairs of the PCR primer region and this uniformity adversely affected base calling, we removed the first 18 sequencing cycles before base calling. The sequencing reads have been submitted to NCBI.

To estimate strain frequencies, we need to know which gene deletion strain uses which barcode (i.e., the gene-barcode map). We retrieved two different versions of the map from Deutschbauer et al. (Deutschbauer et al. 2005) and Smith et al. (Smith et al. 2009), respectively. We found that 4894 gene-barcode relations are consistent between the two maps. However, for 1226 genes, different barcode sequences were shown in the two maps. We decided which map is correct for each of these genes, using the actual read sequences we acquired from the generation 0 population (allowing one sequencing error). Although barcodes were designed to be different from one another by at least 5 nucleotides (Shoemaker et al. 1996), because of DNA synthesis errors, some barcodes are no longer distinguishable after allowing one sequencing error. For this reason, YKL137W and YGL140C were removed from the gene-barcode map. Based on our gene-barcode map, we mapped sequencing reads to gene deletion strains, allowing one sequencing error per read.

The fitness of a strain, relative to the wild-type, is calculated by $w = e^{\ln\left(\frac{P'P_{wt}}{PP'_{wt}}\right)^{1/t}}$, where P, P', P_{WT}, and P'_{WT} are the frequencies of the strain at the beginning of the

competition, the strain at the end of the competition, wild-type at the beginning of the competition, and wild-type at the end of the competition, respectively, and *t* is the number of generations in the competition. To guarantee high accuracy of fitness measurement, strains with fewer than 100 reads in generation 0 were not considered. We quantified the frequency of each strain at the 0^{th}, 3^{rd}, and 26^{th} generation. Thus, for each strain, we could calculate fitness based on frequency changes between generation 0 and 3 or between generation 0 and 26. The fitness is measured more accurately with more generations of competition. However, if the fitness of a strain was so low that the stain disappeared in the 26^{th} generation, we calculated the fitness based on the frequency change between generation 0 and 3; otherwise, the frequency change between generation 0 and 26 was used. Note that sequencing bias does not affect our results.

To estimate the fitness of a gene deletion strain relative of the wild-type, we have to define a wild-type strain that has the same genetic background as the deletion strains, including containing a barcode and the same marker gene as in the deletion strains. *HO*, a site-specific endonuclease gene necessary for mating type conversion, suffers from several severe mutations and is nonfunctional in the strain used for constructing the gene deletion collection (Meiron et al. 1995; Ekino et al. 1999). Thus, the *HO* deletion strain is an ideal wild-type reference. In the deletion collection, there are 10 additional strains in which a pseudogene that lacks an intact open reading frame (ORF) is deleted. We first used the *HO* deletion strain as the wild-type reference to estimate the relative fitness of the 10 additional pseudogene deletion strains. We then calculated the mean fitness of the 11 strains. As expected, the mean fitness is not significantly different from 1 in any medium. We thus merged the reads of all 11 strains and considered them collectively as the wild-type reference. Using this reference, we calculated the fitness of every deletion strain, including the 11 pseudogene deletion strains. To determine whether the fitness of a deletion strain differs significantly from 1, we conducted a *Z*-test using the fitness values of the 11 pseudogene deletion strains as the null distribution. The *P*-values from the *Z*-test were further converted to *Q*-values after the consideration of multiple testing (Storey 2002).

5.9.2 Identification of AP genes

We first used the *HO* deletion strain as the wild-type reference (Meiron et al. 1995; Ekino et al. 1999) to estimate the relative fitness of the 10 additional pseudogene deletion strains. We then calculated the mean fitness of the 11 strains. As expected, the mean fitness is not significantly different from 1 in any medium. We thus merged the reads of all 11 strains and considered them collectively as the wild-type reference. Using this reference, we calculated the fitness of every deletion strain, including the 11 pseudogene deletion strains. To determine whether the fitness of a deletion strain differs significantly from 1, we conducted a *Z*-test using the fitness values of the 11 pseudogene deletion strains as the null distribution. The *P*-values from the *Z*-test were further converted to *Q*-values after the consideration of multiple testing (Storey 2002).

5.9.3 Analysis of the properties of AP genes

Gene loss rates were estimated from 64 strains previously genotyped (Schacherer et al. 2009) (ftp://gen-ftp.princeton.edu/yeast_snps/schacherer2008/all_del.gff). DNA sequences of ORFs in *S. cerevisiae* and *S. paradoxus* were downloaded from (Kellis et al. 2003) (http://www.broadinstitute.org/annotation/fungi/comp_yeasts/downloads.html), and d_N/d_S ratios were calculated using PAML (Yang 2007). We obtained gene expression divergences among species from (Tirosh et al. 2006), expression divergence among mutation accumulation lines from (Landry et al. 2007), and expression noise in *S. cerevisiae* from (Newman et al. 2006). Expression noise is measured by *DM*, which allows the comparison of noise levels among genes with different mean expression levels (Newman et al. 2006). We used GOstats (version 2.20.0) (Falcon and Gentleman 2007) in Bioconductor (http://www.bioconductor.org/) for GO analysis. The GO category enrichment was illustrated with Cytoscape (Shannon et al. 2003).

5.9.4 Relative fitness of four yeast strains in four media

The non-laboratory strains used in our study were described previously (Liti et al. 2009; Schacherer et al. 2009). By competition with a yellow-fluorescent-protein (YFP)-marked reference strain followed with cell counting using flow cytometry (He et al. 2010), we measured the relative fitness ($f_{i,j}$) of each of four strains (i = 1 to 4) in each of

four media (j = 1 to 4). We then calculated the mean fitness of each strain in the four media (g_i) and the mean fitness of the four strains in each medium (h_j). The relative fitness of each strain in each medium was estimated by $f_{i,j}/g_i/h_j$.

5.9.5 Strain constructions

To validate the Bar-seq results, we independently deleted 24 genes in BY4742 (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) (Brachmann et al. 1998), which originated from the lab strain S288C (*MATα SUC2 gal2 mal mel flo1 flo8-1 hap1 ho bio1 bio6*, Saccharomyces Genome Database, SGD, www.yeastgenome.org). The strains were constructed by replacing the ORF of each gene by the auxotroph marker *URA3*, following our previous paper (He et al. 2010). We also constructed 10 *HO* deletion strains by replacing *HO* with *URA3*; these strains have the same genetic background as the 24 gene deletion strains and were regarded as the wild-type in the low-throughput fitness measurement of the 24 deletion strains. The fitness of these 24 strains was measured using a previously published method that can detect a fitness differential of 0.5% (He et al. 2010). Briefly, the fitness of a strain was measured through competition with an YFP-marked strain, followed by cell counting using flow cytometry.

To tag MIG1 with GFP in the wild strain YPS1000, we PCR-amplified the *GFP* coding region together with *kanMX6* from plasmid pFA6a-GFP(S65T)-kanMX6 (Wach et al. 1997) (TaKaRa Ex Taq Polymerase). The PCR products were transformed into YPS1000 **a** cells. After overnight growth in YPD liquid medium, the transformants were spread on YPD plates supplemented with 200 μg/ml geneticin (Invitrogen). We picked single colonies, cultivated them in YPD liquid medium, and extracted genomic DNA (Phenol/Chloroform/Isoamyl Alcohol 25:24:1 Mixture, pH 6.7). We confirmed GFP tagging by PCR

The non-lab strains used in our study were described previously (Liti et al. 2009; Schacherer et al. 2009). For strains K12 and M22, we sporulated the diploid cells, dissected the tetrads, and replaced the *HO* locus by kanMX and hygMX in *α* and **a** haploid cells, respectively, to avoid mating-type switch. To obtain the hybrid strain M22×K12, we crossed M22 **a** cells with K12 *α* cells by mixing them on YPD plates for 24 hours. The cells were further streaked on YPD plates supplemented with 200 μg/ml

geneticin and 300 μg/ml hygromycin B (Invitrogen). We picked single colonies and used PCR amplification to confirm the existence of both *MAT***a** and *MATα* (Huxley et al. 1990).

To obtain *STP4* promoter (P_{STP4}) swapped strains in K12 and M22 backgrounds, we first replaced P_{STP4} (600 bp upstream of the start codon) of both strains with KanMX4, which was amplified from pFA6a-kanMX4 (Wach et al. 1997). Single colonies were picked from YPD plates supplemented with 200 μg/ml geneticin. Strains K12 P_{STP4}*::kanMX* and M22 P_{STP4}*::kanMX* were confirmed by PCR. The 250 bp upstream of the start codon of *STP4* were amplified from K12 and M22 and named K12_12 and M22_12, respectively. The 251-600 bp upstream of the start codon of *STP4* were amplified from K12 and M22 and named K12_34 and M22_34, respectively. NatMX4 was PCR amplified from pAG25 (Goldstein and McCusker 1999) and named nat_56. We PCR fused K12_34 with nat_56 and obtained K12_78. We similarly fused M22_34 with nat_56 to get M22_78. We further PCR fused K12_12 or M22_12 with K12_78 or M22_ to obtain fusion_KK, fusion_KM, fusion_MK, and fusion_MM, respectively. The first letter (K or M) in these names represents the template K12_12 or M22_12, and the second letter (K or M) represents the template K12_78 or M22_78. Fusion_KK and fusion_MK was transformed into K12 P_{STP4}*::kanMX* to obtain strains pK-K12 and pM-K12, respectively. Fusion_KM and fusion_MM were transformed into M12 P_{STP4}*::kanMX* to obtain strains pK-M22 and pM-M22, respectively. Single colonies were picked from YPD plates supplemented with 100 μg/ml nourseothricin. After PCR confirmation, the strains were further verified by DNA sequencing.

5.9.6 Microscopy

Yeast cells were grown in YPD or OAK overnight at 30℃ to the stationary phase. Optical density (OD) of yeast culture was measured at 660 nm with spectrophotometer (GENESYS 5, Thermo Scientific). The yeast culture was diluted to OD_{660} = 0.1 by fresh medium supplemented with 1μg/ml 4',6-diamidino-2-phenylindole (DAPI, Sigma) and was harvested when OD_{660} reached 0.5 (mid-log phase). Yeast cells were washed, condensed, and examined under fluorescence microscopy (DeltaVision Spectris microscope, Applied Precision).

5.9.7 Expression measurement by quantitative PCR (qPCR)

Yeast cells were grown overnight at 30℃ to the stationary phase. The culture was then diluted to OD_{660} = 0.1 and re-grown until OD_{660} = 0.5 (mid-log phase). We extracted the total RNA (RiboPure-Yeast Kit, Ambion) from three independently grown cultures of each genotype. The total RNA was quantified by nanodrop (Thermo Scientific), and 1μg total RNA was reversely transcribed into complimentary DNA (Moloney Murine Leukemia Virus Reverse Transcriptase, Invitrogen) with random hexamer primers. The mRNA levels were measured by qPCR (Power SYBR green PCR mater mix and 7300 Real-Time PCR System, Applied Biosystems) with *ACT1* as an internal control.

5.9.8 Determination of allele-specific gene expression by pyrosequencing

Diploid K12, M22, and M22×K12 strains were individually cultivated in YPG and ETH to mid-log phase. We mixed K12 and M22 in an approximate 1:1 ratio based on optical density. The mixed sample was separated equally into two tubes after vortexing. One was used for RNA extraction and the other DNA extraction. RNA samples were further reversely transcribed. We then performed pyrosequencing following published protocols (Wittkopp et al. 2004). Briefly, appropriate single nucleotide polymorphisms (SNPs) between K12 and M22 were selected, DNA regions surrounding the SNPs were PCR amplified, single strand DNA was isolated by streptavidin sepharose (GE Healthcare), and DNA was sequenced by synthesis (PyroMark Q96 ID System, Qiagen). The expression ratio of a gene from K12 and M22 in the mixed sample was calculated by the pyrosequencing signal ratio from the RNA sample divided by that from the DNA sample. The expression ratio between the K12 and M22 alleles in the hybrid was similarly determined, except that the K12+M22 mixture was replaced with the M22×K12 hybrid.

5.10 Appendices

5.10.1 Precision of fitness measurement by Bar-seq

On average, 22 million reads per sample were mapped, which is equivalent to a mean sequencing depth of 4400 reads per strain. The coefficient of variation in

frequency estimation is $\frac{\sqrt{4400}}{4400}$ =1.5% on average. For all deletion strains fitter than the wild-type, the fitness was estimated from a 26-generation competition. We can estimate the standard deviation of the fitness measurement by computer simulation, based on the assumption that the expected number of reads mapped to a wild-type strain is 4400. We randomly generated read numbers for a wild-type strain and a deletion strain with a relative fitness of 1, following a binomial distribution. We then calculated the fitness of the deletion strain, relative to the wild-type, based on the number of reads observed in the simulation. Based on 1000 simulation replications, the standard deviation of the fitness estimate of the deletion strain is 0.001. Thus, in theory, a fitness differential as small as 0.0026 can be detected with 99% confidence. We discarded strains with initial frequencies lower than 100 reads, because fitness cannot be accurately estimated for these strains. The standard deviation of fitness for a strain with 100 reads at the beginning of the competition could be as high as 0.005 (detectable fitness differential is 0.013 with 99% confidence).

5.10.2 Number of AP genes in each medium

The number of AP genes is the smallest in the OAK medium. This is likely due to the small number of Illumina reads obtained for the generation-26 sample from OAK, which reduces the statistical power for detecting AP genes. For instance, the 11 pseudogene deletion strains show substantially greater fitness variation in OAK than in any other medium (Figure 5.2B). When OAK is disregarded, the number of AP genes in YPD is smaller than that in any other medium and is only 47 percent the average number in the other media.

5.10.3 Sources of false positive errors in identifying AP genes by Bar-seq

To estimate the false positive rate, we randomly selected 24 genes with null allele showing significantly higher fitness than the wild-type in an environment. We independently deleted these genes and measured the fitness of the deletion strains in the same environment using a previously established low-throughput method (FACS) (He et al. 2010). We found that 46% (11/24) of the AP genes identified from Bar-seq can be

confirmed (Figure 5.2C). False positives can be from either (i) secondary mutations accumulated in the gene deletion collection or (ii) fitness measurement errors in Bar-seq. For three reasons, (i) is more likely than (ii). First, a number of secondary mutations have been reported in the yeast gene deletion collection. For example, it has been reported that 8% of the collection exhibit aneuploidy (Hughes et al. 2000). Second, a strong positive correlation exists between Bar-seq based fitness and FACS based fitness for the confirmed AP genes (Figure 5.2C), while such a correlation is lacking for unconfirmed AP genes (Figure 5.2C). Third, under (ii), high fitness strains in Bar-seq should have higher probabilities to be confirmed. However, we found that the confirmed AP genes and unconfirmed AP genes have similar Bar-seq fitness (two-tailed *t* test, $P = 0.85$, two-tailed Mann-Whitney *U* test, $P = 0.49$).

5.10.4 Probability of existence of an open reading frame (ORF) under no selection

The average ORF length is 450 codons in yeast (SGD, http://www.yeastgenome.org/). The probability that a random DNA sequence of this length is an ORF is $(61/64)^{450} = 4.1 \times 10^{-10}$. The yeast genome has $\sim 1.2 \times 10^{6}$ nucleotides and there are two strands. Thus, the probability of having an average-length ORF in the yeast genome simply by chance is $4.1 \times 10^{-10} \times 1.2 \times 10^{6} \times 2 = 0.001$. In other words, ORFs found in genome sequences are most likely maintained by purifying selection in at least one environment. Otherwise, the null allele would be fixed and the functional gene would get lost. When a null allele is fitter than the functional allele in at least one environment and is as fit as the functional allele in the rest of the six examined environments, there is a small probability that the functional allele is not fitter than the null allele in any environment. In such a case, the null allele will quickly replace the functional allele, known as adaptive gene loss (Wang et al. 2006). Because the probability of observing such an event is low, we here ignore this possibility (Wang et al. 2006). This consideration allows the removal of the criterion (ii) in defining AP and is supported by several lines of evidence presented in the main text.

5.10.5 Effective selection coefficient

When the selection coefficient of a mutant, relative to the wild-type, varies among environments, it can be shown that the effective selection coefficient (s_e) is

$$s_e = \sum_{i=1}^{E} (f_i s_i), \qquad [1]$$

where s_i is the selection coefficient in environment i, f_i is the fraction of time the organism lives in environment i, and E is the total number of environments of the organism. We first prove it in haploid organisms and then prove it in diploid sexual organisms.

In haploid organisms, selection coefficient s is defined by

$$\frac{q_T}{p_T} = \frac{q_{T-1}}{p_{T-1}} \times (1+s), \qquad [2]$$

where p_T is the frequency of the wild-type allele A in generation T, q_T =1-p_T is the frequency of the beneficial mutant allele B in generation T, and s is the selection coefficient of allele B relative to allele A. If the environment is constant, we have

$$\frac{q_T}{p_T} = \frac{q_0}{p_0} \times (1+s)^T, \qquad [3]$$

or

$$\ln\left(\frac{q_T}{p_T} \Big/ \frac{q_0}{p_0}\right) = T \ln(1+s). \qquad [4]$$

When $s << 1$, $\ln(1+s) \approx s$. Therefore, we have

$$\ln\left(\frac{q_T}{p_T} \Big/ \frac{q_0}{p_0}\right) \approx Ts. \qquad [5]$$

It can be easily seen from [3] that, when there are multiple environments,

$$\frac{q_T}{p_T} = \frac{q_0}{p_0} \times \prod_{i}^{E} (1+s_i)^{Tf_i}. \qquad [6]$$

When s <<1, [6] can be rewritten as

$$\ln\left(\frac{q_T}{p_T} \Big/ \frac{q_0}{p_0}\right) \approx T \sum_{i=1}^{E} (f_i s_i). \qquad [7]$$

By comparing [5] and [7], one can see that the effective selection coefficient when there are multiple environments is indeed described by [1].

In diploid sexual organisms under genic selection, the fitness values are 1, 1+*s*, and 1+2*s* for homozygous wild-type (AA), heterozygote (AB), and homozygous mutant (BB), respectively. When $s << 1$, the fitness of BB is approximately $(1+s)^2$. We have

$$\begin{aligned} p_T &= \frac{2p_{T-1}^2 + 2p_{T-1}q_{T-1}(1+s)}{2[p_{T-1}^2 + 2p_{T-1}q_{T-1}(1+s) + q_{T-1}^2(1+2s)]} \\ &\approx \frac{2p_{T-1}^2 + 2p_{T-1}q_{T-1}(1+s)}{2[p_{T-1}^2 + 2p_{T-1}q_{T-1}(1+s) + q_{T-1}^2(1+s)^2]} \\ &= \frac{2p_{T-1}[p_{T-1} + q_{T-1}(1+s)]}{2[p_{T-1} + q_{T-1}(1+s)]^2} \\ &= \frac{p_{T-1}}{p_{T-1} + q_{T-1}(1+s)} = \frac{p_{T-1}}{1 + q_{T-1}s} \end{aligned} \quad . \qquad [8]$$

Thus,

$$\frac{q_T}{p_T} = \frac{1-p_T}{p_T} = [1 - \frac{p_{T-1}}{1+sq_{T-1}}] / (\frac{p_{T-1}}{1+sq_{T-1}}) = \frac{q_{T-1}}{p_{T-1}}(1+s). \qquad [9]$$

It can be seen that [9] has the same form as [2]. Hence, the formulae for haploid organisms apply to diploid sexual organisms.

5.10.6 Fixation of AP-alleviating alleles

Let N_e be the effective population size of a diploid population. It is known (Kimura 1983) that a newly arisen AP-alleviating allele has a fixation probability (P) of ~$2s_e$, when $N_e s_e > 1$. The time to fixation (T_f) in a diploid population under strong selection ($N_e s_e >> 1$) is

$$\ln(\frac{2N_e - 1}{1} \Big/ \frac{1}{2N_e - 1}) \approx T_f s_e , \qquad [10]$$

because the allele initially has one copy and can be considered to be fixed when it has $2N_e$-1 copies. From [10], we can see that, when $N_e >> 1$,

$$T_f \approx \frac{2\ln(2N_e)}{s_e} \qquad [11]$$

In other words, T_f is proportional to $\ln(2N_e)$, and increases by ~2.3 fold when N_e increases by 10 fold.

The total waiting time (T) for the fixation of the first AP-alleviating allele is the sum of the waiting time (T_m) for the occurrence of the beneficial mutation that is destined

for fixation and the time required for this mutation to be fixed (T_f). The expected value of T_m is

$$T_m = \frac{1}{2N_eLuP} = \frac{1}{4N_euLs_e}, \quad [12]$$

where L is the equivalent target size in nucleotides for beneficial regulatory mutations that alleviate AP and u is the DNA point mutation rate per nucleotide per generation. Because T_f can be calculated by equation [11], we have

$$T = T_m + T_f = \frac{1}{4N_euLs_e} + \frac{2\ln(2N_e)}{s_e} \quad [13]$$

and

$$T_m / T_f = \frac{1}{8uLN_e \ln(2N_e)}. \quad [14]$$

In yeast, $N_e \sim 10^7$ and $u \sim 3\times10^{-10}$ per site per generation (Lynch et al. 2008). Thus, $T_m / T_f = 2.48 / L$. In other words, as long as $L >> 2.5$, which should not be rare, T is mainly determined by T_f rather than T_m.

Under weak selection (i.e., N_es_e is similar to or smaller than 1), due to drift, T_f will be greater than that given by [11] and will eventually approaches $4N_e$ when $N_es_e << 1$. In such cases, T_m given by [6] tends to be much greater than T_f. Thus, using [13] underestimates T only slightly. We thus continue to use [13] to estimate T even when N_es_e is similar to or smaller than 1. It can be seen from [13] that T decreases with rising s_e, L, and u. It can be shown that when $8N_euL<1$, which is almost always true except when L is very large, T decreases with rising N_e.

5.11 Acknowledgments

Di Ma and Che Xiao helped with experiments; Zhi Wang helped with computational analysis. We thank Arielle Cooley, Calum Maclean, Kai Mao, Brian Metzger, and the University of Michigan Sequencing Core for technical assistance and Xiaoshu Chen, Zhenglong Gu, Calum Maclean, and Jian-Rong Yang for comments. This work was partially supported by Block grant from the Department of Ecology and Evolutionary Biology at University of Michigan.

Figure 5.1 Genome-wide identification of yeast genes subject to antagonistic pleiotropy (AP) among six environments.

(A) High-throughput fitness estimation. All ~5000 nonessential gene deletion strains were grown together in one of six different media. Fitness was estimated from strain frequencies quantified by Bar-seq at the beginning (0th generation) and end (3rd or 26th generation) of each competition. Each color depicts one yeast genotype. (B) The fitness spectrum of gene deletion strains, relative to the wild-type (WT), in YPD. "$>$", significantly fitter; "$<$", significantly less fit; "$\approx$", fitness not significantly different. (C) Genes with null allele fitter than WT allele in at least one medium. Each row represents a gene and each column represents a medium. The color scheme is the same as in (B). The number of genes whose null alleles are significantly fitter than WT in each of the six media is shown in the parentheses following the medium. The numbers of genes whose null alleles are significantly fitter than WT in N = 1, 2, ⋯, and 6 media are indicated in the parentheses below the N values.

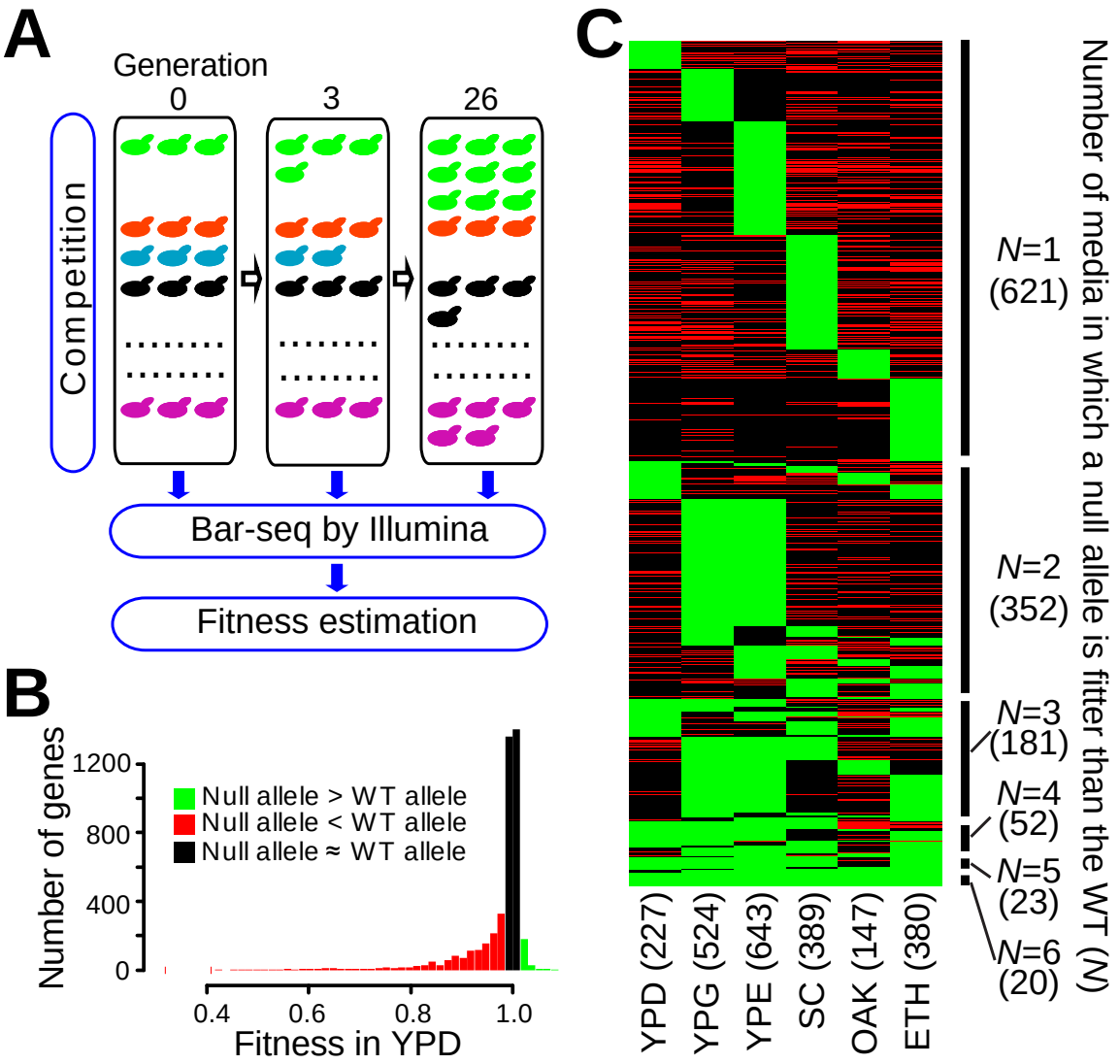


Figure 5.2 Validations of the Bar-seq results.

(A) Fitness estimates in YPD are highly correlated between two biological repeats. Pearson's correlation coefficient $r = 0.94$ ($P < 10^{-4899}$). (B) The fitness of 11 pseudogene deletion strains in six media. The unusually high variation under OAK is caused by a low number of sequencing reads obtained. (C) Fitness values of 24 randomly chosen AP gene deletion strains estimated by Bar-seq and those of their corresponding independently generated deletion strains measured by a more accurate low-throughput method (FACS). Those confirmed by FACS to be subject to AP (blue dots) show highly correlated fitness estimates by two methods ($y = 1.0819\ x - 0.0863$, $r = 0.95$, $P = 6.6 \times 10^{-6}$), while the unconfirmed (red triangles) show no correlation ($y = 0.0003\ x + 1.0004$, $r = 0.002$, $P = 0.97$). Error bars represent one standard error. (D) Compared with gene deletion strains equally fit as the WT in a medium (open bar), those significantly fitter (grey bar) are more likely to be less fit than the WT in other media. Error bars show one standard error. Statistical significance (P-value) is estimated by chi-square test. (E) Number of observed AP genes ($Q < 0.01$) increases with the number of media tested.

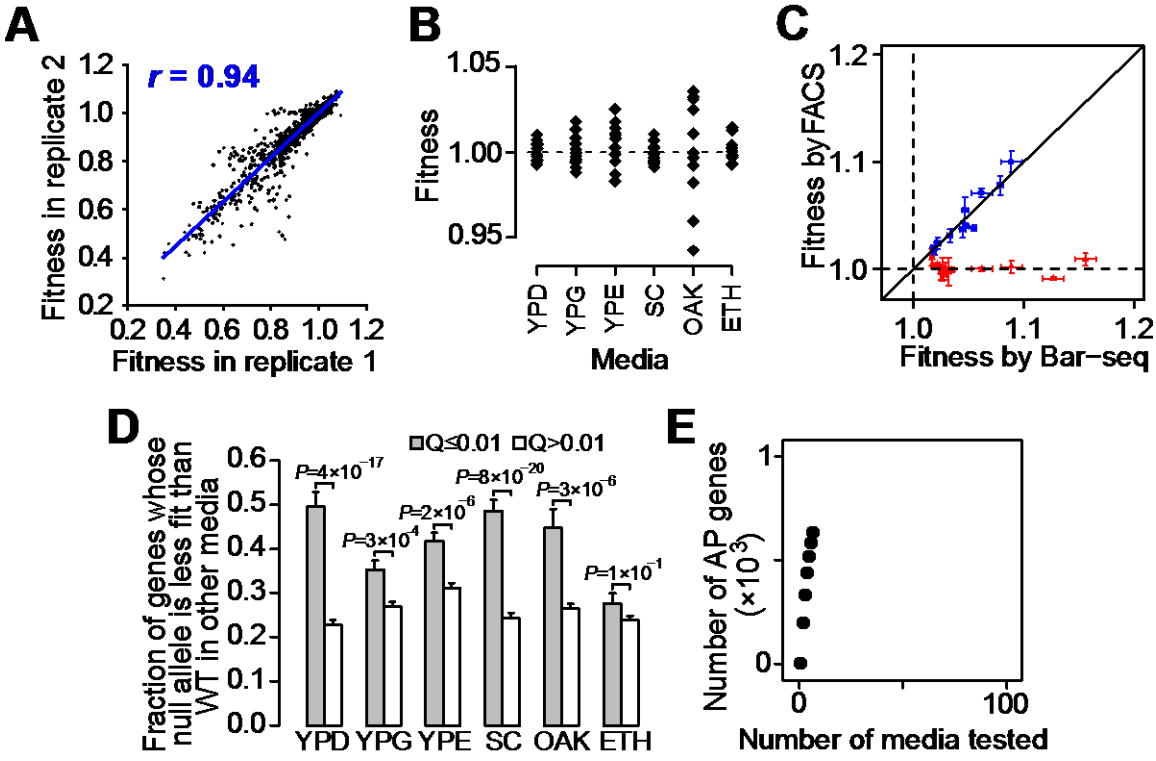


Figure 5.3 Properties of AP genes, compared with neutral genes.

AP genes are those whose null alleles are significantly fitter than the wild-type (WT) in at least one of the six media. Neutral genes are those whose null alleles are not significantly different in fitness from the WT under all six media. (A) Gene loss rates (per gene per strain) in 64 strains of diverse origins are lower among the entire set of 1249 AP genes (black bar) or AP genes identified from individual media (grey bars) than 1344 neutral genes (open bar). In all panels, *P*-values are from Mann-Whitney U test (*, $P<0.05$; **, $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$; *****, $P<0.00001$) or *t* test ($^{+}$, $P<0.05$; $^{++}$, $P<0.01$; $^{+++}$, $P<0.001$; $^{++++}$, $P<0.0001$; $^{+++++}$, $P<0.00001$). Error bars indicate one standard error. (B) Ratios of the number of nonsynonymous substitutions per nonsynonymous site (d_N) and the number of synonymous substitutions per synonymous site (d_S) between *S. cerevisiae* and *S. paradoxus* are lower for AP genes than neutral genes. (C) Gene expression divergences among four *Saccharomyces* species are lower in AP genes than neutral genes. (D) Gene expression divergences among yeast mutation accumulation (MA) lines are lower in AP genes than neutral genes. (E) Gene expression noise is lower for AP genes than neutral genes.

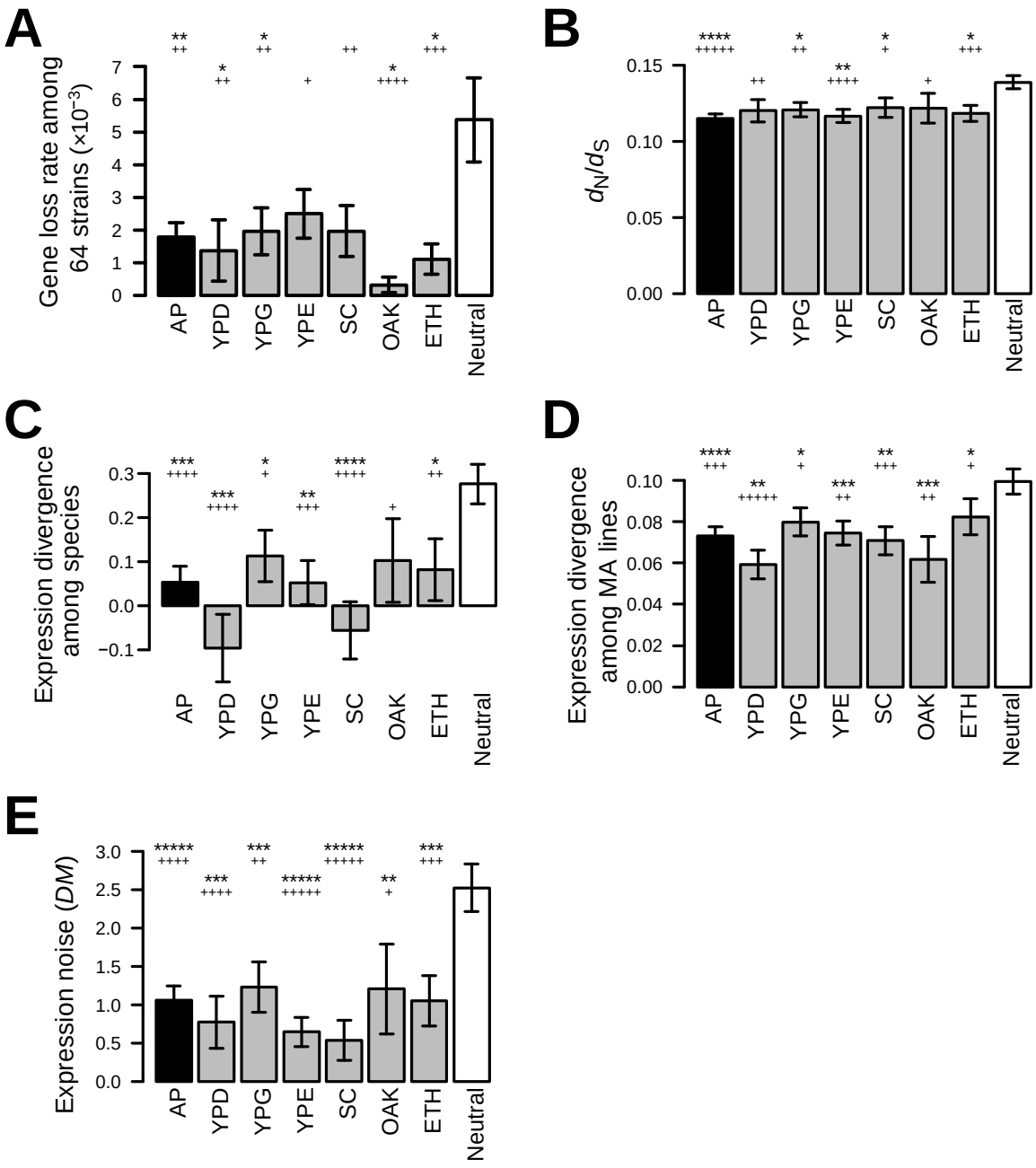

A
Gene loss rate among 64 strains (×10⁻³)
B
d_N/d_S
C
Expression divergence among species
D
Expression divergence among MA lines
E
Expression noise (DM)
AP
YPD
YPG
YPE
SC
OAK
ETH
Neutral

Figure 5.4 Significantly overrepresented Gene Ontology (GO) categories for genes whose null allele are fitter than the wild-type in ETH.

GO categories and their "parents" in the GO hierarchical architecture are connected by arrows. Blue arrows represent the "is a" relation and green arrows represent the "part of" relation. Node colors represent *P*-values of overrepresentation, while the cyan circle around a node indicates a significantly overrepresented GO category (*FDR* < 0.05). Node size reflects the number of genes in the GO category.

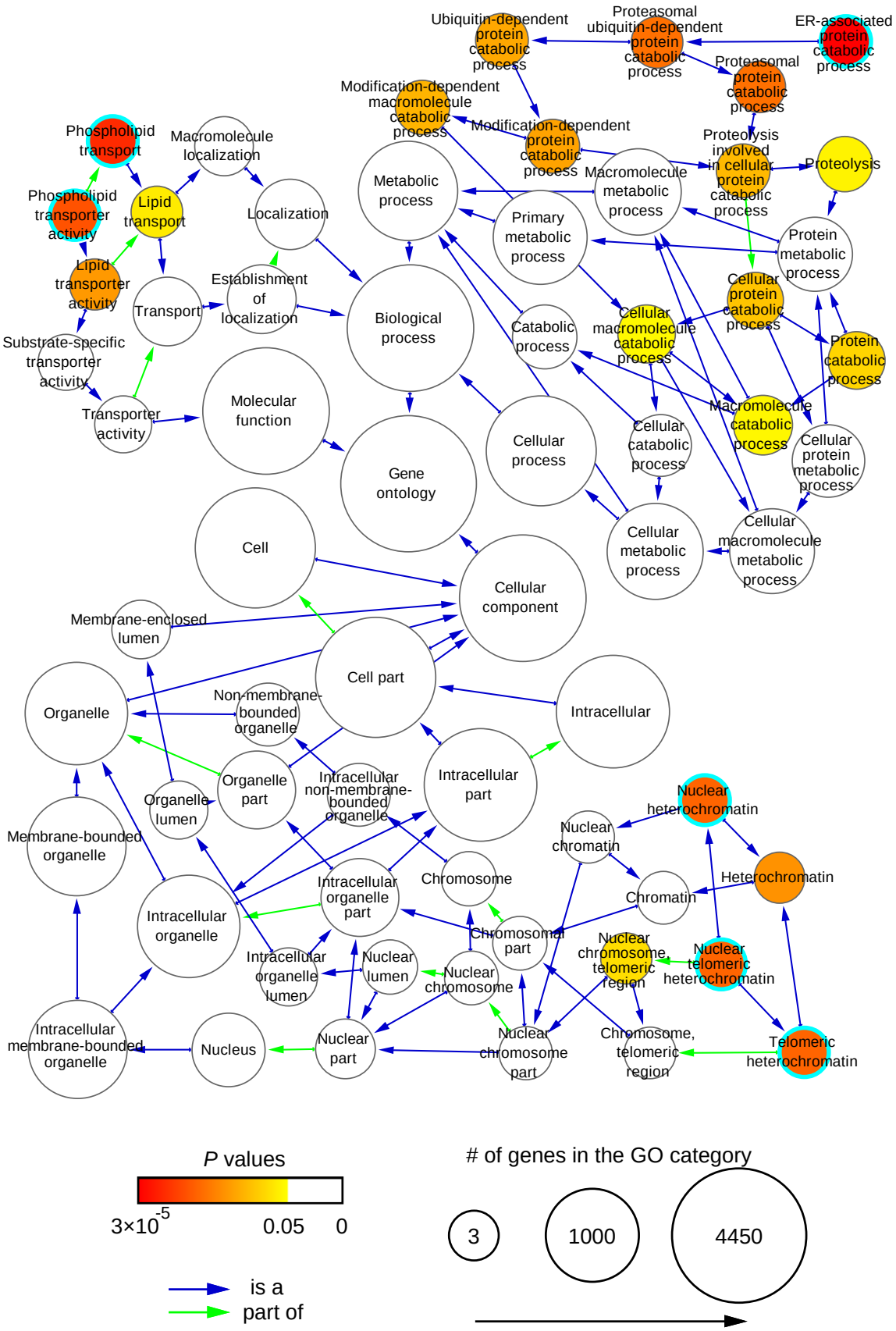

Ubiquitin-dependent protein catabolic process
Proteasomal ubiquitin-dependent protein catabolic process
ER-associated protein catabolic process
Proteasomal protein catabolic process
Modification-dependent macromolecule catabolic process
Modification-dependent protein catabolic process
Proteolysis involved in cellular protein catabolic process
Proteolysis
Phospholipid transport
Macromolecule localization
Phospholipid transporter activity
Lipid transport
Localization
Metabolic process
Macromolecule metabolic process
Primary metabolic process
Protein metabolic process
Lipid transporter activity
Transport
Establishment of localization
Biological process
Catabolic process
Cellular macromolecule catabolic process
Cellular protein catabolic process
Protein catabolic process
Substrate-specific transporter activity
Transporter activity
Molecular function
Gene ontology
Cellular process
Cellular catabolic process
Macromolecule catabolic process
Cellular protein metabolic process
Cell
Cellular component
Cellular metabolic process
Cellular macromolecule metabolic process
Membrane-enclosed lumen
Cell part
Organelle
Non-membrane-bounded organelle
Intracellular
Organelle part
Organelle lumen
Intracellular non-membrane-bounded organelle
Intracellular part
Membrane-bounded organelle
Nuclear heterochromatin
Nuclear chromatin
Heterochromatin
Chromosome
Chromatin
Intracellular organelle part
Intracellular organelle
Chromosomal part
Nuclear chromosome, telomeric region
Nuclear telomeric heterochromatin
Intracellular organelle lumen
Nuclear lumen
Nuclear chromosome
Intracellular membrane-bounded organelle
Nucleus
Nuclear part
Nuclear chromosome part
Chromosome, telomeric region
Telomeric heterochromatin
P values
3×10-5
0.05
0
# of genes in the GO category
3
1000
4450
is a
part of

Figure 5.5 AP is resolved at least partially by gene regulation in the presence of sufficient selection.

(A) Prior knowledge about the native environments of various yeast strains. Respir., respiration. (B) Relative fitness of four yeast strains in four media. The values are directly comparable across rows and across columns. See Experimental Procedures for the estimation of relative fitness. Due to severe flocculation of YPS1000 in ETH, its fitness could not be measured. (C) The null allele of *PDR17* is fitter than the wild-type (WT) in YPG ($P = 2 \times 10^{-5}$), but less fit than the WT in ETH ($P = 6 \times 10^{-8}$). In panels C-F and J, error bars show one standard error. (D) *PDR17* expression under YPG is lower in M22 than in K12 ($P = 2 \times 10^{-4}$). (E) *PDR17* expression of M22 is lower under YPG than under ETH ($P = 3 \times 10^{-3}$). (F) Expression-level difference between YPG and ETH is greater for M22 than for K12 ($P = 0.004$). (G-I) Numbers of examined genes in which AP is at least partially resolved (green) or unresolved (yellow) by transcriptional regulation, based on the same three tests shown for *PDR17* in panels d-f, respectively. (J) The null allele of *MIG1* is less fit than the WT in YPD ($P = 0.05$), but fitter than the WT in OAK ($P = 3 \times 10^{-48}$). (K) In the wild strain YPS1000, MIG1 is localized in the nucleus under YPD but in the cytoplasm under OAK. MIG1-GFP (green fluorescent protein) fusion protein allows the visualization of MIG1's subcellular localization. DAPI (4'-6-Diamidino-2-phenylindole) stains the nucleus in blue. DIC, differential interference contrast microscopy.

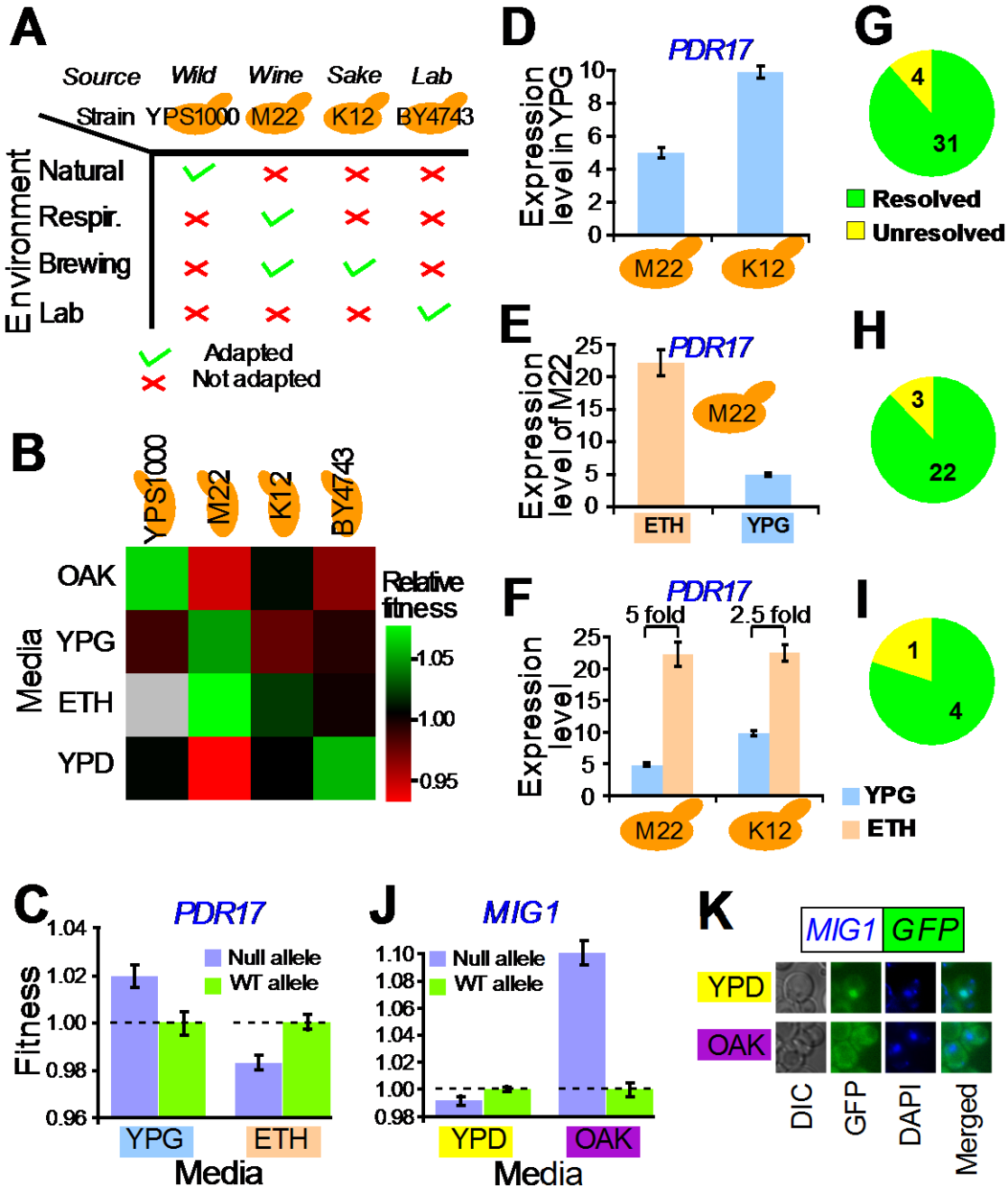

A
Source
Wild
Wine
Sake
Lab
Strain
YPS1000
M22
K12
BY4743
Environment
Natural
Respir.
Brewing
Lab
Adapted
Not adapted
B
YPS1000
M22
K12
BY4743
OAK
YPG
ETH
YPD
Media
Relative fitness
1.05
1.00
0.95
C
PDR17
Null allele
WT allele
Fitness
1.04
1.02
1.00
0.98
0.96
YPG
ETH
Media
D
PDR17
Expression level in YPG
10
8
6
4
2
0
M22
K12
E
PDR17
Expression level of M22
25
20
15
10
5
0
M22
ETH
YPG
F
PDR17
5 fold
2.5 fold
Expression level
25
20
15
10
5
0
M22
K12
YPG
ETH
G
4
31
Resolved
Unresolved
H
3
22
I
1
4
J
MIG1
Null allele
WT allele
1.10
1.08
1.06
1.04
1.02
1.00
0.98
YPD
OAK
Media
K
MIG1
GFP
YPD
OAK
DIC
GFP
DAPI
Merged

Figure 5.6 Genetic mechanisms of AP-alleviating environment-specific expression regulations.

(A) *PDR17* expression ratios between M22 (purple) and K12 (blue) alleles in mixed diploid parents and in hybrids. (B) *APQ12* expression ratios between M22 (purple) and K12 (blue) alleles in mixed diploid parents and in hybrids. (C) The null allele of *STP4* is fitter than the WT in YPG ($P = 5 \times 10^{-4}$), but less fit than the WT in ETH ($P = 5 \times 10^{-21}$). (D) *STP4* expression ratios between M22 and K12 alleles in mixed diploid parents and in hybrids. (E) Localization of causal mutation(s) responsible for the lowered *STP4* expression of M22, compared to K12, under YPG. Nucleotide differences between the two strains in the region between 442 nucleotides upstream and 238 nucleotides downstream of the translation starting site are presented, together with their positions relative to the translation starting site. We swapped between haploid strains of M22 (pM-M22) and K12 (pK-K12) a 250-nucleotide proximate promoter region that contains four single nucleotide differences to create two mosaic strains (pM-K12 and pK-M22). The expression levels of *STP4* in the four strains under YPG and ETH are depicted. In all panels, error bars show one standard error. One red star stands for significantly different expression levels at $P < 0.05$ between two genotypes connected by a grey line, while double stars indicate $P < 0.01$.

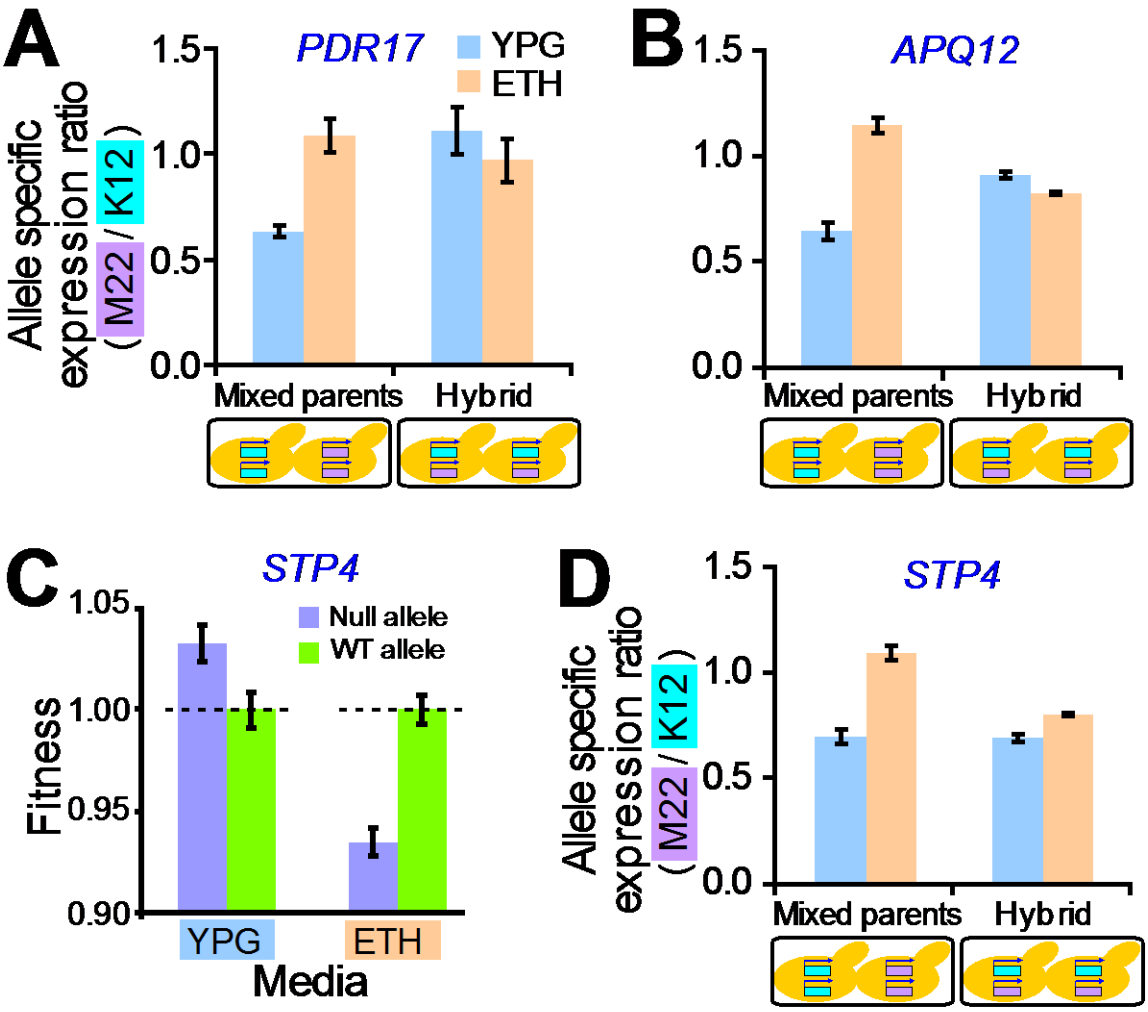
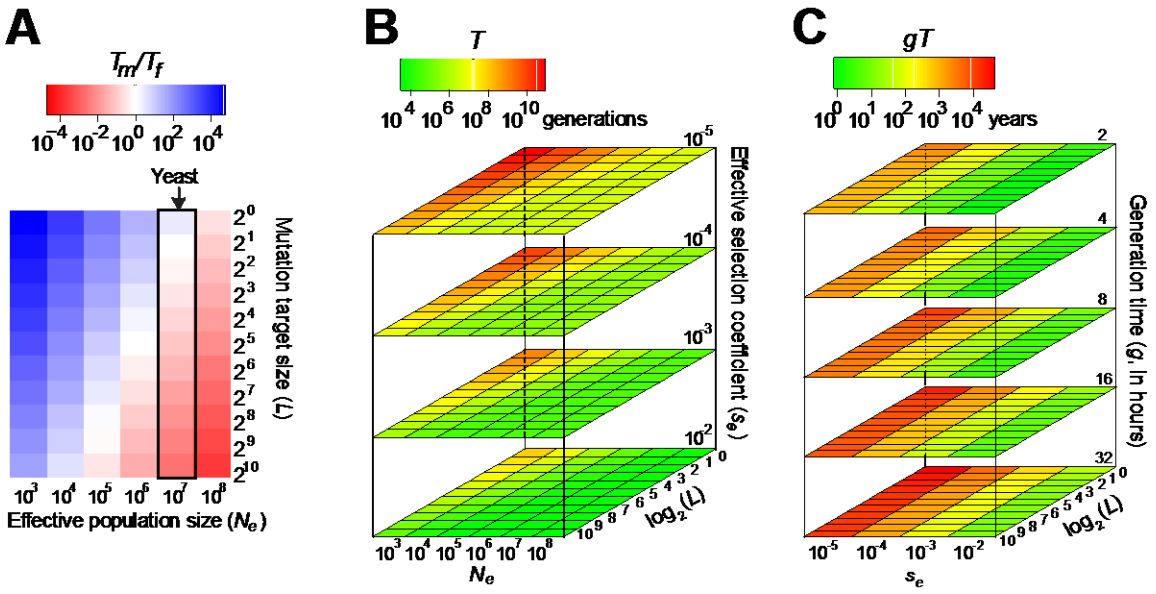

A
PDR17
YPG
ETH
Allele specific expression ratio (M22 / K12)
1.5
1.0
0.5
0.0
Mixed parents
Hybrid
B
APQ12
C
STP4
Null allele
WT allele
Fitness
1.05
1.00
0.95
0.90
YPG
ETH
Media
D
STP4
E
Expression level in YPG
8 6 4
Strains
M22
K12
pK-K12
pM-K12
pK-M22
pM-M22
STP4
-410 -324 -256 -134 -126 -93 -75 5 160 174
G T A A G A G C T T
C A - G A T A T C C
K12 K12 K12
K12 M22 K12
M22 K12 M22
M22 M22 M22
Expression level in ETH
10 20 30 40

Figure 5.7 Expected fixation times of AP-alleviating alleles.

(A) The ratio between the expected waiting time for the appearance of the first AP-alleviating allele that is destined for fixation (T_m) and the expected time required for this allele to get fixed from its first appearance (T_f) decreases with rising effective population size (N_e) and mutation target size (L). Yeast has an effective population size of ~10^7, as indicated by the box. (B) Expected total number of generations ($T = T_m+T_f$) required for the appearance and fixation of the first AP-alleviating allele decreases with rising L and effective selection coefficient (s_e). In almost all cases, T also decreases with rising N_e. (C) Expected number of years (gT) required for the appearance and fixation of the first AP-alleviating allele in yeast under different L, s_e, and generation time (g).



5.12 References

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**(8): 945-949.

Berlocher SH, Feder JL. 2002. Sympatric speciation in phytophagous insects: moving beyond controversy? *Annu Rev Entomol* **47**: 773-815.

Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD. 1998. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**(2): 115-132.

Brooks PJ. 1997. DNA damage, DNA repair, and alcohol toxicity--a review. *Alcohol Clin Exp Res* **21**(6): 1073-1082.

Carter AJ, Nguyen AQ. 2011. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med Genet* **12**(1): 160.

Cooper VS, Lenski RE. 2000. The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature* **407**(6805): 736-739.

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**(4): 1915-1925.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**(16): e105.

Donohue TM, Jr. 2009. Autophagy and ethanol-induced liver injury. *World J Gastroenterol* **15**(10): 1178-1185.

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.

Ekino K, Kwon I, Goto M, Yoshino S, Furukawa K. 1999. Functional analysis of HO gene in delayed homothallism in Saccharomyces cerevisiae wy2. *Yeast* **15**(6): 451-458.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**(2): 257-258.

Farrelly C. 2012. 'Positive biology' as a new paradigm for the medical sciences. Focusing on people who live long, happy, healthy lives might hold the key to improving human well-being. *EMBO Rep* **13**(3): 186-188.

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

Foster KR, Shaulsky G, Strassmann JE, Queller DC, Thompson CR. 2004. Pleiotropy as a mechanism to stabilize cooperation. *Nature* **431**(7009): 693-696.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**(6896): 387-391.

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM et al. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**(5987): 52-56.

Goldstein AL, McCusker JH. 1999. Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. *Yeast* **15**(14): 1541-1553.

Gruber JD, Vogel K, Kalay G, Wittkopp PJ. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects and dominance. *PLoS Genet*: e1002497.
He X, Qian W, Wang Z, Li Y, Zhang J. 2010. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**(3): 272-276.
He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. *Genetics* **173**(4): 1885-1891.
Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**(5874): 362-365.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**(1346): 119-124.
Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ et al. 2000. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet* **25**(3): 333-337.
Huxley C, Green ED, Dunham I. 1990. Rapid assessment of S. cerevisiae mating type by PCR. *Trends Genet* **6**(8): 236.
Ingram LO, Buttke TM. 1984. Effects of alcohols on micro-organisms. *Adv Microb Physiol* **25**: 253-300.
Innocenti P, Morrow EH. 2010. The sexually antagonistic genes of Drosophila melanogaster. *PLoS Biol* **8**(3): e1000335.
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**(6937): 241-254.
Kimura M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.
Komeili A, O'Shea EK. 2000. Nuclear transport and transcription. *Curr Opin Cell Biol* **12**(3): 355-360.
Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic properties influencing the evolvability of gene expression. *Science* **317**(5834): 118-121.
Lang GI, Murray AW, Botstein D. 2009. The cost of gene expression underlies a fitness trade-off in yeast. *Proc Natl Acad Sci U S A* **106**(14): 5755-5760.
Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* **4**: 170.
Levins R. 1968. *Evolution in Changing Environments* Princeton University Press, Princeton, NJ.
Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**(7236): 337-341.
Lynch M. 2007. *The Origins of Genome Architecture*. Sinauer, Sunderland, Mass.
Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**(27): 9272-9277.
Magwire MM, Yamamoto A, Carbone MA, Roshina NV, Symonenko AV, Pasyukova EG, Morozova TV, Mackay TF. 2010. Quantitative and molecular genetic

analyses of mutations increasing Drosophila life span. *PLoS Genet* **6**(7): e1001037.
Mather K, Harrison BJ. 1949. The manifold effect of selection. *Heredity* **Pt. 2 3**: 131-162.
Meiron H, Nahon E, Raveh D. 1995. Identification of the heterothallic mutation in HO-endonuclease of S. cerevisiae using HO/ho chimeric genes. *Curr Genet* **28**(4): 367-373.
Mishra P, Prasad R. 1988. Role of phospholipid head groups in ethanol tolerance of Saccharomyces cerevisiae. *J Gen Microbiol* **134**(12): 3205-3211.
Murphy HA, Kuehne HA, Francis CA, Sniegowski PD. 2006. Mate choice assays and mating propensity differences in natural yeast populations. *Biol Lett* **2**(4): 553-556.
Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**(7095): 840-846.
Orr HA. 2000. Adaptation and the cost of complexity. *Evolution* **54**(1): 13-20.
Ostrowski EA, Rozen DE, Lenski RE. 2005. Pleiotropic effects of beneficial mutations in Escherichia coli. *Evolution* **59**(11): 2343-2352.
Rice WR. 1992. Sexually antagonistic genes: experimental evidence. *Science* **256**(5062): 1436-1439.
Rodier F, Campisi J, Bhaumik D. 2007. Two faces of p53: aging and tumor suppression. *Nucleic Acids Res* **35**(22): 7475-7484.
Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* **458**(7236): 342-345.
Schuller HJ. 2003. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr Genet* **43**(3): 139-160.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**(11): 2498-2504.
Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14**(4): 450-456.
Sliwa P, Korona R. 2005. Loss of dispensable genes is not adaptive in yeast. *Proc Natl Acad Sci U S A* **102**(49): 17670-17674.
Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C. 2009. Quantitative phenotyping via deep barcode sequencing. *Genome Res* **19**(10): 1836-1842.
Smith EN, Kruglyak L. 2008. Gene-environment interaction in yeast gene expression. *PLoS Biol* **6**(4): e83.
Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ et al. 2002. Systematic screen for human disease genes in yeast. *Nat Genet* **31**(4): 400-404.
Storey JD. 2002. A direct approach to false discovery rates. *J Roy Statist Soc Ser B* **64**(3): 479-498.
Talebi AR, Sarcheshmeh AA, Khalili MA, Tabibnejad N. 2011. Effects of ethanol consumption on chromatin condensation and DNA integrity of epididymal spermatozoa in rat. *Alcohol* **45**(4): 403-409.

Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet* **38**(7): 830-834.
Wach A, Brachat A, Alberti-Segui C, Rebischung C, Philippsen P. 1997. Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in Saccharomyces cerevisiae. *Yeast* **13**(11): 1065-1075.
Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22**(6): 1365-1374.
Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**(3): 204-213.
Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS Biol* **4**(3): e52.
Wang Z, Liao BY, Zhang J. 2011. Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci U S A* **107**(42): 18034-18039.
Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc Natl Acad Sci U S A* **108**(16): E67-76.
Warringer J, Zorgo E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ et al. 2011. Trait variation in yeast is defined by population history. *PLoS Genet* **7**(6): e1002111.
Wenger JW, Piotrowski J, Nagarajan S, Chiotti K, Sherlock G, Rosenzweig F. 2011. Hunger artists: yeast adapted to carbon limitation show trade-offs under carbon sufficiency. *PLoS Genet* **7**(8): e1002202.
Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**: 398-411.
Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430**(6995): 85-88.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-1591.

CHAPTER 6
CONCLUSIONS

6.1 Concluding remarks

6.1.1 Summary

Functional genomics originated at the turn of the century (Eisenberg et al. 2000; Lockhart and Winzeler 2000). At the time, although a number of genome had been sequenced, it was unclear how to interpret the sequences. Thus, the focus of research started to switch from sequencing genomes to studying the function of genes in a genome, including their temporal and spatial expression patterns (Spellman et al. 1998), subcellular localizations (Huh et al. 2003), interactions with other molecules in the cell (Uetz et al. 2000; Ito et al. 2001; Gavin et al. 2002; Ho et al. 2002; Gavin et al. 2006; Krogan et al. 2006), fitness effects upon deletion (Giaever et al. 2002), and genetic interactions with each other (Tong et al. 2001; Tong et al. 2004; Costanzo et al. 2010). This wave of functional genomics has had a significant impact on evolutionary genetics. First, traditional evolutionary genomics focused on identifying nucleotide sites in a genome that are under natural selection, including negative selection and positive selection. However, the selective agent as well as the phenotypic consequence of the selection is often unclear. For example, it is unknown if evolution is repeatable and reversible. Second, traditional functional evolutionary genetics has provided interesting results in a small number of genes. However, the relative importance of various evolutionary mechanisms identified from a handful of genes remains unclear at the genomic scale. Functional evolutionary genomics combines the strengths of functional genomics and evolutionary genomics and promises to offer deep and grand insights into evolutionary processes.

The past few years has seen a series of studies in evolutionary functional genomics. For example, DNA polymorphisms were associated with phenotypic variations among wild strains of yeast or plant (Steinmetz et al. 2002; Atwell et al. 2010;

Ehrenreich et al. 2010). In addition, gene expression evolution has been extensively studied at the genomic scale (Wittkopp et al. 2008; Tirosh et al. 2009; Emerson et al. 2010; McManus et al. 2010). Many interesting results were also obtained from evolutionary analyses of protein-protein interaction data (Wagner 2001; He and Zhang 2005b; He and Zhang 2006c) and datasets of fitness effects of gene deletions (Gu et al. 2003; He and Zhang 2005a; He and Zhang 2006a; He and Zhang 2006b; He and Zhang 2006c; Liao and Zhang 2007; Liao and Zhang 2008).

My studies were built on these and other studies. I focused on understanding evolutionary processes through both large-scale functional genomic analysis and small-scale experimental validation. I studied the evolutionary rate of protein-protein interaction, the distribution as well as the underlying mechanisms of genetic interactions, and the prevalence and evolutionary resolution of antagonistic pleiotropy.

My studies suggest that one should be cautious in using various functional genomic data. For example, in Chapter 2, I found high false negative and false positive rates in high-throughput protein interaction data. These errors make such data produced by different teams or for different species incomparable. To avoid such problems, I used high-quality protein interaction data from *S. cerevisiae* and experimentally examined the corresponding interaction in a related yeast species. Further, I examined the protein interaction data of *S. cerevisiae* experimentally to confirm any apparent difference between the two species. Such caution is absolutely required for an accurate measure of the evolutionary rate of protein interaction.

I combined genome-wide patterns revealed by computational analysis or high-through-put experiments with small-scale experimental validation. For example, in Chapter 3, I experimentally validated our theoretical and computational predictions of epistasis. In Chapter 4, to test our model of translational efficiency by unequal codon usage, I constructed four synonymous versions of *mCherry* and measured their impacts on the cellular translational efficiency. In Chapter 5, I tested two competing hypotheses about the presence of antagonistic pleiotropy by measuring expression levels of the relevant genes in multiple wild yeast strains that have been adapted to different environments. I also studied the detailed genetic mechanisms of evolutionary resolution

of antagonistic pleiotropy in a few genes. Such experiments strengthened my conclusions.

Many of my studies have been further confirmed by other studies. In Chapter 3, I studied the genome-wide distribution of epistasis, and my result was echoed by a recent experimental study (Jakubowska and Korona 2012). In Chapter 4, I found that synonymous codons have similar translational elongation rates. Recent studies in mouse and *E. coli* also discovered this phenomenon (Ingolia et al. 2011; Li et al. 2012). It is interesting to note that mouse study was based on a different experimental strategy (Ingolia et al. 2011).

6.1.2 Implications

My findings have multiple implications in evolutionary genetics. First, my results illustrate the importance of understanding the genetic basis of evolutionary phenomena such as antagonistic pleiotropy.

Second, my research emphasizes the importance of understanding the molecular functions of genes in order to understanding their evolution. For instance, the evolutionary rates of protein sequences have been studied extensively in the past 50 years and many theories have been proposed (Zukerkandl and Pauling 1965; Kimura 1968; Li 1997; Nei and Kumar 2000; Koonin and Galperin 2003; Drummond and Wilke 2008; Wolf et al. 2009; Doolittle 2010). However, it is still mysteries how protein sequence changes give rise to morphological, physiological, or behavioral evolution that we observe. Because the connection between protein sequence and phenotype is often hard to investigate, it is interesting to study this question at an intermediate level of protein function. In Chapter 2, I aimed to connect between protein sequence evolution and protein function evolution. I found that protein interaction evolution is extremely slow, and no correlation with protein sequence evolution was observed.

Third, my research showcases the power of functional genomics. It had been known that both positive or negative epistasis exist among genes (Tong et al. 2001; Tong et al. 2004; Costanzo et al. 2010). The overall sign of epistasis is important for various evolutionary theories (Crow and Kimura 1979; Kondrashov 1988; Kondrashov and Crow 1991; Coyne 1992; Barton and Charlesworth 1998; Jasnos and Korona 2007). The

overall pattern cannot be summarized from individual studies, because of various biases in data collection. Measuring all pairwise epistasis, however, is extremely time-consuming. In Chapter 3, an *in silico* prediction of epistasis based on metabolomics offered a strategy to study epistasis at the genomic scale.

Fourth, my research emphasizes the importance of studying evolution from a systems perspective. For example, the phenomenon of codon usage bias has been studied for decades. It was noticed that highly expressed genes preferentially use a set of synonymous codons (Ikemura 1981). Furthermore, such codons are recognized by a set of high concentration tRNAs. It was then natural to assume that the preferred synonymous codons are more efficient. However, those tRNAs are not necessarily more efficient than the others when an equilibrium between tRNA and codon usage is achieved. In Chapter 4, I provided multiple lines of evidence that natural selection has acted on the balance between codon usage and tRNA concentration and all synonymous codons have similar translation efficiency.

Fifth, my research has potential implications for synthetic biology. Understanding the basis of evolutionary interesting traits can help design new traits in an organism, which is the aim of synthetic biology. For example, in Chapter 4, by understanding how nature has selected synonymous codons, we can design synthetic cells or genomes following the same rule. In Chapter 5, I investigated how the expression regulation of a gene in different environments evolved. The same strategy could be used to introduce a new gene into the genome that has antagonistic effects in multiple environments.

6.1.3 Future directions

My research covers a few interesting questions on functional evolutionary genomics; more questions can be addressed in the future.

First, numerous other types of functional genomic data can potentially be integrated into the current study of functional evolutionary genomics. In my thesis, I used a broad range of functional genomic data, including gene expression levels, protein levels, protein-protein interactions, genetic interactions, gene-environment interactions and growth rates. Several other functional genomic data also became available in recent

years, e.g., nucleosome binding (Lee et al. 2007), mRNA folding (Kertesz et al. 2010), chromosome territory (Duan et al. 2010), and transcription factor binding (Lee et al. 2002). It would be interesting and important to understand the evolution of such molecular traits and how the molecular traits influence the evolution process.

It is worth noting that such studies are essential for understanding the evolution with extensive pleiotropy. A single mutation in the coding sequence of a gene can have multiple effects. It may change the protein sequence, codon usage, mRNA secondary structure, post-translational modification, and nucleosome binding, among other things. These changes may influence the dynamics of gene expression, e.g., transcription elongation rate, translation initiation rate, translation elongation rate, and translational accuracy. Such dynamic parameters of gene expression could further impact the amount, variation, and timing of gene expression. All aforementioned aspects of gene expression are potentially important to fitness. Such pleiotropic effects of a mutation on different aspects of gene expression are poorly understood. Because of the tremendous functional genomic data that are available in yeast, yeast serves as an excellent model eukaryote for studying the genetic architecture of gene expression.

Second, the effects of point mutations need to be emphasized in the future. Whereas epistasis and pleiotropy of loss-of-function mutations have been studied recently at the genomic scale (Tong et al. 2001; Tong et al. 2004; Costanzo et al. 2010), epistasis and pleiotropy of point mutations have not been accurately measured in a systematic way, despite that point mutations are more prevalent and important than null mutations in nature. This deficiency is mainly due to the difficulty in generating and quantifying the fitness of a large number of mutants. Some recent studies circumvented the hurdle and provided some interesting results (Weinreich et al. 2006; Shultzaberger et al. 2010). However, epistasis and pleiotropy of point mutations deserve more efforts.

Third, it is important to validate if the conclusions obtained in unicellular organisms are correct in multicellular organisms. Unicellular organisms may be different from multicellular organisms in several aspects. For example, the growth rate of unicellular organisms may be limited mainly by the environment, whereas the growth rate of multicellular organisms is more constrained by the development process that is genetically determined. Such differences may make our conclusion in yeast less

applicable to multicellular organisms. It is thus interesting to study how our conclusions in yeast can be expanded to multicellular organisms.

6.2 Reference

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**(7298): 627-631.

Barton NH, Charlesworth B. 1998. Why sex and recombination? *Science* **281**(5385): 1986-1990.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S et al. 2010. The genetic landscape of a cell. *Science* **327**(5964): 425-431.

Coyne JA. 1992. Genetics and speciation. *Nature* **355**(6360): 511-515.

Crow JF, Kimura M. 1979. Efficiency of truncation selection. *Proc Natl Acad Sci U S A* **76**(1): 396-399.

Doolittle RF. 2010. The roots of bioinformatics in protein evolution. *PLoS Comput Biol* **6**(7): e1000875.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**(2): 341-352.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**(7296): 363-367.

Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, Gresham D, Caudy AA, Kruglyak L. 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**(7291): 1039-1042.

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. 2000. Protein function in the post-genomic era. *Nature* **405**(6788): 823-826.

Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. 2010. Natural selection on cis and trans regulation in yeasts. *Genome Res* **20**(6): 826-836.

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**(7084): 631-636.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**(6868): 141-147.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**(6896): 387-391.

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**(6918): 63-66.

He X, Zhang J. 2005a. Gene complexity and gene duplicability. *Curr Biol* **15**(11): 1016-1021.

He X, Zhang J. 2005b. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.
He X, Zhang J. 2006a. Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol* **23**(1): 144-151.
He X, Zhang J. 2006b. Toward a molecular understanding of pleiotropy. *Genetics* **173**(4): 1885-1891.
He X, Zhang J. 2006c. Why do hubs tend to be essential in protein networks? *PLoS Genet* **2**(6): e88.
Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**(6868): 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**(6959): 686-691.
Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**(3): 389-409.
Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**(4): 789-802.
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**(8): 4569-4574.
Jakubowska A, Korona R. 2012. Epistasis for growth rate and total metabolic flux in yeast. *PLoS One* **7**(3): e33132.
Jasnos L, Korona R. 2007. Epistatic buffering of fitness loss in yeast double deletion strains. *Nat Genet* **39**(4): 550-554.
Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**(7311): 103-107.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**(5129): 624-626.
Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**(6198): 435-440.
Kondrashov AS, Crow JF. 1991. Haploidy or diploidy: which is better? *Nature* **351**(6324): 314-315.
Koonin E, Galperin M. 2003. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* Kluwer Academic Publishers, Boston.
Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP et al. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**(7084): 637-643.
Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I et al. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**(5594): 799-804.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39**(10): 1235-1244.
Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**(7395): 538-541.
Li W. 1997. *Molecular Evolution*. Sinauer, Sunderland, Mass.
Liao BY, Zhang J. 2007. Mouse duplicate genes are as essential as singletons. *Trends Genet* **23**(8): 378-381.
Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.
Lockhart DJ, Winzeler EA. 2000. Genomics, gene expression and DNA arrays. *Nature* **405**(6788): 827-836.
McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**(6): 816-825.
Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
Shultzaberger RK, Malashock DS, Kirsch JF, Eisen MB. 2010. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLoS Genet* **6**(7): e1001042.
Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**(12): 3273-3297.
Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW. 2002. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**(6878): 326-330.
Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**(5927): 659-662.
Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**(5550): 2364-2368.
Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303**(5659): 808-813.
Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**(6770): 623-627.
Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**(7): 1283-1292.
Weinreich DM, Delaney NF, Depristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**(5770): 111-114.
Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* **40**(3): 346-350.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. Inaugural Article: The universal distribution of evolutionary rates of genes and distinct

characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A* **106**(18): 7273-7280.

Zukerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, (ed. V Bryson, HJ Vogel), pp. 97-166. Academic Press, New York.