Fitness and Epistatic Effects of Synonymous Mutations in Yeast

by

Xukang Shen

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
2023

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor Timothy James
Professor Anuj Kumar
Professor Patricia Wittkopp

Xukang Shen

xukang@umich.edu

ORCID iD: 0000-0002-9446-4644

Dedication

To the *Saccharomyces cerevisiae* cells in my experiments

Acknowledgements

I could not finish my Ph.D. journey without the help from numerous people along the way.

I am very lucky to have worked with George. I still remember my interview in February 2017 on a sunny but very chilly day. Rex and I walked past our lab which was still in the Kraus Building. George waved to us from his office and we waved to him. In my mind, it happened yesterday, although it was actually six years ago. I learned two significant merits from George. First, I should take a step back to see the whole picture whenever I study a question. In our lab meetings, journal clubs, and discussions, George usually reminds us where we are in the context of evolution, biology, or even science. These reminders act as compasses on ships, allowing us to know the importance of the questions we are talking about and where we can go beyond that. Second, in every discussion with George, we never stopped until we have a clear idea of what to do next. It requires curiosity, perseverance, and intellectual courage.

I would also like to thank all my committee members, Professor Trisha Wittkopp, Professor Timothy James, and Professor Anuj Kumar for their time and effort in serving on my thesis committee and for providing many constructive suggestions. I taught an Authentic Research class on experimental evolution with George and Tim for three terms. I participated from its first iteration in Winter 2018 to the unforgettable last term in Winter 2020. Retrospectively, teaching that class was one of the most delightful experiences in the last six years. I still remember that after the lectures, we climbed up to the lab room, Lucas (Winter

2018), Rebecca, Alex (Winter 2019 and Winter 2020), Haiqing (Winter 2020), Kevin (Winter 2019), Tim, George and I discussed the experimental setup in a very relaxing way, and we laughed often. I luckily received a Teaching Excellence Award, which I have kept along with my other vital documents. I learned that the experimental evolution class would be back in Fall 2023 in a somewhat different but more exciting form, and am very happy for its return.

I definitely want to thank the lab mates in our lab. I learned a lot from them, about evolution, experiments, biology, science, basketball, good restaurants, etc. Especially, I want to thank Haiqing Xu who taught me how to do experiments. Whenever I met a difficult problem, he was the first person I went to. Outside the lab, we are also very close friends. We had a nice trip to Florida in 2017, traveled in Japan after attending SMBE 2018 in Yokohama, and drove together to the Evolution conference of 2022 in Cleveland. I also want to thank Anjali Mahilkar, Mengyi Sun, Siliang Song, Piaopiao Chen, Haoxuan Liu, Chuan Li, Daohan 'Rex' Jiang, Zhengting Zou, Xinzhu Wei, Chuan Xu, Daniel Lyons, and Wei-Chin Ho. I have very close contacts with some of them.

I want to thank my friends in and outside the EEB department. Their companionship makes my Ph.D. journey easier and more enriching. I would like to thank Henry Ertl, Nikesh Dahal, Yi-hong Ke, Youjia Wu, Nancy Barlett, Qi Geng, Jiaqian Li, Yihao Yang, and the MB2 neighborhood. I am very thankful for the past and current graduate coordinators in EEB: Cindy Carl, Kati Ellis, and Nathan Sadowsky.

I also want to thank one of my college professors – Xuefu Zhang, who was the instructor of my Ancient Greek and Roman Philosophy class in 2015. In that class, we read the Republic by Plato chapter by chapter and discussed it every week. He was such a good teacher that his passion and knowledge led me to thousands of years ago, and I realized that humans worry

similar things throughout the history. Although philosophy cannot give definite answers to questions as science can, it helps me live in a more self-examined way. That class was a start. After it, I friended with Xuefu and began reading philosophy books and philosophers' biographies under his guidance. Through reading these great thinkers' books and discussions with Xuefu, I learned how excellent a human can be and what I should focus on. Also, with the influence from Xuefu, I started reading anthropology books, especially, social anthropology. I become more curious about this world, travel to those places I have never been to in mind, and talk with those people in books. Xuefu encourages me to seize an important question and not to let it go. He also encourages me to be a global citizen, transcending the limitation of nationality and country and living as an independent individual. His encouragements sometimes sound a little outdated in this gradually de-globalized and disappointing world but they really support me through these years and make me believe that there is something good to follow.

At last, I want to thank Wei Chen and my parents for their support and love. Although my parents are thousands of miles away from Ann Arbor, their words of encouragement can reach and embrace me. When I was a kid, we three used to travel in Zhejiang province in our small Citroen car. Zhejiang is half the size of Michigan, but it is a mountainous province and the cultures can be very diverse even in two cities only one or two hundred kilometers away. I enjoyed listening to various accents (though I hardly understood them) and appreciated the beautiful views along the trips. After I grew up a little bit, we drove and traveled around many places in China. Even now, I sometimes think about the people we met in those trips and imagine what they are doing now. My parents make me believe that there is always a bigger world outside of my current life and that I should be confident in myself. I spent most of my time with Wei in the past five years. Her companionship made many difficult moments much less stressful.

She is so smart and curious that with her I explored more of this world. One day during the lockdown in 2020, she suggested that birding could be interesting, which started our birding. Now, we have seen nearly 200 bird species in Washtenaw County and more than 300 species in the US. With her, I am a better me. We shall see as many birds as possible in the coming decades in this great world.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Synonymous mutations have long been considered neutral, but there is accumulating evidence that they influence many biological processes, ranging from transcription to translation, except that they do not alter amino acid sequences. In this thesis, I directly tested the neutral assumption of synonymous mutations by measuring the fitness effects of more than 8,000 mutations, including 1,866 synonymous ones, in 21 genes in the budding yeast *Saccharomyces cerevisiae*. I found that synonymous and nonsynonymous mutations have similar distributions of fitness effects and that most synonymous and nonsynonymous mutations are non-neutral. Both types of mutations can influence the mRNA level and thereby affect fitness. Nonsynonymous mutations have larger across-environment fitness variations than synonymous mutations. As a result, nonsynonymous mutations are more likely to be purged in fluctuating environments, which may explain why the nonsynonymous to synonymous substitution rate ratio (d_N/d_S) between species is below 1 for almost all genes. To confirm that the fitness effects observed are not artifacts, I performed whole-genome sequencing of the progenitor strain, gene deletion strains, wild-type control strains, and on average ~28 mutants per gene. I found no off-target genome editing but observed secondary mutations in some mutants. Notwithstanding, all results were verified in the subset of genes with negligible effects of secondary mutations. To study whether the fitness effect of a synonymous mutation depends on the presence/absence of other mutations in the gene, I studied intragenic epistasis between mutations by using double mutants created in my experiments. I found that synonymous mutations can genetically interact with

synonymous mutations and nonsynonymous mutations and that 8.5% to 26.1% of instances of epistasis vary significantly between two environments. I also found that epistasis between nonsynonymous mutations is more variable than that between synonymous mutations across the four environments examined. Together, these studies deepen our understanding of the fitness and epistatic effects of synonymous mutations and demand a reconsideration of many previous conclusions dependent on the neutral assumption of synonymous mutations.

Chapter 1 Introduction

The discovery of the non-neutrality of synonymous mutations

In the 1960s, when the genetic code was deciphered, it was found that 18 amino acids were each encoded by multiple codons. These different codons that code for the same amino acid were named synonymous codons. Consequently, nucleotide mutational changes between synonymous codons were called synonymous mutations or silent mutations. By contrast, nucleotide mutations that alter the encoded amino acids were called nonsynonymous mutations or missense mutations.

Also in the 1960s, a new evolutionary theory—the neutral theory—was developed (Kimura, 1968; Kimura, 1983; King & Jukes, 1969). The neutral theory asserts that most nucleotide differences between species result from random fixations of neutral mutations and most intraspecific polymorphisms are also neutral. Because synonymous mutations do not change amino acid sequences, these mutations were assumed neutral. As a result, evolutionists regard synonymous changes as neutral markers in evolution and this concept is behind many evolutionary methods and analyses. For example, widely used methods for detecting natural selection such as the McDonald-Kreitman test (McDonald & Kreitman, 1991) and d_N/d_S test (Kimura, 1983) use the frequencies of synonymous changes as the neutral baseline. Effective population size (N_e) is an important concept in population genetics and conservation biology; synonymous polymorphisms are used for estimating N_e because these polymorphisms are believed to be neutral (Gillespie, 2004).

However, in the 1970s, the first evidence that synonymous mutations are not all neutral was discovered. At that time, some genes were sequenced and it was found that some codons were used more frequently than their synonymous codons (Grantham et al., 1980). This phenomenon is called codon usage bias (CUB) and this pattern is more profound for highly expressed genes. Consequently, synonymous codons more prevalently used in highly expressed genes were named “preferred codons” while less-used ones were named “unpreferred codons” (Grantham et al., 1981). The positive correlation between CUB and gene expression level suggested that synonymous mutations might affect translation (Sharp & Li, 1986). Subsequent research has provided mounting evidence supporting this idea, with studies showing that synonymous mutations can impact various aspects of translation, including initiation (Kudla et al., 2009), efficiency (Ikemura, 1981), accuracy (Akashi, 1994), and co-translational protein folding (Buhr et al., 2016). Besides the impact on translation, synonymous mutations have also been reported to participate in transcription (Stergachis et al., 2013; Zhou et al., 2016) and pre-mRNA splicing (Chamary et al., 2006). Thanks to the advancements in molecular biology tools and sequencing techniques, scientists can now measure the fitness effects of synonymous mutations with greater precision. Surprisingly, many studies have shown that these assumed neutral mutations are, in fact, non-neutral in both bacteria (Kristofich et al., 2018; Lind et al., 2010) and eukaryotes (Sharon et al., 2018).

Biological processes affected by synonymous mutations

Transcription initiation

Stergachis *et al.* (Stergachis et al., 2013) found that ~15% of human codons not only encode amino acids but also specify transcription factor (TF) binding sites, so named these codons 'duons'. Synonymous mutations in the duons may alter TF binding.

Zhou *et al.* performed codon optimization of eight *Neurospora* genes, two heterologous reporter genes, firefly luciferase, and *S. cerevisae I-sel* gene based on the *Neurospora* codon usage (Zhou et al., 2016). Then, these codon-optimized genes were inserted into a *Neurospora crassa* chromosomal locus. Codon optimization increased the mRNA abundance of these genes by promoting enrichment of RNA polymerase II.

Pre-mRNA splicing

Synonymous mutations near introns can affect mRNA splicing (Chamary et al., 2006). For example, *SMN2* and *SMN1* are two paralogous genes in the human genome. *SMN2* has an almost identical sequence as *SMN1* but has a synonymous difference in Exon 7. Because of this synonymous difference, 80% of *SMN2* mRNA skips Exon 7 and produces truncated, unstable protein products (Pagani & Baralle, 2004).

mRNA folding and stability

Synonymous mutations change mRNA sequences and potentially their folding. Park *et al.* found that natural selection for mRNA folding is stronger in more highly expressed genes and that random mutations are more likely to decrease mRNA folding in highly than lowly expressed genes (Park et al., 2013). It was found that changes in mRNA folding can influence the functional mRNA half-life and protein concentration (Mauger et al. 2019) as well as the translational elongation speed and accuracy (Yang et al., 2014).

Presnyak *et al.* found that using preferred codons alone increases the mRNA stability and thereby increases the mRNA level (Presnyak et al., 2015). Later, the same group (Buschauer et al., 2020) found that unpreferred codons are associated with less abundant tRNA molecules, which slows down the ribosomal translation process. The Ccr4-Not pathway monitors this delay and degrades the mRNAs occupied by slow ribosomes.

Translational initiation

Kudla *et al.* built a mutant library by randomizing the synonymous codons of the GFP gene (Kudla et al., 2009). The protein expression level of GFP varied by ~250 fold among the variants. Surprisingly, they found that the mRNA folding near the translation initiation site explains one half of the expression variation. Synonymous mutations in the first 30 coding nucleotides can change the mRNA folding and ribosomal binding so may alter the translation initiation.

Translational efficiency

In a genome, the cognate tRNA gene copy number varies among the synonymous codons of an amino acid. Preferred codons are associated with high-gene-copy-number cognate tRNAs. In the translation process, the waiting time for an abundant tRNA is shorter than that for a rare tRNA. So, preferred codons are predicted to be translated faster than unpreferred codons (Ikemura, 1981). Ribo-seq (Ingolia et al., 2011) could sequence mRNA segments protected by ribosomes. So, efficiently translated codons are less likely than inefficiently translated codons to be captured in ribo-seq. This pattern was observed and indicated that preferred codons are indeed translated more efficiently (Hussmann et al., 2015; Weinberg et al., 2016).

Translational accuracy

The first evidence supporting the hypothesis that preferred codons are translated more accurately than unpreferred codons was from a comparative analysis published in 1994 (Akashi, 1994). Akashi compared the amino acid sequences of 38 genes between *Drosophila melanogaster* and *D. virilis* or *D. pseudoobscura*. He found that, for a given amino acid, conserved residues are more likely than unconserved ones to be encoded by preferred codons. Because conserved residues are important for protein functions, translational errors must be less tolerable. Therefore, Akashi concluded that preferred codons are translated more accurately.

With the advancement of proteomics, Sun and Zhang analyzed proteome-wide mistranslation events in *E. coli*, reporting that preferred codons are translated more accurately than unpreferred ones (Sun & Zhang, 2022).

Co-translational protein folding

Synonymous mutations can change the translation dynamics and alter protein folding during translation (Buhr et al., 2016). Synonymous mutations in the chloramphenicol acetyltransferase (CAT) gene alter the co-translational protein folding via changing the translational elongation speed, leading to enhanced protein degradation in vivo and dramatically lowered *E. coli* fitness (Walsh et al., 2020).

Experimental evidence of the non-neutrality of synonymous mutations from past systematic studies

Because synonymous mutations have been found to influence many biological processes, the assumption that synonymous mutations are neutral is questionable. To directly test this in a systematic way, several authors made synonymous mutations in a genome and measured their fitness effects.

In 2010, Lind *et al.* (Lind et al., 2010) did mutagenesis in two ribosomal protein genes in the bacterium *Salmonella typhimurium*. They created 38 synonymous mutants and 88 nonsynonymous mutants. Each mutant carries only one mutation in the genome. The fitness effects of these mutations were measured in LB and M9 glucose minimum media. In both media, the fitness distributions of synonymous and nonsynonymous mutants are quite similar. In LB medium, the average growth rate of nonsynonymous mutants relative to the wild type was 0.94, while the average relative growth rate of synonymous mutants was 0.92. In the M9 glucose minimum medium, the average relative growth rates were 0.95 for both synonymous and nonsynonymous mutants.

With the development of CRISPR/Cas9 genome editing, larger-scale mutagenesis became possible. Sharon *et al.* (Sharon et al., 2018) compared the genomes of the BY strain of *S. cerevisiae* and a vineyard strain (RM) of *S. cerevisiae*. They identified genomic differences between these two strains, individually replaced the alleles in the BY strain with the corresponding alleles of the RM strain, and measured the fitness effects of these replacements using a sequencing-based method. They found similar fractions of synonymous and nonsynonymous replacements with significant fitness effects.

Potential impacts of the non-neutrality of synonymous mutations on evolutionary biology

Synonymous mutations have long been assumed to be neutral markers in population genetic and evolutionary studies and in conservation biology. Invalidation of this assumption would have broad impacts. Here, I will mainly focus on the effective population size estimation and tests of natural selection.

Effective population size (N_e) is the size of an idealized population that would have the same effect of random sampling on gene frequency as in the actual population. N_e influences the effectiveness of selection relative to genetic drift so is important in conservation biology. N_e is usually calculated by dividing the genetic diversity (π) by 2μ (haploids) or 4μ (diploids), where μ is the neutral mutation rate. Genetic diversity is usually estimated based on the synonymous variants with the presumption that they are neutral in evolution. If many synonymous mutations are not neutral, genetic diversity is reduced and thereby N_e is underestimated.

In tests of natural selection, the frequency of synonymous variants is usually regarded as the neutral baseline. In the d_N/d_S test (Kimura, 1983), d_N is the number of nonsynonymous substitutions per nonsynonymous site and d_S is the number of synonymous substitutions per synonymous site. If many synonymous mutations are non-neutral, d_S is lower than the neutral expectation so d_N/d_S is inflated. In the McDonald-Kreitman test (McDonald & Kreitman, 1991), the amount of variation within species (polymorphism) is compared with the variation between species (divergence). Synonymous polymorphisms and substitutions are used as neutral baselines. If many synonymous mutations are non-neutral, the results of the McDonald-Kreitman test are confounded because synonymous polymorphisms and substitutions no longer reflect the neutral baselines.

Epistasis (G×G) and variation in epistasis across environments (G×G×E)

In the systematic study of the fitness effects of synonymous and nonsynonymous mutations, we study one mutation at a time. To have a more comprehensive understanding of the mutational effect, we need to take epistasis (G×G) into consideration.

'Epistasis' was first coined by William Bateson in 1909 to describe the interaction of mutations in influencing the phenotype (Bateson, 1909). Among various definitions of epistasis, the most commonly used one is $\varepsilon = f_{AB} - f_A f_B$, where f_{AB} is the fitness of a double mutant and f_A and f_B are the fitness of the two corresponding single mutants. With the advent of next-generation sequencing and other technologies, systematic quantifications of epistasis became possible. Intragenic epistasis between mutations was measured for the whole gene (Li et al., 2016; Puchta et al., 2016) or one segment of a gene (Olson et al., 2014). Intergenic epistasis was measured by large-scale gene knockout experiments (Costanzo et al., 2010).

It is known that epistasis varies across environments (G×G×E). In a systematic multi-environment fitness landscape study (Li & Zhang, 2018), it was found that 5.1% to 48.7% of epistatic interactions are significantly different between any two of the four environments examined but epistasis in one environment can be predicted from the fitness data collected in another environment.

Thesis overview

In Chapter 2, I created mutant libraries of 21 yeast genes, consisting of more than 8000 mutants (including 1866 synonymous mutants). I first measured the fitness effects of mutations in YPD and found most synonymous and nonsynonymous mutations to be non-neutral. I then experimentally measured the mRNA level of each mutant and compared it with the wild-type gene expression level, finding that both synonymous and nonsynonymous mutations can change the mRNA abundance and that using preferred codons increases the mRNA level. Because the distribution of the fitness effects of mutations is similar for synonymous and nonsynonymous mutations, it is difficult to explain $d_N/d_S << 1$ for most genes in almost all species. I proposed that nonsynonymous mutants have more variable fitness than synonymous mutants across environments, so nonsynonymous mutants are more likely to have low fitness in some environments and be purged in evolution when the environment fluctuates. As a result, nonsynonymous mutations are less likely to get fixed than synonymous mutations. I experimentally verified this hypothesis by measuring the fitness effects of mutations in three additional environments.

In Chapter 3, I sequenced the genomes of the BY4742 progenitor strain, wild-type control strain used the competition, 21 gene knockout strains, and ~28 mutant strains per gene. I found that in the genes with negligible effects of secondary mutations, the conclusions from Chapter 2 hold.

In Chapter 4, I took advantage of the double mutants produced by oligo synthesis errors and estimated intragenic epistasis between mutations in four environments. I found that synonymous mutations have pervasive epistatic interactions with synonymous or nonsynonymous mutations. I also found that 8.5% to 26.1% of epistatic interactions vary

significantly across environments. Furthermore, I found that epistasis between nonsynonymous mutations is more variable than that between synonymous mutations across environments. Epistasis between nonsynonymous mutations involves the interaction between protein sequence changes which synonymous mutations do not cause, so this adds to the complexity and variability of epistasis between nonsynonymous mutations.

References

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

Bateson, W. (1909). *Mendel's principles of heredity*. Cambridge University Press.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., . . . Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol Cell*, *61*(3), 341-351. https://doi.org/10.1016/j.molcel.2016.01.008

Buschauer, R., Matsuo, Y., Sugiyama, T., Chen, Y. H., Alhusaini, N., Sweet, T., . . . Beckmann, R. (2020). The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science*, *368*(6488). https://doi.org/10.1126/science.aay6912

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, *7*(2), 98-108. https://doi.org/10.1038/nrg1770

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., . . . Boone, C. (2010). The Genetic Landscape of a Cell. *Science*, *327*(5964), 425-431. https://doi.org/10.1126/science.1180823

Gillespie, J. H. (2004). *Population genetics : a concise guide* (2nd ed.). Johns Hopkins University Press.

Grantham, R., Gautier, C., & Gouy, M. (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res*, *8*(9), 1893-1912. https://doi.org/10.1093/nar/8.9.1893

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*, *9*(1), r43-74. https://doi.org/10.1093/nar/9.1.213-b

Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., & Press, W. H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet*, *11*(12), e1005732. https://doi.org/10.1371/journal.pgen.1005732

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. https://doi.org/10.1016/0022-2836(81)90003-6

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789-802. https://doi.org/10.1016/j.cell.2011.10.002

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624-626. https://doi.org/10.1038/217624a0

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, *164*(3881), 788-798. https://doi.org/10.1126/science.164.3881.788

Kristofich, J., Morgenthaler, A. B., Kinney, W. R., Ebmeier, C. C., Snyder, D. J., Old, W. M., . . . Copley, S. D. (2018). Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genet*, *14*(8), e1007615. https://doi.org/10.1371/journal.pgen.1007615

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science*, *324*(5924), 255-258. https://doi.org/10.1126/science.1170160

Li, C., Qian, W., Maclean, C. J., & Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, *352*(6287), 837-840. https://doi.org/10.1126/science.aae0568

Li, C., & Zhang, J. (2018). Multi-environment fitness landscapes of a tRNA gene. *Nat Ecol Evol*, *2*(6), 1025-1032. https://doi.org/10.1038/s41559-018-0549-8

Lind, P. A., Berg, O. G., & Andersson, D. I. (2010). Mutational Robustness of Ribosomal Protein Genes. *Science*, *330*(6005), 825-827. https://doi.org/10.1126/science.1194617

Mauger, D. M., Cabral, B. J., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., . . . McFadyen, I. J. (2019). mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci U S A*, *116*(48), 24075-24083. https://doi.org/10.1073/pnas.1908052116

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652-654. https://doi.org/10.1038/351652a0

Olson, C. A., Wu, N. C., & Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol*, *24*(22), 2643-2651. https://doi.org/10.1016/j.cub.2014.09.072

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet*, *5*(5), 389-396. https://doi.org/10.1038/nrg1327

Park, C., Chen, X., Yang, J. R., & Zhang, J. (2013). Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*, *110*(8), E678-686. https://doi.org/10.1073/pnas.1218066110

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., . . . Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111-1124. https://doi.org/10.1016/j.cell.2015.02.029

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., & Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. *Science*, *352*(6287), 840-844. https://doi.org/10.1126/science.aaf0965

Sharon, E., Chen, S.-A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, *175*(2), 544-557.e516. https://doi.org/10.1016/j.cell.2018.08.057

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol*, *24*(1-2), 28-38. https://doi.org/10.1007/BF02099948

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., . . . Stamatoyannopoulos, J. A. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, *342*(6164), 1367-1372. https://doi.org/10.1126/science.1243490

Sun, M., & Zhang, J. (2022). Preferred synonymous codons are translated more accurately: Proteomic evidence, among-species variation, and mechanistic basis. *Sci Adv*, *8*(27), eabl9812. https://doi.org/10.1126/sciadv.abl9812

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc Natl Acad Sci U S A*, *117*(7), 3528-3534. https://doi.org/10.1073/pnas.1907126117

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep*, *14*(7), 1787-1799. https://doi.org/10.1016/j.celrep.2016.01.043

Yang, J. R., Chen, X., & Zhang, J. (2014). Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol*, *12*(7), e1001910. https://doi.org/10.1371/journal.pbio.1001910

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C. H., Fu, J., . . . Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A*, *113*(41), E6117-E6125. https://doi.org/10.1073/pnas.1606724113

Chapter 2 Most Synonymous Mutations in 21 Representative Yeast Genes are Strongly Nonneutral

2.1 Abstract

Synonymous mutations in protein-coding genes do not alter protein sequences so are generally presumed neutral or nearly so (Graur et al., 2016; Kimura, 1968; King & Jukes, 1969; Li, 1997; Nei & Kumar, 2000). To experimentally verify this presumption, we constructed 8,341 yeast mutants each carrying a synonymous, nonsynonymous, or nonsense mutation in one of 21 endogenous genes with diverse functions and expression levels, and measured their fitness relative to the wild-type in a rich medium. Surprisingly, three-quarters of synonymous mutations reduce the fitness significantly, and the distribution of fitness effects is overall similar albeit nonidentical between synonymous and nonsynonymous mutations. We find that both synonymous and nonsynonymous mutations frequently disturb the mutated gene's mRNA level and that the extent of the disturbance partially predicts the fitness effect. Investigations in additional environments reveal greater across-environment fitness variations for nonsynonymous than synonymous mutants despite their similar fitness distributions in each environment, suggesting a smaller proportion of nonsynonymous than synonymous mutants that are always non-deleterious in a changing environment to permit fixation, potentially explaining substantially lower nonsynonymous than synonymous substitution rates commonly observed. The strong non-neutrality of most synonymous mutations, if true in diverse organisms, would require reexamining numerous biological conclusions about mutation, selection, effective population

size, divergence time, and disease mechanism that rely on the neutral assumption of synonymous mutations.

2.2 Introduction

The cracking of the genetic code in the 1960s revealed that between a quarter and a third of single nucleotide mutations in protein-coding genes do not alter protein sequences (Kimura, 1968; King & Jukes, 1969). Although these synonymous mutations are not strictly neutral because they could influence many processes (Chamary et al., 2006; Hershberg & Petrov, 2008; Plotkin & Kudla, 2011) such as transcription factor (TF) binding (Stergachis et al., 2013), transcription (Zhou et al., 2016), pre-mRNA splicing (Chamary et al., 2006), mRNA folding (Park et al., 2013) and stability (Chen et al., 2017; Presnyak et al., 2015), translational initiation (Kudla et al., 2009), efficiency (Frumkin et al., 2018; Qian, Yang, et al., 2012) , and accuracy (Akashi, 1994; Drummond & Wilke, 2008), and co-translational protein folding (Buhr et al., 2016; Walsh et al., 2020), the vast majority of them are presumed to be at least nearly neutral (Graur et al., 2016; Kimura, 1968; King & Jukes, 1969; Li, 1997; Nei & Kumar, 2000), contrasting nonsynonymous mutations, which alter protein sequences and frequently the fitness (Graur et al., 2016; Li, 1997; Nei & Kumar, 2000). The (near) neutrality of synonymous mutations is widely assumed in inferring mutation rate, pattern, and mechanism, testing natural selection, estimating effective population sizes (N_e) and neutral genetic diversities commonly considered in conservation policymaking in addition to population and evolutionary biology, and dating evolutionary events such as population or species divergences and gene or genome duplication (Graur et al., 2016; Li, 1997; Nei & Kumar, 2000). This assumption also diverts the mechanistic study of disease from synonymous mutations (Gilissen et al., 2012).

Nevertheless, synonymous mutations affecting the fitness by >1% are known (Agashe et al., 2013; Frumkin et al., 2018; Kristofich et al., 2018; Lebeuf-Taylor et al., 2019; Walsh et al., 2020). Some even reported comparable fitness effects of synonymous and nonsynonymous

mutations (Lind et al., 2010; Sharon et al., 2018; She & Jarosz, 2018). These reports, however, were based on either relatively few genes and mutations (Lind et al., 2010) or many natural polymorphisms (Sharon et al., 2018; She & Jarosz, 2018) that may not represent random mutations. Here we test the (near) neutrality of synonymous mutations by measuring the fitness effects of thousands of coding mutations in 21 genes in the budding yeast *Saccharomyces cerevisiae*.

2.3 Result

2.3.1 Quantifying mutational fitness effects

The 21 chosen genes participate in diverse biological processes such as metabolism, chromatin remodeling, transcription, translation, and cell wall synthesis (**Table 2-1**) and vary by 1000 times in their expression levels (**Fig. 2-1a**). These genes are nonessential but their deletions lower the fitness by discernable amounts(Qian, Ma, et al., 2012) such that the mutational fitness effects are quantifiable. In each gene, we picked an approximately 150-nucleotide coding sequence and chemically synthesized all 450 possible variants that deviate from the wild-type by a point mutation (**Fig. 2-1b**). The wild-type sequence at its native genomic location was replaced by the variant sequences using CRISPR/Cas9 genome editing of a haploid strain, followed by confirmation of the respiratory function of the mutant library (**Fig. A-1**). All mutants of a gene, together with a wild-type control that went through the same CRISPR/Cas9 editing (**Fig. A-2**), were competed *en masse* in a rich medium (YPD) at 30°C, with no diploidization observed (Fig. S3). Four separate competitions were performed using a common starting population (T_0), and the focal gene was respectively amplified from T_0 and the four replicate populations at 12 (T_{12}) and 48 (T_{48}) hrs, followed by 250-nucleotide paired-end Illumina sequencing (**Fig. 2-1b**). The sequences informed genotypes and allowed tabulating genotype frequencies in each population(Li et al., 2016).

For the 21 genes, we identified a total of 8,341 variants with read counts ≥50 at T_0, including 1,866 synonymous, 6,306 nonsynonymous, and 169 nonsense mutants, respectively. The observed relative numbers of synonymous and nonsynonymous mutants reflect those designed (**Fig. A-4**). Changes in genotype frequencies between T_0 and T_{48} (or T_{12}) were used to estimate the fitness of each mutant relative to the wild-type. The fitness estimates were highly

correlated between replicates, with a mean Pearson's *r* of 0.92 (**Fig. 2-1c**, **Fig. A-5**). Fitness estimates from the *en masse* competitions agreed well with those measured from monoculture growths for 24 reconstructed synonymous and nonsynonymous mutants (**Fig. 1-1d**).

2.3.2 Comparing mutational fitness effects

The median fitness of the 169 nonsense mutants is 0.940 (**Fig. A-6**). As expected, the corresponding value for the 6,306 nonsynonymous mutants is much higher, reaching 0.988 (**Fig. 2-2a**). Surprisingly, the median fitness of the 1,866 synonymous mutants is 0.989, much closer to that of nonsynonymous mutants than to the neutral expectation of 1; the same trend holds for mean fitness (**Fig. 2-2a**). While the fitness distributions look similar for synonymous and nonsynonymous mutants (**Fig. 2-2a**), they are statistically distinct due to a higher density of nonsynonymous than synonymous mutants in the fitness range of 0.91-0.97 but the reverse in the range of 0.97-0.99 (**Fig. 2-2b**, **Fig. A-7**). A significant fitness difference was observed between synonymous and nonsynonymous mutants in only five of the 21 genes, with all five exhibiting higher fitness for synonymous than nonsynonymous mutants (**Fig. 2-2c**, **Fig. A-8**). Even in these five genes, however, the median fitness of synonymous mutants is much closer to that of nonsynonymous mutants than to 1 (**Fig. 2-2c**).

Classifying each mutant into one of three bins based on whether its fitness is significantly below 1 (nominal $P < 0.05$, *t*-test), above 1, or neither, we found similar distributions for synonymous and nonsynonymous mutants (**Fig. 2-2d**). Among synonymous mutations, 75.9% are significantly deleterious while 1.3% are significantly beneficial. The corresponding values are 75.8% and 1.6% for nonsynonymous mutations. Slightly lower values were obtained at the false discovery rate (FDR) of 0.05 (Fig. 1-2d legend). The smallest absolute fitness effect found

significant in our study is 0.001, orders of magnitude greater than the sensitivity (10^{-7}) of natural selection in yeast (Chen & Zhang, 2021) (see Methods). Hence, all mutations with significant fitness effects are strongly nonneutral. Mutant fitness is lower when the mutation is unobserved in the genomes of related yeast species than when it is observed (**Fig. 2-2e**, **Fig. A-9**), indicating that our laboratory fitness estimates are evolutionarily relevant.

2.3.3 Mechanisms of mutational fitness effects

Because synonymous codon usage bias is stronger in more highly expressed genes probably due to translational selection (Hershberg & Petrov, 2008), synonymous mutations from the wild-type are thought to be more deleterious in more highly expressed genes (Plotkin & Kudla, 2011). However, we did not detect a significant negative correlation between the expression level of a gene and the mean fitness of its synonymous mutants (**Fig. 2-3**). Because synonymous mutations in a gene can alter its mRNA level (Chen et al., 2017; Presnyak et al., 2015; Zhou et al., 2016), which could affect fitness (Keren et al., 2016), we measured the relative expression level (*REL*) of the mutated gene in each mutant in four replicates by dividing its mRNA level by that of the wild-type. Briefly, from a population of cells including the wild-type and all mutants of a gene, we amplified and sequenced the DNAs of the focal gene as well as the cDNAs made from the mRNAs of the focal gene (**Fig. 1-4a**). *REL* is the number of cDNA-derived sequencing reads divided by the number of DNA-derived reads for a mutant, relative to that for the wild-type.

We obtained mutant *REL*s for 20 of the 21 genes. Mutant *REL*s are highly correlated between replicates (**Fig. A-10**), confirming the quality of the expression estimates. *REL* deviates significantly from 1 (nominal $P < 0.05$, *t*-test) in 53.8% of synonymous and 55.0% of

nonsynonymous mutants (39.7% and 39.6% at FDR <0.05, respectively), indicating that both synonymous and nonsynonymous mutations frequently alter the mRNA level. The *REL* distribution is not significantly different between synonymous and nonsynonymous mutants (**Fig. A-11**; **see Fig. A-12** for individual genes) and is more or less symmetrical around 1 (**Fig. 2-4b**). By contrast, the mean *REL* is only 0.301 for nonsense mutants (**Fig. A-13**), likely owing to nonsense-mediated mRNA decay (Chang et al., 2007).

Because reducing *REL* from 1 to 0, equivalent to gene deletion, has different fitness effects for different genes (Qian, Yang, et al., 2012), we rescaled mutant fitness F to $f = (F-F_0)/(1-F_0)$, where F_0 is the fitness of the strain lacking the focal gene. Consequently, $1-f$ measures the fitness effect of a mutation relative to that of deleting the focal gene, permitting analyzing the relationship between *REL* and fitness across mutants of different genes. *REL* and rescaled fitness are significantly positively correlated for both synonymous and nonsynonymous mutants under $REL < 1$, but the correlation is much weakened under $REL > 1$ (**Fig. 2-4c**). These observations suggest that influencing the mRNA level is likely a general mechanism underlying the fitness effects of coding mutations and that expression reduction from the wild-type level typically imposes a stronger fitness effect than the opposite (see Methods).

To understand how coding mutations impact the mRNA level, we identified TF-binding sites in the mutated region of each gene (Monteiro et al., 2020), but mutations within and outside TF-binding sites do not show significantly different magnitudes of expression effects (**Fig. A-14**).

Previous manipulative experiments showed that increasing the codon adaptation index (*CAI*) (Sharp & Li, 1987) of a gene through synonymous mutations can boost its mRNA level by slowing mRNA degradation (Chen et al., 2017; Presnyak et al., 2015; Radhakrishnan et al.,

2016) and perhaps enhancing transcription (Zhou et al., 2016). Because nonsynonymous mutations can also alter *CAI*, we computed the relative *CAI* (*rCAI*) of each mutant gene by dividing its *CAI* by that of the wild-type. Indeed, a significant positive correlation exists between *rCAI* and *REL* among synonymous mutants as well as among nonsynonymous mutants (**Fig. 2-4d**). The same is true between *rCAI* and rescaled fitness, especially under *rCAI* <1 (**Fig. A-15**).

Due to the increased prevalence of preferred codons in more highly expressed genes (Hershberg & Petrov, 2008), synonymous mutations decreasing *CAI* (**Fig. 2-5a**) and lowering the mRNA level (**Fig. 1-5b**) are both more abundant in more highly expressed genes. Similar trends are seen for nonsynonymous mutations (**Fig. 2-6a**, **Fig. 2-6b**), because a random nonsynonymous mutation from a preferred codon of an amino acid will likely arrive at a less preferred codon of another amino acid. Consequently, synonymous (**Fig. 2-5c**) and nonsynonymous (**Fig. 2-6c**) mutants of more highly expressed genes have lower mean rescaled fitness.

Because of the demand for mRNA folding strength (*MFS*) (Park et al., 2013), which is at least in part related to translational accuracy(Yang et al., 2014) and co-translational protein folding(Faure et al., 2016), a change in *MFS* caused by a coding mutation may affect fitness(Lind et al., 2010). Indeed, we found a significant positive correlation between the relative *MFS* of a mutant and its rescaled fitness among mutants with reduced *MFS* (**Fig. A-16**), although the correlation is substantially weaker than that between *REL* and rescaled fitness (**Fig. 2-4c**), suggesting that coding mutations' fitness effects are likely conferred more by their influences of the mRNA level than those of the mRNA folding strength.

2.3.4 Fitness effects across environments

Interspecific comparisons have shown that the nonsynonymous to synonymous substitution rate ratio (d_N/d_S) is substantially below 1 for most genes in almost all organisms (Graur et al., 2016; Li, 1997; Nei & Kumar, 2000) including yeast (Goncalves et al., 2011), indicating that the probability of fixation of nonsynonymous mutations is generally much lower than that of synonymous mutations in long-term evolution, seemingly at odds with their similar distributions of fitness effects (DFEs) observed here. One possible explanation is that the two DFEs are highly dissimilar in the range of absolute fitness effects undetectable by our method, which is generally below 5×10^{-3}. For example, when beneficial mutations are ignored as in the neutral theory (Kimura, 1983), if the fraction of nonsynonymous mutations with deleterious fitness effects smaller than the sensitivity of natural selection in yeast (10^{-7}) is 10% of the corresponding fraction of synonymous mutations, a d_N/d_S of ~0.1 will result. This hypothesis is, however, difficult to test because of the much lower sensitivity of experiments than natural selection.

We wondered whether the low d_N/d_S can also be caused by a difference between synonymous and nonsynonymous mutants in their fitness variation among environments (Gillespie, 1975; Lewontin & Cohen, 1969). Considering this variation is relevant because the fixation of a neutral mutation takes on average $4N_e$ generations (Kimura & Ohta, 1969), during which the environment is highly likely to have changed many times. In addition to influencing the mRNA level and/or mRNA folding strength that can exert a fitness effect, nonsynonymous mutations also alter the protein sequence and potentially function, which synonymous mutations do not. Because each of the molecular phenotypic effects could be environment-dependent, nonsynonymous mutants may naturally have a larger across-environment fitness variance than

synonymous mutants, especially given recent reports that amino acid substitutions often show environment-specific fitness effects (Chen & Zhang, 2020; Dandage et al., 2018; Flynn et al., 2020). Under the most extreme scenario, the fraction of deleterious mutations is identical between synonymous and nonsynonymous mutations in each environment, but the specific deleterious mutations vary across environments for nonsynonymous but not synonymous mutations. Consequently, when the environment of a population fluctuates within the typical fixation time, some synonymous mutations are never deleterious so may be fixed, while virtually every nonsynonymous mutation is deleterious under some environments so cannot be fixed, resulting in $d_N/d_S << 1$. We quantitatively investigated this model using computer simulation. Assuming the YPD-based DFEs in each environment, we varied the fitness of a mutant among environments with the coefficient of variation (*CV*) greater for nonsynonymous than synonymous mutants. A mutant is selectively purged if its fitness is lower than a preset cutoff (e.g., 0.99 given the fitness estimation error in our experiments) in any environment, and d_N/d_S is inferred from the fraction of nonsynonymous mutants unpurged relative to that of synonymous mutants unpurged. As predicted, d_N/d_S drops precipitously with the number of different environments experienced by the population (**Fig. 2-7a**, **Fig. A-17).**

To verify the key assumption on *CV* in the above model, we measured the DFEs of the same yeast synonymous and nonsynonymous mutations in three additional environments that differ in nutrient and stress, with three biological replicates per environment (**Fig. A-18 – A-23**). As in YPD, in each of these three environments, the median fitness of synonymous mutants is much closer to that of nonsynonymous mutants than to 1 (**Fig. 2-7b-d**) and 52.9-62.2% of synonymous mutants are significantly nonneutral (**Fig. A-24**). These fractions are lower than that in YPD likely because of the reduced sensitivity of our fitness measurement caused by the

use of fewer replicates (see Methods). For each mutant, we computed its *CV* in fitness across the four environments. Indeed, *CV* is significantly greater for nonsynonymous than synonymous mutants with ($P < 10^{-5}$) or without (**Fig. 2-7e**) the control of the mean fitness in the four environments (see Methods). Additionally, the fraction of neutral mutations in one environment that become deleterious in any of the other three environments is greater for nonsynonymous than synonymous mutations (**Fig. A-25**). We then used the empirical DFEs and fitness estimation errors in the four environments to estimate the expected d_N/d_S after purging mutants whose fitness is lower than a cutoff in any of the environments. Indeed, comparing the four populations respectively staying in one of the four constant environments with the fifth population whose environment fluctuates among the four conditions (see Methods), we found that, in terms of d_N/d_S, the fifth population is either significantly lower than or is not statistically distinguishable from the lowest of the first four (**Fig. 2-7f**). It is expected from the simulation result (**Fig. 2-7a**) that d_N/d_S in the fifth population will further decline as the number of different environments experienced rises.

2.4 Discussion

Our characterization of the DFE of thousands of coding mutations in diverse yeast genes under four environments showed that, under any environment, most synonymous mutations are strongly nonneutral and that the DFEs of synonymous and nonsynonymous mutations are overall similar. There is no particular reason why our results would be restricted to yeast, but confirmations in diverse organisms are required to verify the generality of our findings. Because our experiments were performed in haploids, future studies should assess whether synonymous and nonsynonymous mutations also have similar DFEs in the heterozygous state.

Our results suggest a general mechanism through which coding mutations affect fitness—disturbing the mRNA level of the mutated gene, but do not preclude other mechanisms such as impacting mRNA folding and translation. It is currently difficult to demonstrate and quantify the causal contributions of a coding mutation's various molecular phenotypic effects to its fitness effect, because this would require the difficult experiment of mimicking each molecular phenotypic effect of a coding mutation without disturbing the cell in any other aspect that might influence fitness. For instance, to mimic coding mutations' influences on the mRNA level of a gene, we could use an inducible promoter to drive gene expression and adjust the promoter activity by altering the concentration of the inducer in the medium(Azizoglu et al., 2021), but this alteration disturbs the medium composition, which could affect fitness more than through the inducible promoter. Additionally, the induction of the promoter may influence the expressions of neighboring genes. Use of tunable degrons, short amino acid sequences that regulate protein degradation (Natsume & Kanemaki, 2017), is another method, but degrons may also affect fitness by altering protein function or mRNA folding and tuning degrons could disturb the medium.

The mRNA level of a gene has a strong influence on the evolutionary rate of its protein sequence, and several mechanisms of this influence have been demonstrated (Wu et al., 2022; Zhang & Yang, 2015). Our finding that the fraction of nonsynonymous mutations reducing the mRNA level rises with the mRNA level of the gene (**Fig. 2-6b**) and the fitness ramification of this trend (**Fig. 2-6c**) suggest an additional mechanism (**Fig. 2-8**).

Because many biological conclusions rely on the presumption that synonymous mutations are (nearly) neutral (Graur et al., 2016; Li, 1997; Nei & Kumar, 2000), its invalidation has broad implications. For example, many tests infer selection on a gene by comparing its

synonymous and nonsynonymous polymorphisms and/or substitutions. Given that most synonymous mutations are deleterious, making the same inference would require assuming that synonymous and nonsynonymous mutations are subject to equal selections that are unrelated to protein sequence and function. While seemingly reasonable, this assumption may not always hold (Park et al., 2013), so further empirical verifications are needed. That most synonymous mutations are strongly nonneutral means that mutation rate, pattern, and mechanism inferred from synonymous polymorphisms or substitutions may have been distorted. For the same reason, N_e inferred from synonymous polymorphisms in natural populations is likely substantially underestimated, impacting evolutionary studies and certain conservation-related decisions. Similarly, synonymous substitution-based dating of evolutionary divergences may be unjustifiable in some cases. Our results also imply that synonymous mutations are nearly as important as nonsynonymous mutations in causing disease and call for strengthened effort in predicting and identifying pathogenic synonymous mutations (Sauna & Kimchi-Sarfaty, 2011). Given that gene expression anomaly can cause disease (Lee & Young, 2013), our results further suggest the disturbance of the mRNA level as a potentially common disease mechanism of coding mutations.

2.5 Materials and Methods

2.5.1 Data source

The mRNA expression levels of yeast genes in YPD (**Fig. 2-1a**) were from Chou *et al.* (Chou et al., 2017). The fitness values of yeast gene deletion strains under YPD (**Table. 2-1**) were from Qian *et al.* (Qian, Ma, et al., 2012). Yeast gene functions (**Table. 1-1**) were based on *Saccharomyces* Genome Database (https://www.yeastgenome.org/).

2.5.2 Media

Standard media of YPD (1% yeast extract, 2% peptone, and 2% glucose), YPD + 0.375 mM H_2O_2, YPE (1% yeast extract, 2% peptone, and 2% ethanol), and YPG (1% yeast extract, 2% peptone, and 2% glycerol) were used. Synthetic complete (SC) media contained 0.017% yeast nitrogen base without amino acids, 0.5% sulfate, and 2% glucose, with the addition of appropriate SC mix or SC drop-out mix. 5-FOA (5-fluoroorotic acid) plates contained 0.017% yeast nitrogen base without amino acids, 0.5% sulfate, 2% glucose, SC mix, and 0.15% 5-FOA.

2.5.3 Construction of yeast gene deletion strains

We had three primary considerations in choosing the genes for study. First, because a previous study of DFEs of synonymous and nonsynonymous mutations analyzed only two ribosomal protein genes (Lind et al., 2010), we wanted to include genes with a larger array of functions to complement that study. Second, knowing that synonymous mutations' fitness effects may depend on the gene expression level (Plotkin & Kudla, 2011), we wanted to choose genes with a wide range of expression levels to gain a broad picture. Third, because our experiment involved deleting the gene of choice, we must study nonessential genes.

Furthermore, the deletions must alter the fitness by detectable amounts such that the mutational fitness effects are quantifiable. The decision of using a 150-nucleotide region per gene was based on the read length of paired-end Illumina sequencing. The starting site of the 150-nucleotide region was randomly chosen in the first half of the coding sequence of a gene as long as the chosen 150 nucleotides are entirely within the coding region. Two exceptions were *RPL39* and *RPS7A*, where 147 nucleotides and 141 nucleotides were respectively studied because of these genes' short coding sequences.

For each chosen gene, we used CRISPR/Cas9 to delete from the genome of wild-type (BY4742) cells the 150-nucletide target sequence and its 25-nucleotide downstream sequence that would be used as a primer binding site to amplify the gene (see **Table. A-1** for all primer sequences). In the deletion step, the wild-type sequence was replaced by a 23-nucleotide designed sequence (20-nucleotide Cas9 target sequence plus 3-nucleotide PAM site) that would be used as the CRISPR/Cas9 recognition site in the mutant sequence insertion step. The deletion was then verified by Sanger sequencing.

2.5.4 Chemical synthesis of gene variants

For each gene, we had GENEWIZ (https://www.genewiz.com/en) synthesize in an oligo-mix format all 450 variants that each deviate from the wild-type by a single point mutation (except for *RPL39* that had 441 variants and *RPS7A* that had 423 variants due to their shorter sequences). With the exception of oligos for *RPL39* and *RPS7A*, each oligo has 200 nucleotides, including the 150-nucleotide target sequence and its 25-nucleotide upstream and 25-nucleotide downstream flanking sequences. The flanking sequences would serve as primer binding sites for

the amplification of the variant sequences. The guaranteed amount of each oligo was 3 nmol, more than enough as the DNA template for polymerase chain reaction (PCR) amplification.

2.5.5 Construction of mutant libraries

The pool of the synthesized single-strand variant oligos of each gene was amplified from the oligo-mix by PCR. High-fidelity Q5 polymerase (NEB) was used in all PCR reactions. The PCR-amplified double-stranded mutant sequences were transformed along with a CRISPR/Cas9 plasmid (pML104-URA3)(Laughery et al., 2015) into the strain with the wild-type gene deleted. The Cas9 protein would recognize the aforementioned 23-nucleotide sequence and cause double-stranded breaks. The variant sequences were inserted into the genome at the native genomic location of the focal gene via homologous recombination repair. For each gene, over 10,000 colonies were collected on SC minus uracil plates by washing with sterile water. The large number of colonies collected ensured the inclusion of most mutational variants of each gene. The variant cells were then counter-selected on the 5-FOA plates to get rid of the CRISPR/Cas9 plasmid. The cells were then stored in 30% glycerol at -80°C.

2.5.6 Construction of the wild-type control

We amplified the wild-type *ASC1* gene from the genome of the haploid strain BY4742 by PCR and inserted it into the *ΔASC1* cell using CRISPR/Cas9. Three colonies were picked and the insertion was confirmed by Sanger sequencing. The cells were then counter-selected on 5-FOA plates to remove the CRISPR/Cas9 plasmid. These three independently reconstituted wild-type strains (WT1, WT2, and WT3) were then stored in 30% glycerol at -80°C.

We measured the maximum growth rate of BY4742 and each of the three reconstituted wild-type strains using Biotek Gen5™ Microplate Reader. The cells were first grown overnight. About 5000 cells were added into 0.1 mL YPD in a well of a Costar™ 96-well plate, which was in continuous shaking at 30°C. Sixteen replicate growth curves were collected per strain, except that one replicate of BY4742 was contaminated so was discarded. The maximum growth rate was calculated following a previous protocol (Warringer et al., 2003). The maximum growth rate was not significantly variable among the four strains (**Fig. A-2**). For instance, the maximum growth rate of WT1 was not significantly different from that of WT2, WT3, or BY4742 (**Fig. A-2**). WT1 was used as the wild-type control in *en masse* competitions and mutant fitness estimation. Our results would remain virtually the same should the growth rate of WT2 or WT3 be used in mutant fitness calculation.

2.5.7 En masse competitions in YPD

A frozen sample of cells carrying the variants of a gene and a frozen sample of the wild-type control cells were revived at 30°C in YPD (with shaking at 250 RPM) for 3 hrs. These cells were then mixed in an approximately 1:50 ratio of wild-type control cells to all mutant cells combined (i.e., the population should contain about 2% wild-type control cells). Four replicate competitions were then started by dilution of this common starting population into four 14 mL Falcon tubes, each containing 6 mL of YPD medium. Upon dilution, the cell density of the starting population was 1×10^5 cells/mL. The competition was performed in a shaking incubator (250 RPM) at 30°C. Every 12 hrs, the cell culture was diluted to 1×10^5 cells/mL by transferring to 6 mL fresh YPD. The competition lasted for 48 hrs. The population aliquots at 0 (T_0), 12

(T_{12}), and 48 (T_{48}) hrs were stored in 30% glycerol at -80°C. We performed a total of 84 competitions for the 21 genes (4 × 21).

2.5.8 Library preparation and Illumina sequencing

Genomic DNA was extracted from population aliquots (Masterpure^{TM} Yeast DNA Purification Kit), followed by amplification of gene variants by PCR. One primer was targeted at the 25-nucleotide sequence immediately downstream of the mutated region while the other primer was annealed upstream of the mutated region beyond the homologous recombination repair sequence. This design ensured that only those variant sequences that were inserted at the native genomic location of the focal gene were amplified. The primers included Illumina sequencing adapter and i5/i7 index sequences. The amplicons were sequenced by 250-nucleotide paired-end Illumina sequencing (HiSeq2500). Paired reads for variant sequences were required to be identical to be counted. To ensure relative accuracy in fitness estimation, we considered only those genotypes with at least 50 read pairs in T_0.

2.5.9 Sequencing-based fitness estimation

We estimated the fitness of each mutant relative to the wild-type control by $(P'_{MT}P_{WT})/(P_{MT}P'_{WT})^{(1/G)}$, where P_{MT} and P_{WT} are the respective frequencies of the mutant and wild-type control at the beginning of the competition, P'_{MT} and P'_{WT} are the corresponding frequencies at the end of the competition, and G is the number of generations of the wild-type control in the competition and equals 7.25 for 12 hrs and 29 for 48 hrs. In theory, the above formula works in an *en masse* competition under the assumption of no strain-strain interaction,

as was confirmed by our computer simulation. The strong correlation between mutant fitness estimated from *en masse* competition and that estimated from monoculture growth (**Fig. 1-1d**) supports the assumption of no strain-strain interaction. To estimate G, we first allowed a frozen sample of wild-type control cells to revive at 30°C in YPD at 250 RPM for 3 hrs. We then started a monoculture of the wild-type control at 1×10^5 cells/mL in 6 mL of YPD. The growth continued for 12 hrs in a shaking incubator (250 RPM) at 30°C. We then estimated G in the 12 hrs based on the culture's optical density change. G in 48 hrs is 4 times G in 12 hrs. Mutant fitness is estimated more accurately with longer competitions. However, if the fitness of a mutant was so low that the strain disappeared in T_{48}, we calculated the fitness using T_0 and T_{12}; otherwise, we used T_0 and T_{48}. Note that only for 36 mutants were the fitness estimated using T_{12} instead of T_{48}. Based on four biological replicates, we used a t-test to determine if the fitness of a mutant deviates from 1 at the nominal P-value of 5%. The average standard error of the estimated mutant fitness was 0.005, considered as the mean detection limit of our fitness measurement. The absolute value of the smallest fitness effect with nominal $P < 0.05$ was 0.001. It has been estimated based on the level of synonymous polymorphism that N_e is approximately 10^7 in *S. cerevisiae*(Chen & Zhang, 2021), suggesting that natural selection can detect a fitness effect of 10^{-7} or greater in yeast. However, if most synonymous mutations are deleterious, as the present study shows, the actual N_e would be greater than 10^7 and natural selection more sensitive than considered in this study.

2.5.10 Verifying the respiratory function of mutants

Cells from each mutant library were first serially diluted. Equal numbers of cells were then spread on YPD and YPG plates, where respiratory functions were respectively unneeded

and needed for cell growth. We allowed cell growth for two days on YPD and three days on YPG, because of faster cell growth with glucose as the carbon source. Colonies were then counted on each plate. This experiment was repeated three times for the mutant library of each gene. BY4742 was used as a positive control in the respiratory function test. As a negative control, we simultaneously deleted *TOM6* and *TOM7* from BY4742, because TOM6 and TOM7 are components of the TOM (translocase of outer membrane) complex that is responsible for import of mitochondrially directed proteins and is important for respiration(Honlinger et al., 1996).

2.5.11 Quantifying ploidy after competition

One T_{48} population for each gene was randomly chosen and examined for ploidy. Approximately 10^7 cells were collected, washed with 1.5 mL of water, and fixed by a gentle addition of 3.5 mL of 95% ethanol and incubation for 2 hrs at room temperature. Fixed cells were collected by centrifugation for 15 s at 10,000*g*, followed by resuspension of the pellet in 1 mL water and transfer to a 1.5-mL microcentrifuge tube. After a brief centrifugation, we re-suspended cells in 0.5 mL RNase solution (2 mg/mL RNase A in 50 mM Tris pH 8.0, 15 mM NaCl, boiled for 15 min and then cooled to room temperature) and incubated the cells for at least 2 hrs at 37°C. We then collected cells from the RNase solution by centrifugation for 15 s at 10,000*g*. Cells were incubated in 0.2 mL protease solution (5 mg/mL pepsin and 4.5 µl/mL concentrated HCl in H_2O) for 20 min at 37°C and then collected by centrifugation. Cells were re-suspended in 0.5 mL 50 mM Tris pH 7.5, and were either stored at 4°C for a few days or analyzed immediately. For analysis, 50 µl of cell suspension was transferred to 1 mL of 1 µM SYTOX Green staining solution. All samples were analyzed using iQue Screener Plus flow

cytometry. First, we used the forward scatter area and side scatter area with a clustering package to remove non-cell particles. Second, we used forward scatter area and forward scatter height to remove doublets. Third, we plotted DNA content histograms of the distribution of the amount of DNA per cell. We used haploid (BY4742) and diploid (BY4743) yeast cells as controls to determine ploidy. In each of these two control profiles, there are two peaks, respectively representing cells in the G1 and G2/M cell-cycle stages (1C and 2C DNA content for haploids and 2C and 4C for diploids).

2.5.12 Impact of PCR and sequencing errors

The following error analysis followed Li *et al.* (Li et al., 2016). The error rate for Illumina sequencing is 3×10^{-4} per site per read (http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf). Thus, due to sequencing error, a genotype is expected to lose $U = [1-(1-3\times10^{-4})^{2\times150}]M_0$ read pairs, where M_0 is the true number of read pairs of the genotype and 150 is the sequence length considered. Because the fractional loss $U/M_0 = 0.086$ is a constant for all genotypes including the wild-type in each sample, the loss of reads due to sequencing error does not affect fitness estimation. Sequencing error also causes the genotype to gain on average $V = (3\times10^{-4}/3)^2M_1 = 10^{-8}M_1$ read pairs, where M_1 is the total number of read pairs for all neighbors of the focal genotype (i.e., the genotypes that differ from the focal genotype by one nucleotide). Thus, the fractional gain of read pairs for the genotype is expected to be $V/M_0 = 10^{-8}M_1/M_0$, which has virtually no impact on fitness estimation in our study. For instance, at T_0, M_1/M_0 is expected to be 50 for the wild-type and 11 for any mutant. Hence, the fractional gain of read pairs is $<10^{-6}$ for any genotype.

We similarly estimated the impact of PCR error. Q5 DNA polymerase used in PCR has a very low error rate of 5.3×10^{-7} per nucleotide incorporated (Potapov & Ong, 2017). The PCR used in sequencing library preparation had 25 cycles. Thus, due to PCR error, a genotype is expected to lose $U = (5.3\times10^{-7}\times150\times25)M_0$ molecules, where M_0 is the true number of DNA molecules of the genotype, 150 is the sequence length in nucleotides, and 25 is the number of PCR cycles. Because the fractional loss $U/M_0 = 0.002$ is a constant for all genotypes in each sample, the loss of molecules due to PCR error does not affect fitness estimation. PCR error also causes the genotype to gain on average $V = (5.3\times10^{-7}\times25/3)M_1 = 4.4\times10^{-6}M_1$ molecules, where M_1 is the total number of molecules for all neighbors of the focal genotype. Thus, the fractional gain of molecules for the genotype is expected to be $V/M_0 = 4.4\times10^{-6}M_1/M_0$, which has little impact on fitness estimation in our study. As mentioned, at T_0, M_1/M_0 is expected to be 50 for the wild-type and 11 for any mutant. Hence, the fractional gain in the number of molecules is 2.2×10^{-4} for the wild-type and 4.9×10^{-5} for any mutant.

2.5.13 Growth curve-based fitness estimation of reconstructed mutants

We used maximum growth rates estimated from monoculture growth curves to verify the mutant fitness estimated by *en masse* competition followed by sequencing. We chose nine synonymous mutants of *RPL29*, *RAD6*, or *RPS7A* and 15 nonsynonymous mutants of *TSR2*, *RAD6*, *RPS7A*, or *BUD23* with relatively large ranges of sequencing-based fitness estimates. We resynthesized these gene variants and remade the corresponding mutant strains. Using the method described earlier for measuring the growths of reconstituted wild-type strains, we measured the growth curves of each of these mutants as well as the wild-type control on the same 96-well plate, with eight replicates per strain. The relative fitness of a mutant was

calculated by $F = 2^{\text{relative growth rate-1}}$, where the relative growth rate is the maximum growth rate of the mutant divided by that of the wild-type control. The maximum growth rate was calculated following a previous protocol (Warringer et al., 2003). The above formula of F is derived as follows. Let r be the mutant growth rate and R be the wild-type growth rate. Let T be the wild-type generation time. By definition, mutant fitness relative to the wild-type (per generation) is $F = e^{rT}/e^{RT}$. Hence, $\ln F = (r\text{-}R)T$. Because by definition $e^{RT} = 2$, $T = (\ln 2)/R$. Combining the above two equations yielded $\ln F = (r\text{-}R)(\ln 2)/R = (r/R\text{-}1)\ln 2$. Therefore, $F = 2^{r/R\text{-}1} = 2^{\text{relative growth rate-1}}$. If mutant cells do not divide so that its population growth rate is 0, the mutant fitness relative to the WT is 0.5. If the mutation kills cells in addition to preventing mitosis, the mutant population growth rate is negative (i.e., the population shrinks), which would lead to a mutant fitness that is lower than 0.5.

CRISPR/Cas9 could generate off-target mutations. However, the high fitness correlation (**Fig. 1d**) between two independently constructed sets of 24 mutants suggests that this potential off-target effect did not influence our result.

2.5.14 Identifying orthologs of the 21 S. cerevisiae genes in five other yeast species

To examine whether a mutation examined in *S. cerevisiae* is present in the genomes of other yeast species, we attempted to identify the orthologs of the 21 genes studied in our experiment in *S. paradoxus*, *S. mikatae*, *S. uvarum*, *S. castellii*, and *Candida glabrata*, all of which diverged from *S. cerevisiae* after the whole-genome duplication in yeast. We retrieved genomic coding sequence (CDS) data from the NCBI genome assembly database (https://www.ncbi.nlm.nih.gov/assembly/) if they are available (*S. paradoxus*, *C. glabrata,* and *S. castellii*); otherwise, we retrieved genomic DNA data (*S. mikatae* and *S. uvarum*) from the same

database. For species with CDS data, we built a local blast library and performed tblastn using protein sequences of the 21 genes from *S. cerevisiae* as query sequences. The E-value threshold was set at 10^{-10}. If there was a full-length-query match, the matched subject was recorded as an ortholog. If the query was partially matched to the subject, the subject was inspected manually to ensure the orthologous relationship. For each species and gene, only the hit with the lowest E-value was examined to prevent the inclusion of paralogs. For species with genomic DNA data, we similarly built a local blast library and performed tblastn under the same E-value threshold. If there was a full-length-query match, the matched subject sequence was recorded as an ortholog. If the query was partially matched to the subject (likely due to introns), the matched subject sequence was extended 100-2000 bp upstream and downstream to ensure that it included all exons of the gene; the exact length of the extension was determined manually based on the length of the unmatched part of the query as well as genomic structure. We then used AUGUSTUS(Stanke & Morgenstern, 2005) to predict the coding region of the gene in the extended subject sequence, and manually inspected the sequence to ensure the orthologous relationship. We successfully identified almost all orthologs of the 21 genes in the five yeast species, except for *EST1*, for which we only identified an ortholog in *S. paradoxus*. We therefore excluded *EST1* from the downstream analysis. We also failed to identify the *EOS1* ortholog in *S. castellii* and *IES6* ortholog in *S. mikatae*, but decided to include these two genes in downstream analysis except for the missing species. The orthologous coding sequences of the six yeasts were then aligned using MACSE v2 (Ranwez et al., 2018). A mutation examined in *S. cerevisiae* is considered observed in the other yeasts if it appears in the genome of any of the other five yeasts and if no other nucleotide difference from *S. cerevisiae* exists in that genome in the codon harboring the mutation; otherwise, it is considered unobserved.

2.5.15 Estimating the mRNA levels of mutated genes

A frozen sample of cells carrying the variants of a focal gene and a frozen sample of the wild-type control cells were revived at 30°C in YPD with shaking at 250 RPM for 3 hrs. These cells were then mixed in an approximately 1:50 ratio of wild-type control cells to all mutant cells combined. Four replicate cultures were then started by diluting this common starting population into four 14 mL Falcon tubes, each containing 6 mL of YPD medium. The cell density of the starting population was 1×10^5 cells/mL. When the cells were in the log phase after 12 hrs of growth at 30°C in a shaking incubator (250 RPM), we extracted DNA and RNA from the cell cultures (Masterpure™ Yeast DNA Purification Kit and RNeasy Mini Kit, respectively). The mRNA of the focal gene was reverse transcribed (SuperScript® III First-Strand Synthesis System for RT-PCR) using about 20 nucleotides within the 25-nucleotide sequence immediately downstream of the variant sequence as the gene-specific primer.

We amplified the mutant gene segments by 25 cycles of PCR from genomic DNA and cDNA, respectively. The cDNA libraries of *EST1* were not successfully amplified, which may be because *EST1* has the lowest expression level among the 21 genes studied (**Fig. 1-1a**). As described earlier, one primer was targeted within the 25-nucleotide sequence downstream of the variant sequence while the other primer was upstream of the variant sequence and beyond the homologous recombination repair sequence. There were Illumina-adapter and i5/i7 index sequences on the primers. The amplicons were subjected to 250-nucleotide paired-end Illumina sequencing (NovaSeq). Paired reads for variant gene sequences must be identical to be counted. To ensure accuracy in expression estimation, we excluded genotypes with fewer than 50 read pairs from the genomic DNA.

The relative mRNA expression level (*REL*) of a mutant is the number of cDNA-derived read pairs divided by the number of DNA-derived read pairs for the mutant, relative to the corresponding value of the wild-type control. We estimated the *REL* for 7,795 mutants with fitness estimates in YPD. With the four replicates in *REL* estimation, we used a *t*-test to determine if the *REL* of a mutant significantly deviates from 1 at a nominal *P*-value of 5%. Virtually identical results were obtained when *REL* was first log-transformed before the *t*-test.

Following the sequencing and PCR error analyses presented earlier, we estimated the impact of reverse transcription errors on *REL* estimation. The reverse transcriptase used is a version of M-MLV RT, with an error rate of 4×10^{-5} per nucleotide incorporated (https://www.thermofisher.com/us/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/rt-education/reverse-transcriptase-attributes.html). Due to reverse transcription error, a genotype is expected to lose $U = (4\times10^{-5}\times150)M_0$ molecules, where M_0 is the expected number of cDNA molecules of the genotype and 150 is the sequence length. Because the fractional loss $U/M_0 = 0.006$ is a constant for all genotypes in each sample, the loss of molecules due to reverse transcription error does not affect expression estimation. Reverse transcription error also causes the genotype to gain on average $V = 4\times10^{-5}/3\ M_1 = 1.3\times10^{-5}M_1$ molecules, where M_1 is the expected total number of cDNA molecules for all neighbors of the focal genotype. Thus, the fractional gain of molecules for the genotype is expected to be $V/M_0 = 1.3\times10^{-5}M_1/M_0$, which has little impact on expression estimation in our study. M_1/M_0 is expected to be about 50 for the wild-type and 11 for mutants whose expression levels are comparable with that of the wild-type. The corresponding fractional gains of molecules are 6.5×10^{-4} and 1.4×10^{-4}, respectively. Even if a mutant has a *REL* as low as 0.1, M_1/M_0 is 110 and the fractional gain of the number of molecules is 1.4×10^{-3}. As described,

PCR and sequencing errors had virtually no effect. Hence, the overall error from reverse transcription, PCR, and sequencing is negligible in expression estimation.

In addition to correlating mutant *REL* with rescaled fitness (**Fig. 1-4c**), we used a linear mixed model to assess the relative importance of *REL* and mutation type (synonymous vs. nonsynonymous) to rescaled fitness, with gene identity added as a random effect. We separately analyzed mutants with *REL* <1 and those with *REL* >1, because of their apparently different relationships with rescaled fitness (**Fig. 1-4c**). For mutants with *REL* <1, the fraction of variance of rescaled fitness explained by *REL* is 61.5% ($P < 2.2\times10^{-16}$), while that explained by mutation type is only 0.2% ($P = 0.0002$). For mutants with *REL* >1, the fraction of variance of rescaled fitness explained by *REL* is 7.4% ($P < 2.2\times10^{-16}$), while that explained by mutation type is only 0.4% ($P = 1.5\times10^{-7}$). These results demonstrate that *REL* explains a substantially larger fraction of variance of rescaled fitness than does mutation type.

Additionally, after accounting for gene-specific effects using a mixed-effect model, we found the positive correlation between the rescaled fitness and *REL* to remain significant when *REL* <1 ($P = 2.3\times10^{-47}$). There is a marginally significant negative correlation between the rescaled fitness and *REL* when *REL* >1 ($P = 0.048$). We also attempted to fit a quadratic model using $\log_2(REL)$ as an independent variable and accounted for a random effect of gene identity. Indeed, the hypothesis that the fitness peak is at *REL* = 1 could not be rejected.

2.5.16 Codon adaption index (CAI)

We computed *CAI* for the entire coding sequence of each wild-type or mutant gene, using previously reported yeast relative synonymous codon usage (*RSCU*) estimates(Sharp & Li,

1987), which are highly correlated with those derived from the 200 most highly expressed genes (r = 0.995) (Qian, Yang, et al., 2012).

2.5.17 mRNA folding strength (MFS)

The minimum free energy at 30°C was calculated for each wild-type or mutant mRNA sequence using RNAfold in ViennaRNA (2.4.17) with default parameters except for the temperature(Hofacker et al., 1994). We define mRNA folding strength (*MFS*) as the absolute value of the minimum free energy.

2.5.18 TF-binding sites

TF-binding sites were searched in the wild-type for the 150-nucleotide target sequence plus the 20-nucleotide flanking sequence on each side using the database Yeastract(Monteiro et al., 2020).

2.5.19 DFE estimation in SC + 37°C, YPD + 0.375mM H_2O_2, and YPE

The experiment followed that in DFE estimation in YPD, except that the competitions lasted for 20 generations (cells were transferred 6.5 and 13 generations after the start of the competition) and had three replicates per environment. Sequencing library preparation was unsuccessful for mutants of *EST1* and *PAF1* likely because of primer degradations. Therefore, we acquired the fitness data of mutants of 19 genes in these three additional environments. The fraction of mutants whose fitness is significantly different from 1 is lower here than in YPD, likely because of the reduced statistical power due to the lowered number of replications.

Indeed, when we randomly sampled three of the four replicates from YPD, the fraction of mutants whose fitness is significantly different from 1 (nominal P <0.05) decreased to an average of 0.63 and 0.64 for synonymous and nonsynonymous mutants, respectively, similar to those observed in these three additional environments **(Fig. A-24**). To examine whether the difference between synonymous and nonsynonymous mutants in fitness *CV* across the four environments is entirely due to a potential difference in mean fitness, we controlled the mean fitness in the four environments when comparing the across-environment fitness *CV* between synonymous and nonsynonymous mutants. Specifically, we used an identity index of 0 for each synonymous mutant and 1 for each nonsynonymous mutant. The partial Spearman's correlation between the identity index and *CV* upon the control of the mean fitness in the four environments is 0.052 ($P = 7.7\times10^{-6}$).

2.5.20 Simulation of the impact of environmental changes on d_N/d_S

Our simulation assumed that the DFEs of synonymous and nonsynonymous mutations estimated from YPD hold in each environment, but the fitness effect of a mutation can vary across environments. We respectively constructed cumulative fitness distribution functions (CFDFs) of synonymous and nonsynonymous mutants from the corresponding fitness data collected in YPD. We started from all synonymous mutants with fitness measured in YPD, and ranked these mutants from low to high by their YPD fitness. We then added a random noise drawn from the normal distribution $N(0, \sigma^2)$ to each fitness value, and ranked the mutants by their new fitness values. Let us assume that, after the addition of noise, the mutant originally ranked i now had a rank of j. We then randomly sampled M synonymous mutants from the CFDF and ranked them by their fitness, where M is the number of synonymous mutants with

fitness measured in YPD. We assigned the fitness of the mutant ranked the *j*th in these *M* sampled mutants to mutant *i* as its fitness in a new environment. The above procedure was repeated for each environment considered. Fitness *CV* among environments was controlled by adjusting σ^2, with larger σ^2 yielding greater *CV*. Many σ^2 values were tried to achieve a target *CV* (difference between observed and target *CV* <0.0001). The same was done for nonsynonymous mutants. We set a higher *CV* for nonsynonymous than synonymous mutants. We set a fitness cutoff (0.98 or 0.99) and assumed that any mutant with fitness below the cutoff in any environment was purged. We then computed d_N/d_S by the fraction of unpurged nonsynonymous mutants divided by the fraction of unpurged synonymous mutants. Under each parameter set, we repeated the simulation 1000 times and reported the mean d_N/d_S and its 95% confidence interval.

2.5.21 Expected d_N/d_S in the four environments examined

To predict the expected d_N/d_S in long-term evolution in each of the four environments where DFEs were measured here, we considered all of the synonymous and nonsynonymous mutants with fitness measured in the environment. Because the fitness measures contained measurement errors, we added a random error term drawn from the normal distribution $N(0, \sigma_{se}^2)$ to the measured fitness, where σ_{se} is the mutant-specific standard error of the measured fitness estimated from the experimental replicates in the environment. We set a fitness cutoff and assumed that any mutant with fitness in the environment below the cutoff was purged. We then computed d_N/d_S by the fraction of unpurged nonsynonymous mutants divided by the fraction of unpurged synonymous mutants. In an environment that varies among the four individual conditions, we assumed that any mutant with fitness below the cutoff in any condition was

purged. Because of random measurement errors considered, we repeated the prediction 1000 times and presented the 95% confidence interval of the predicted d_N/d_S.

2.5.22 Data availability

Sequencing data generated in this study have been deposited into NCBI with the Bioproject ID PRJNA750109. Public data used include gene function annotations in the *Saccharomyces* Genome Database (https://www.yeastgenome.org/) and genomic coding sequences of *S. paradoxus*, *C. glabrata*, and *S. castellii* and genomic sequences of *S. mikatae* and *S. uvarum* from the NCBI genome assembly database (https://www.ncbi.nlm.nih.gov/assembly/). Source data are provided with this paper.

2.5.23 Code availability

Custom code is available at https://github.com/song88180/ Mutational-Fitness-Effects and https://doi.org/10.5281/ zenodo.5908478.

2.6 References

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol*, *30*(3), 549-560. https://doi.org/10.1093/molbev/mss273

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

Azizoglu, A., Brent, R., & Rudolf, F. (2021). A precisely adjustable, variation-suppressed eukaryotic transcriptional controller to enable genetic discovery. *Elife*, *10*. https://doi.org/10.7554/eLife.69549

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., . . . Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol Cell*, *61*(3), 341-351. https://doi.org/10.1016/j.molcel.2016.01.008

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, *7*(2), 98-108. https://doi.org/10.1038/nrg1770

Chang, Y. F., Imam, J. S., & Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, *76*, 51-74. https://doi.org/10.1146/annurev.biochem.76.050106.093909

Chen, P., & Zhang, J. (2020). Antagonistic pleiotropy conceals molecular adaptations in changing environments. *Nat Ecol Evol*, *4*(3), 461-469. https://doi.org/10.1038/s41559-020-1107-8

Chen, P., & Zhang, J. (2021). Asexual Experimental Evolution of Yeast Does Not Curtail Transposable Elements. *Mol Biol Evol*, *38*(7), 2831-2842. https://doi.org/10.1093/molbev/msab073

Chen, S., Li, K., Cao, W., Wang, J., Zhao, T., Huan, Q., . . . Qian, W. (2017). Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. *Mol Biol Evol*, *34*(11), 2944-2958. https://doi.org/10.1093/molbev/msx229

Chou, H. J., Donnard, E., Gustafsson, H. T., Garber, M., & Rando, O. J. (2017). Transcriptome-wide Analysis of Roles for tRNA Modifications in Translational Regulation. *Mol Cell*, *68*(5), 978-992 e974. https://doi.org/10.1016/j.molcel.2017.11.002

Dandage, R., Pandey, R., Jayaraj, G., Rai, M., Berger, D., & Chakraborty, K. (2018). Differential strengths of molecular determinants guide environment specific mutational fates. *PLoS Genet*, *14*(5), e1007419. https://doi.org/10.1371/journal.pgen.1007419

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, *134*(2), 341-352.

Faure, G., Ogurtsov, A. Y., Shabalina, S. A., & Koonin, E. V. (2016). Role of mRNA structure in the control of protein folding. *Nucleic Acids Res*, *44*(22), 10898-10911. https://doi.org/10.1093/nar/gkw671

Flynn, J. M., Rossouw, A., Cote-Hammarlof, P., Fragata, I., Mavor, D., Hollins, C., 3rd, . . . Bolon, D. N. (2020). Comprehensive fitness maps of Hsp90 show widespread environmental dependence. *Elife*, *9*. https://doi.org/10.7554/eLife.53810

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci U S A*, *115*(21), E4940-E4949. https://doi.org/10.1073/pnas.1719375115

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet*, *20*(5), 490-497. https://doi.org/10.1038/ejhg.2011.258

Gillespie, J. H. (1975). Natural selection for within-generation variance in offspring number II. Discrite haploid models. *Genetics*, *81*(2), 403-413.

Goncalves, P., Valerio, E., Correia, C., de Almeida, J. M., & Sampaio, J. P. (2011). Evidence for divergent evolution of growth temperature preference in sympatric Saccharomyces species. *PLoS One*, *6*(6), e20739. https://doi.org/10.1371/journal.pone.0020739

Graur, D., Sater, A. K., & Cooper, T. F. (2016). *Molecular and genome evolution*. Sinauer Associates, Inc.

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annu Rev Genet*, *42*, 287-299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast Folding and Comparison of Rna Secondary Structures. *Monatshefte Fur Chemie*, *125*(2), 167-188. https://doi.org/Doi 10.1007/Bf00818163

Honlinger, A., Bomer, U., Alconada, A., Eckerskorn, C., Lottspeich, F., Dietmeier, K., & Pfanner, N. (1996). Tom7 modulates the dynamics of the mitochondrial outer membrane translocase and plays a pathway-related role in protein import. *EMBO J*, *15*(9), 2125-2137.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., . . . Segal, E. (2016). Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell*, *166*(5), 1282-1294 e1218. https://doi.org/10.1016/j.cell.2016.07.024

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res*, *11*(3), 247-269. https://doi.org/10.1017/s0016672300011459

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, M., & Ohta, T. (1969). The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics*, *61*(3), 763-771.

King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, *164*(881), 788-798.

Kristofich, J., Morgenthaler, A. B., Kinney, W. R., Ebmeier, C. C., Snyder, D. J., Old, W. M., . . . Copley, S. D. (2018). Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genet*, *14*(8), e1007615. https://doi.org/10.1371/journal.pgen.1007615

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science*, *324*(5924), 255-258. https://doi.org/324/5924/255 [pii]
10.1126/science.1170160

Laughery, M. F., Hunter, T., Brown, A., Hoopes, J., Ostbye, T., Shumaker, T., & Wyrick, J. J. (2015). New vectors for simple and streamlined CRISPR-Cas9 genome editing in Saccharomyces cerevisiae. *Yeast*, *32*(12), 711-720. https://doi.org/10.1002/yea.3098

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife*, *8*. https://doi.org/10.7554/eLife.45952
Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237-1251. https://doi.org/10.1016/j.cell.2013.02.014
Lewontin, R. C., & Cohen, D. (1969). On population growth in a randomly varying environment. *Proc Natl Acad Sci U S A*, *62*(4), 1056-1060. https://doi.org/10.1073/pnas.62.4.1056
Li, C., Qian, W., Maclean, M., & Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, *352*, 837-840.
Li, W. (1997). *Molecular Evolution*. Sinauer.

Lind, P. A., Berg, O. G., & Andersson, D. I. (2010). Mutational Robustness of Ribosomal Protein Genes. *Science*, *330*(6005), 825-827. https://doi.org/10.1126/science.1194617
Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., . . . Teixeira, M. C. (2020). YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Res*, *48*(D1), D642-D649. https://doi.org/10.1093/nar/gkz859
Natsume, T., & Kanemaki, M. T. (2017). Conditional Degrons for Controlling Protein Expression at the Protein Level. *Annu Rev Genet*, *51*, 83-102. https://doi.org/10.1146/annurev-genet-120116-024656
Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.
Park, C., Chen, X., Yang, J. R., & Zhang, J. (2013). Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*, *110*(8), E678-686. https://doi.org/10.1073/pnas.1218066110
Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, *12*(1), 32-42. https://doi.org/10.1038/nrg2899
Potapov, V., & Ong, J. L. (2017). Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*, *12*(1), e0169774. https://doi.org/10.1371/journal.pone.0169774
Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., . . . Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111-1124. https://doi.org/10.1016/j.cell.2015.02.029
Qian, W., Ma, D., Xiao, C., Wang, Z., & Zhang, J. (2012). The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell Rep*, *2*(5), 1399-1410. https://doi.org/10.1016/j.celrep.2012.09.017
Qian, W., Yang, J. R., Pearson, N. M., Maclean, C., & Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet*, *8*(3), e1002603. https://doi.org/10.1371/journal.pgen.1002603
PGENETICS-D-11-02081 [pii]
Radhakrishnan, A., Chen, Y. H., Martin, S., Alhusaini, N., Green, R., & Coller, J. (2016). The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell*, *167*(1), 122-132 e129. https://doi.org/10.1016/j.cell.2016.08.053
Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol Biol Evol*, *35*(10), 2582-2584. https://doi.org/10.1093/molbev/msy159

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet*, *12*(10), 683-691. https://doi.org/10.1038/nrg3051

Sharon, E., Chen, S. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, *175*(2), 544-557 e516. https://doi.org/10.1016/j.cell.2018.08.057

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, *15*(3), 1281-1295. https://doi.org/10.1093/nar/15.3.1281

She, R., & Jarosz, D. F. (2018). Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change. *Cell*, *172*(3), 478-490 e415. https://doi.org/10.1016/j.cell.2017.12.015

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*, *33*(Web Server issue), W465-467. https://doi.org/10.1093/nar/gki458

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., . . . Stamatoyannopoulos, J. A. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, *342*(6164), 1367-1372. https://doi.org/10.1126/science.1243490

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc Natl Acad Sci U S A*, *117*(7), 3528-3534. https://doi.org/10.1073/pnas.1907126117

Warringer, J., Ericson, E., Fernandez, L., Nerman, O., & Blomberg, A. (2003). High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci U S A*, *100*(26), 15724-15729. https://doi.org/10.1073/pnas.2435976100

Wu, Z., Cai, X., Zhang, X., Liu, Y., Tian, G. B., Yang, J. R., & Chen, X. (2022). Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nat Ecol Evol*, *6*(1), 103-115. https://doi.org/10.1038/s41559-021-01578-x

Yang, J. R., Chen, X., & Zhang, J. (2014). Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol*, *12*(7), e1001910. https://doi.org/10.1371/journal.pbio.1001910

Zhang, J., & Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nat Rev Genet*, *16*(7), 409-420. https://doi.org/10.1038/nrg3950

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C. H., Fu, J., . . . Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A*, *113*(41), E6117-E6125. https://doi.org/10.1073/pnas.1606724113

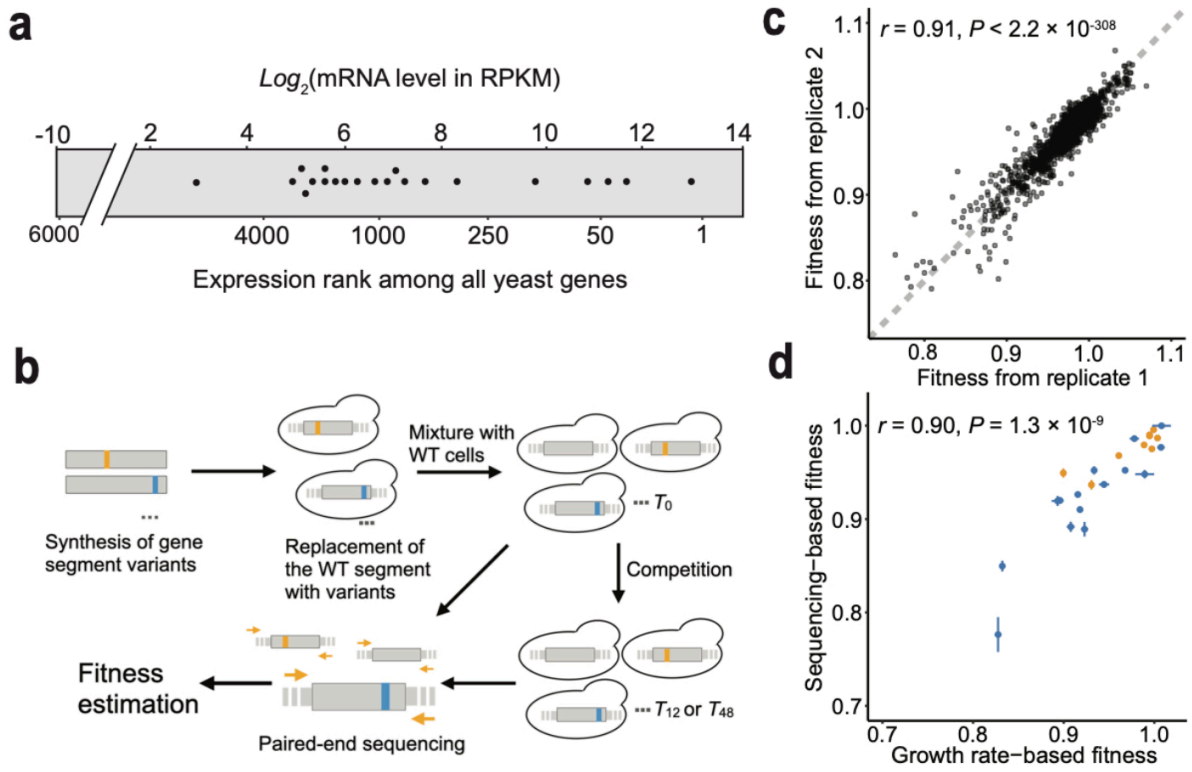

Figure 2-1. Estimating the fitness effects of coding mutations in 21 yeast genes.a, The mRNA expression levels in YPD of the 21 genes (dots) measured by RPKM (Reads Per Kilobase of transcript per Million mapped reads) and their ranks among all yeast genes. **b**, Experimental procedure. WT, wild-type. T_0, T_{12}, and T_{48} respectively refer to 0, 12, and 48 hrs after competition. **c**, Mutant fitness estimated in the first two of four biological replicates. Each dot is a mutant (n = 8,341 mutants) and the dotted line indicates the diagonal. Pearson's correlation (r) and its associated P-value are presented. **d**, Sequencing-based and growth rate-based fitness estimates are highly correlated. Each dot represents a synonymous (yellow) or nonsynonymous (blue) mutant. Mutants used in monoculture growth rate-based fitness estimation and those used in *en masse* competition followed by sequencing-based fitness estimation are independently constructed. Error bars show the standard error of the mean. Pearson's correlation r and its associated P-value are presented (r = 0.89 and 0.90 for the 9 synonymous and 15 nonsynonymous mutants, respectively).

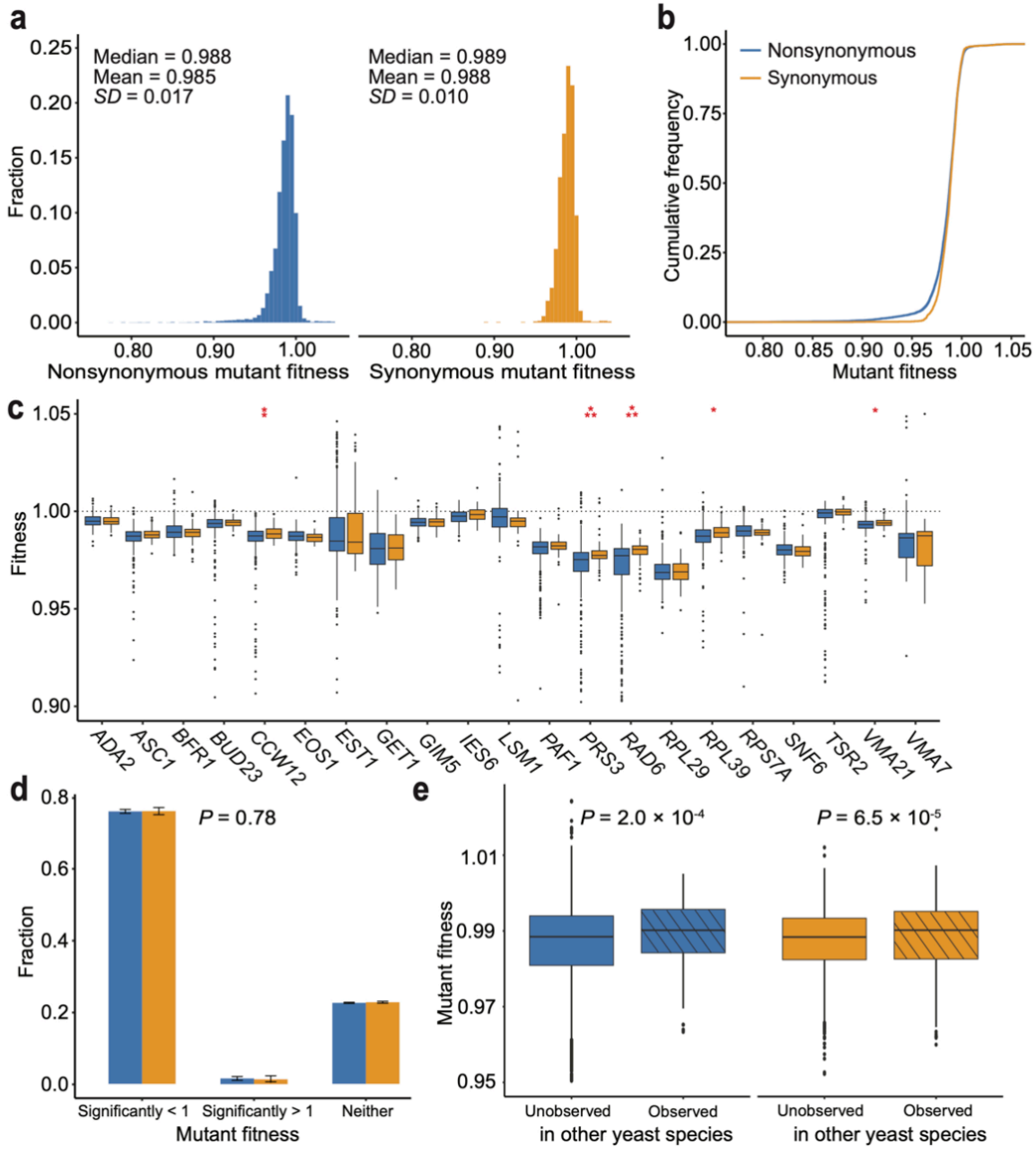

Figure 2-2. Mutant fitness in YPD. Distributions of the fitness of 6,306 nonsynonymous (blue) and 1,866 synonymous (yellow) mutants. The two distributions are significantly different ($P = 6.1\times10^{-5}$, two-tailed Wilcoxon rank-sum test; $P = 1.3\times10^{-6}$, Kolmogorov–Smirnov test). **b**, Cumulative frequency distributions of fitness of nonsynonymous and synonymous mutants. **c**, Fitness distributions of nonsynonymous and synonymous mutants of 21 individual genes shown by box plots. Nonsynonymous and synonymous distributions of each gene are compared by a two-tailed Wilcoxon rank-sum test followed by FDR correction (*, $P < 0.05$; ⁑, $P < 0.01$, ⁂, $P < 0.001$). Mutants with fitness <0.9 are not shown (see **Fig. A-8** for the complete figure). **d**,

Fractions of nonsynonymous and synonymous mutants with fitness significantly below 1 (nominal P <0.05), significantly above 1, and neither, respectively. Error bars show one standard error. Nonsynonymous and synonymous mutants are not significantly differentially distributed among the three bins (two-tailed Fisher's exact test). Under FDR = 0.05, 72.7% and 1.5% of nonsynonymous mutations are significantly deleterious and beneficial, respectively. The corresponding values are 72.5% and 1.1% for synonymous mutations. **e**, Mutant fitness is lower when the mutation is not observed than when it is observed in the genomes of five related yeast species. There are 5839, 169, 1087, 714 mutants in the four bins, respectively. *P*-values are from two-tailed Wilcoxon rank-sum test. Mutants with fitness <0.95 or >1.025 are not shown (see **Fig. A-9** for the complete figure). In c and e, each data point is a mutant. The lower and upper edges of a box represent the first (qu_1) and third (qu_3) quartiles, respectively, the horizontal line inside the box indicates the median (*md*), the whiskers extend to the most extreme values inside inner fences, $md \pm 1.5(qu_3\text{-}qu_1)$, and the dots show outliers.

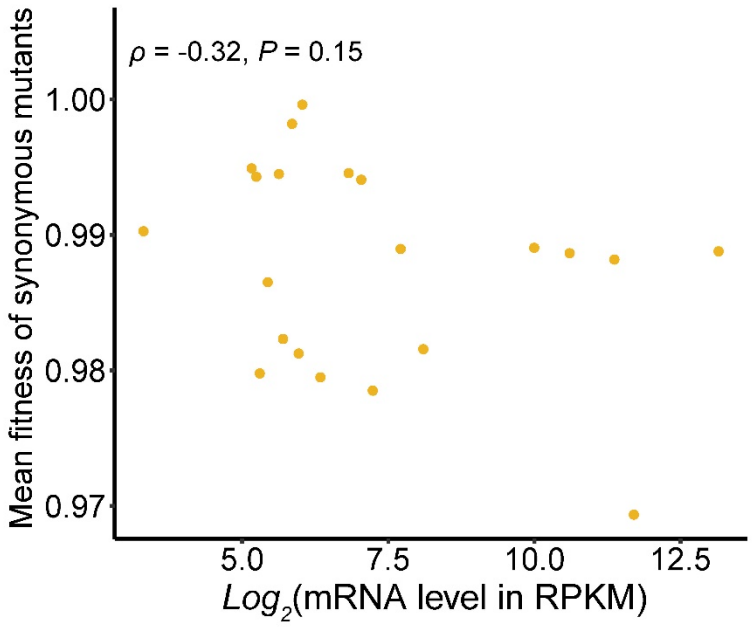

Figure 2-3. Non-significant negative correlation between the mean fitness of synonymous mutants of a gene and the expression level of the gene. Each dot represents a gene. Spearman's correlation ρ and associated *P*-value are presented.

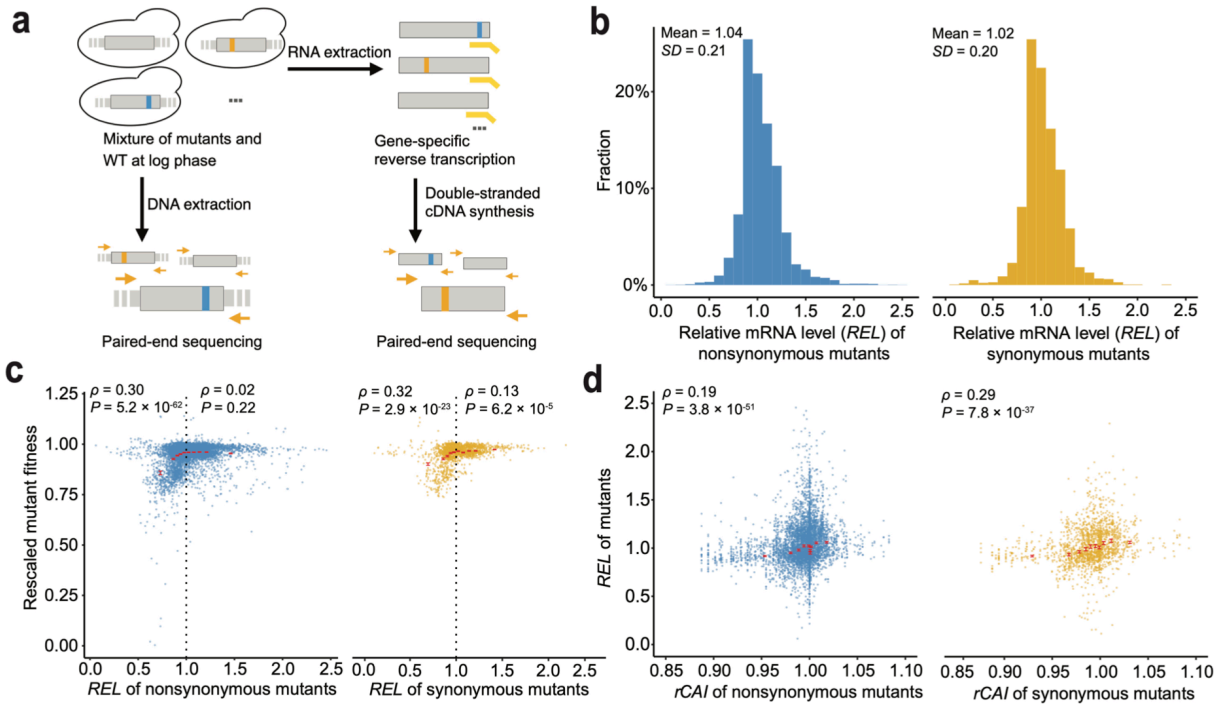

Figure 2-4. Coding mutations alter mRNA level of the mutated gene. **a**, High-throughput quantification of the mRNA levels of a focal gene in all mutants of the gene. WT, wild-type. *REL*, the mRNA level in a mutant relative to that in the WT, is estimated from the number of cDNA-derived sequencing reads divided by the number of DNA-derived reads for the mutant, relative to that for the WT. **b**, Frequency distributions of *REL* for 5927 nonsynonymous (blue) and 1783 synonymous (yellow) mutants, respectively. The two distributions are not significantly different ($P = 0.11$, two-tailed Wilcoxon rank-sum test). **c**, Correlation between *REL* and rescaled fitness among mutants. The correlation is significantly different between mutants with *REL* <1 and >1 ($P < 0.0001$ for both nonsynonymous and synonymous mutants based on *z*-test after Fisher's *r*-to-*z* transformation). **d**, Positive correlation between *rCAI*, the *CAI* of a mutant relative to that of the wild-type, and *REL* among mutants. For visualization, in c and d, we group all mutants into 10 equal-size bins by their X-values and present the mean X- and Y-values of each bin (red dot) and the standard error of the mean Y-value (error bar).

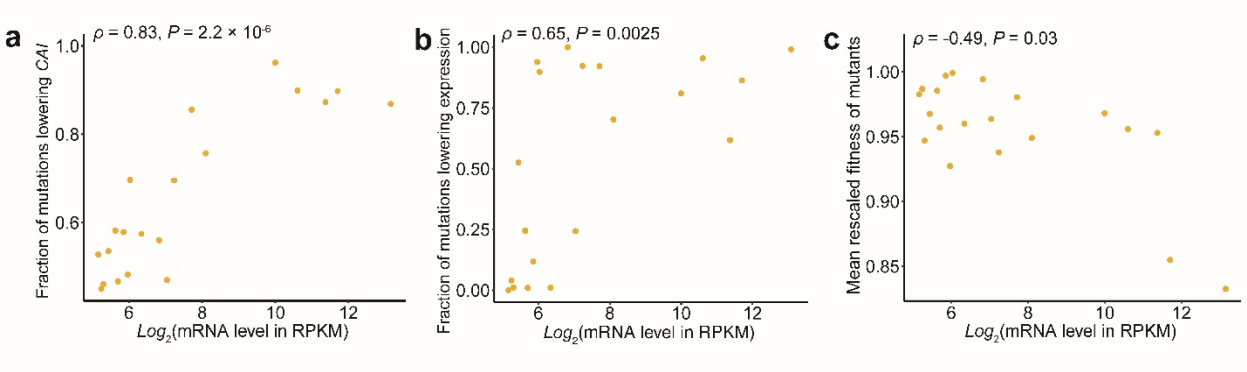

Figure 2-5. Relationship between the mRNA level of a gene and the effects of synonymous mutations in the gene on CAI, expression level, and rescaled fitness. **a**, Fraction of mutations lowering *CAI* increases with the expression level of the gene. **b**, Fraction of mutations lowering the expression level increases with the expression level of the gene. **c**, Mean rescaled fitness of mutants declines with the expression level of the gene. Each dot represents a gene. Spearman's correlation (ρ) and associated *P*-value are presented.

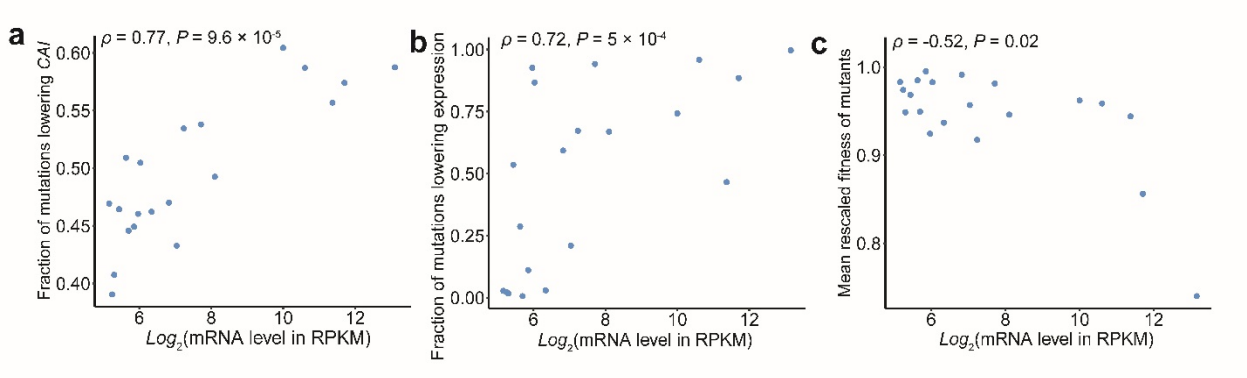

Figure 2-6. Relationship between the mRNA level of a gene and the effects of nonsynonymous mutations in the gene on CAI, expression level, and rescaled fitness. a, Fraction of mutations lowering *CAI* increases with the expression level of the gene. **b**, Fraction of mutations lowering the expression level increases with the expression level of the gene. **c**, Mean rescaled fitness of mutants declines with the expression level of the gene. Each dot represents a gene. Spearman's correlation (ρ) and associated *P*-value are presented. Because deleting a more highly expressed gene tends to cause a greater fitness reduction[56], the present finding means that the mean fitness reduction caused by a nonsynonymous mutation should rise with the expression level of the gene.

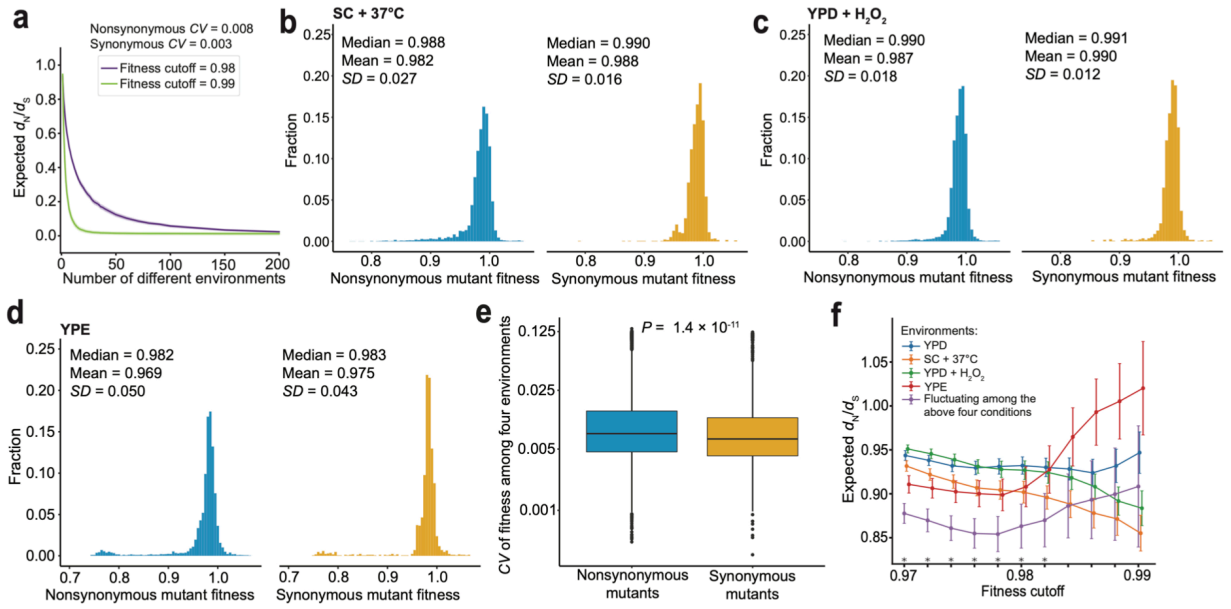

Figure 2-7. A higher fitness CV across environments for nonsynonymous than synonymous mutants can create dN/dS <<1 despite similar DFEs of synonymous and nonsynonymous mutations in each environment. a, Expected d_N/d_S from 1000 simulations of a population that experiences multiple different environments. A mutant is purged if its fitness is below a preset cutoff such as 0.98 or 0.99 in any environment. Shaded areas represent 95% confidence intervals. **b-d**, Distributions of nonsynonymous and synonymous mutant fitness are significantly different in SC + 37°C ($P = 1.8\times10^{-12}$, two-tailed Wilcoxon rank-sum test; $P = 1.5\times10^{-9}$, Kolmogorov–Smirnov test; **b**), YPD + 0.375 mM H_2O_2 ($P = 1.9\times10^{-7}$ and 7.0×10^{-8}, respectively; **c**), and YPE ($P = 9.9\times10^{-5}$ and 2.9×10^{-9}, respectively; **d**). **e**, Box plots showing distributions of fitness *CV* across the four environments for 5,671 nonsynonymous and 1,696 synonymous mutants. Box plot symbols follow those in Fig. 2e. The mean *CV* is 0.0163 for nonsynonymous and 0.0124 for synonymous mutants. The two distributions are significantly different (two-tailed Wilcoxon rank-sum test). **f**, Expected d_N/d_S when the population stays in a constant environment or a changing environment. Actual DFEs in the four individual environments are used and various fitness cutoffs as in panel a are considered. Fitness measurement error is considered through 1000 random samples of error per mutant. The mean expected d_N/d_S and the 95% confidence interval of the expected d_N/d_S are presented. Dots and error bars are slightly shifted horizontally to help visualization. * indicates that d_N/d_S is significantly lower in the fifth population, whose environment fluctuates among the four conditions, than in each of the four constant-environment populations (P <0.05). For the cutoffs where no * is shown, d_N/d_S is not significantly different between the fifth population and the constant-environment population with the lowest d_N/d_S.

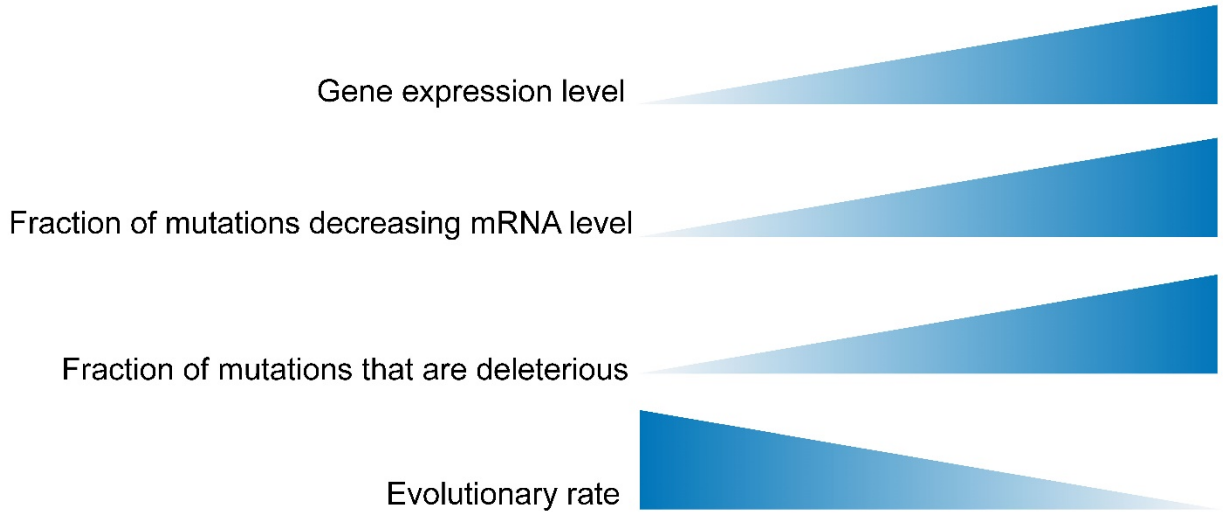

Figure 2-8. A new model explaining the negative correlation between the evolutionary rate of a protein and its mRNA level. Compared with nonsynonymous mutations in lowly expressed genes, those in highly expressed genes tend to reduce the gene expression level and hence tend to be deleterious. As a result, the protein evolutionary rate is negatively correlated with its gene expression level. The height of a symbol represents the quantity considered.

Table 2- 1. Properties of the 21 genes studied.

Gene name	Expression level (RPKM)	Fitness of the gene deletion strain in YPD	Coding sequence length (bp)	Region subject to mutation (nt.)
ADA2	35.81	0.70	1305	304 - 453
ASC1	2647.43	0.75	960	13 - 162
BFR1	209.76	0.43	1413	376 - 525
BUD23	37.87	0.57	828	205 - 354
CCW12	9113.10	0.93	402	84 - 234
EOS1	43.38	0.58	1101	193 - 342
EST1	9.94	0.54	2100	238 - 387
GET1	62.56	0.74	708	201 - 350
GIM5	49.52	0.62	492	117 - 266
IES6	57.87	0.39	501	130 - 279
LSM1	113.24	0.69	519	58 - 207
PAF1	51.98	0.59	1338	292 - 441
PRS3	150.72	0.66	963	214 - 263
RAD6	81.02	0.49	519	70 - 219
RPL29	3334.86	0.79	180	7 - 156
RPL39	1023.08	0.65	156	7 - 153
RPS7A	1557.77	0.74	573	4 - 144
SNF6	39.46	0.62	999	217 - 266

TSR2	65.40	0.57	618	46 - 195
VMA21	131.56	0.84	234	38 - 187
VMA7	274.51	0.65	357	52 - 201

(Continued)

Gene name	Observed no. of nonsynonymous mutants	Observed no. of synonymous mutants	Observed no. of nonsense mutants
ADA2	330	92	3
ASC1	269	104	2
BFR1	336	90	24
BUD23	302	100	0
CCW12	320	114	16
EOS1	323	99	0
EST1	298	65	23
GET1	341	83	26
GIM5	332	98	1
IES6	325	102	12
LSM1	124	26	7
PAF1	324	105	13
PRS3	285	93	12
RAD6	312	94	0
RPL29	336	93	21
RPL39	323	82	7
RPS7A	305	94	1
SNF6	346	89	1
TSR2	325	89	0

VMA21	309	115	0
VMA7	141	39	0

(Continued)

Gene name	Gene function
ADA2	Transcription coactivator; component of the ADA and SAGA transcriptional adaptor/HAT (histone acetyltransferase) complexes
ASC1	G-protein beta subunit and guanine dissociation inhibitor for Gpa2p; ortholog of RACK1 that inhibits translation; core component of the small (40S) ribosomal subunit; required to prevent frameshifting at ribosomes stalled at repeated CGA codons; regulates P-body formation induced by replication stress; represses Gcn4p in the absence of amino acid starvation; controls phosphorylation of multiple proteins
BFR1	Component of mRNP complexes associated with polyribosomes; involved in localization of mRNAs to P bodies; implicated in secretion and nuclear segregation; multicopy suppressor of BFA (Brefeldin A) sensitivity
BUD23	Methyltransferase that methylates residue G1575 of 18S rRNA; required for rRNA processing and nuclear export of 40S ribosomal subunits independently of methylation activity; functions with DEAH-box RNA helicase Ecm16p
CCW12	Cell wall mannoprotein; plays a role in maintenance of newly synthesized areas of cell wall
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

EST1	TLC1 RNA-associated factor involved in telomere length regulation; recruitment subunit of telomerase
GET1	Subunit of the GET complex; involved in insertion of proteins into the ER membrane
GIM5	Subunit of the heterohexameric cochaperone prefoldin complex
IES6	Component of the INO80 chromatin remodeling complex; critical for INO80 function; involved in regulation of chromosome segregation and maintenance of normal centromeric chromatin structure
LSM1	Lsm (Like Sm) protein; forms heteroheptameric complex (with Lsm2p, Lsm3p, Lsm4p, Lsm5p, Lsm6p, and Lsm7p) involved in degradation of cytoplasmic mRNAs
PAF1	Component of the Paf1p complex involved in transcription elongation
PRS3	5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase; synthesizes PRPP, which is required for nucleotide, histidine, and tryptophan biosynthesis
RAD6	Ubiquitin-conjugating enzyme (E2); involved in postreplication repair as a heterodimer with Rad18p, regulation of K63 polyubiquitination in response to oxidative stress, DSBR and checkpoint control as a heterodimer with Bre1p, ubiquitin-mediated N-end rule protein degradation as a heterodimer with Ubr1p, ERAD with Ubr1p in the absence of canonical ER membrane ligases, and Rpn4p turnover as part of proteasome homeostasis, in complex with Ubr2p and Mub1p
RPL29	Ribosomal 60S subunit protein L29
RPL39	Ribosomal 60S subunit protein L39
RPS7A	Protein component of the small (40S) ribosomal subunit

SNF6	Subunit of the SWI/SNF chromatin remodeling complex; involved in transcriptional regulation; functions interdependently in transcriptional activation with Snf2p and Snf5p; relocates to the cytosol under hypoxic conditions
TSR2	Protein with a potential role in pre-RNA processing
VMA21	Integral membrane protein required for V-ATPase function; not an actual component of the vacuolar H+-ATPase (V-ATPase) complex; diverged ortholog of human XMEA (X-linked Myopathy with Excessive Autophagy); functions in the assembly of the V-ATPase; localized to the yeast endoplasmic reticulum (ER)
VMA7	Subunit F of the V1 peripheral membrane domain of V-ATPase; part of the electrogenic proton pump found throughout the endomembrane system; required for the V1 domain to assemble onto the vacuolar membrane; the V1 peripheral membrane domain of vacuolar H+-ATPase (V-ATPase) has eight subunits

Chapter 3 Experimental Validation of the Non-neutrality of Synonymous Mutation

3.1 Abstract

In Chapter 2, we found that about three-quarters of synonymous mutations in 21 endogenous yeast genes are nonneutral in YPD. There is a hypothesis that the observation could be due to the fitness effects of CRISPR/Cas9 off-target edits or secondary mutations in the experiment. To test this hypothesis, we sequenced the genomes of BY4742 progenitor strain, 21 gene knockout strains, *ASC1* wild-type control strain, and 579 mutant strains (on average ~28 per gene). By comparing the mutations identified and the potential off-target sites of the gRNAs used, we confirmed the lack of any off-target edits. The mutations identified are thus likely to be natural spontaneous mutations that occurred in the experiments. We found that such mutations are either absent or have negligible fitness effects in the mutants of seven genes studied. In these genes, 61.6% of synonymous mutations are significantly nonneutral in YPD. We further confirmed all other observations in Chapter 2 for these seven genes.

3.2 Introduction

We recently constructed 8,341 mutants each carrying a synonymous, nonsynonymous, or nonsense mutation in one of 21 *Saccharomyces cerevisiae* genes (Shen et al., 2022). We found that most synonymous and most nonsynonymous mutants are significantly less fit than the wild-type control, although nonsynonymous mutations are overall more detrimental than synonymous mutations (Shen et al., 2022). It has been suggested that our observations may have arisen from the fitness effects of potential CRISPR/Cas9 off-target edits and/or secondary mutations (Kruglyak et al., 2023). To assess this hypothesis, we sequenced the genomes of relevant strains to find potential off-target edits or secondary mutations.

3.3 Results

3.3.1 Identifying off-target edits and secondary mutations in relevant genomes

To construct mutants in Chapter 2, we performed two rounds of CRISPR/Cas9 genome editing (Shen et al., 2022). In the first round, a segment of the wild-type sequence was replaced by an artificially designed landing pad (DLP) (**Fig. 3-1a**). In the second round, the landing pad was replaced with variant sequences (**Fig. 3-1b**). To identify potential off-target edits or secondary mutations from the first round of editing, we sequenced the genomes of the BY4742 progenitor strain and 21 gene knockout strains. To find potential off-target edits or secondary mutations which emerged in the second round of editing, we sequenced ~28 randomly picked mutants of each gene. Note that, upon the second round of editing, each mutant is made up of multiple (~25 on average) independently edited cells in our mutant pool, and the sequencing-based fitness of the mutant is the average fitness of these cells.

CRISPR/Cas9 off-target edits are mutations in untargeted regions that bear a high sequence similarity to the target region. In our study, gene-deletion gRNAs were designed using Benchling (www.benchling.com/crispr) to minimize potential off-target editing. Specifically, 20 of the 21 gRNAs in the first round of editing were regarded by Benchling as good guides while the remaining one approached the cutoff for good guides (**Table. 3-1**). Benchling could also predict potential off-target sites of a gRNA based on sequence similarity (**Fig. 3-2**). None of the identified mutations occurred in the potential off-target sites. The gRNA used in the second round of editing is especially good because the Cas9 cutting site was created artificially and Benchling predicted no potential off-target sites in the yeast genome.

The wild-type control used in the competition in our study in Chapter 2 was created by replacing the wild-type *ASC1* gene with the landing pad, followed by the replacement of the

landing pad with the *ASC1* wild-type sequence. We found that this wild-type control carried no mutations relative to the BY4742 progenitor strain (**Table. 3-2**, **Table. 3-3**).

In 7 genes (*ADA2*, *ASC1*, *BFR1*, *EOS1*, *IES6*, *RPL39* and *TSR2*), no more than two mutant strains carried secondary mutations when compared with the respective gene knockout strains, with typically no more than one secondary mutation per mutant (**Table. 3-3**). The average fitness effect of mutations in protein coding regions is about -0.01 (in the 21 genes studied) (Shen et al., 2022) and each mutant genotype consisted of ~25 independently edited cells. Hence, if fewer than 10% of cells of a genotype carry a secondary mutation of a mean fitness effect of -0.01, the secondary mutations lower the fitness of the genotype by no more than 0.001, which is much smaller than the mean fitness effect of mutations measured as well as the mean standard error of the fitness effect estimates. In other words, this level of error is negligible.

Of the 7 genes examined, the gene knockout strains of *BFR1*, *EOS1*, and *IES6* each harbored 2, 1, and 1 secondary mutations, respectively, while the knockout strains of the other four genes had no secondary mutations. We inserted the wild-type sequences into these three knockout strains and found their maximum growth rates not significantly different from the wild-type control previously used in our competition (**Fig. 3-3**). Hence, the secondary mutations in the knockout strains of these three genes have negligible fitness effects in YPD. This is not surprising, because the 4 secondary mutations occurred in genes whose individual deletions have an average fitness cost of 0.06 (Qian et al., 2012). By contrast, the average fitness cost of individual deletions of *BFR1*, *EOS1*, and *IES6* is 0.53 while the corresponding value for the 21 genes studied is 0.36 (Qian et al., 2012).

3.3.2 The results in Ch. 2 are replicated for the seven genes

The mean fitness of 2231 nonsynonymous mutants of the aforementioned seven genes is 0.991 (**Fig. 3-4a**). The mean fitness of 658 synonymous mutants of these genes is 0.992 (**Fig. 3-4a**), which is much closer to that of the nonsynonymous mutants than to the neutral expectation of 1. However, overall, nonsynonymous mutants have significantly lower fitness than synonymous mutants (**Fig. 3-4b**). We classified all mutations into three bins: significantly beneficial, significantly deleterious, and neutral (i.e., neither of the above categories) (**Fig. 3-4c**). Among synonymous mutations, 61.6% are significantly deleterious, while 1.4% are significantly beneficial. The corresponding values are 63.0% and 0.9% for nonsynonymous mutations. Consistent with our previous finding (Shen et al., 2022), mutant fitness is lower when the mutation is unobserved in the genomes of related yeast species than when it is observed (**Fig. 3-4d, Fig. B-1**), indicating that our laboratory fitness estimates are evolutionarily relevant.

In Chapter 2, we found that *REL* and rescaled fitness are significantly positively correlated for both synonymous and nonsynonymous mutants when $REL < 1$ but the correlation is much weakened when $REL > 1$, suggesting that reducing gene expression from the wild-type level is more likely to be deleterious than increasing gene expression. This pattern was also observed in these seven genes (**Fig. 3-5**).

We previously found that the *CV* of fitness across four environments is significantly greater for nonsynonymous than synonymous mutants. We used simulations to show that environmental changes could explain $d_N/d_S << 1$ (**Fig. 2-7**). Using YPD-based fitness effects of the mutations of the aforementioned 7 genes, we replicated this result (**Fig. 3-6a**). Among the seven genes, *ADA2*, *ASC1*, *RPL39*, and *TSR2* do not carry secondary mutations in the knockout

strains so their multi-environment mutant fitness estimates are valid. For these 4 genes, the nonsynonymous mutants have larger fitness variances across the four environments than those of the synonymous mutants (**Fig. 3-6b**). d_N/d_S in the population whose environment fluctuates among the four environments is lower than any population which stays in one of the four environments (**Fig. 3-6c**).

3.4 Discussion

Consistent with previous reports from yeast studies (Jakociunas et al., 2015; Ryan et al., 2014), we did not find any CRISPR/Cas9 off-target editing in a total of 601 strains, including 21 knockout strains, 579 mutant strains, and 1 reconstructed wild-type control strain. Although CRISPR/Cas9 off-target edits are known in plants and animals, the rate of such edits is low and well-designed experiments can avoid off-target edits even in the human genome (Cho et al., 2014). The genome is 250 times smaller and the efficiency of non-homologous end-joining relative to homologous recombination (a predictor of off-target editing) is drastically lower in yeast than in humans, rendering off-target editing trivial in yeast (Jakociunas et al., 2015; Ryan et al., 2014).

Our results showed that secondary mutations can occur during mutant construction. After the first editing, the 21 gene knockout strains are known to have low fitness, so secondary mutations compensating the deleterious effect of gene deletion would be positively selected. However, such mutations are beneficial, so they must become detrimental after the second editing via sign epistasis to potentially explain the "lower-than-expected" mutant fitness. After the second editing, beneficial secondary mutations would be fewer because the mutant genes would be only one mutational step away from the wild-type. Random mutations can occur in the experiment such as the growth on the SC-URA plates, counter-selection on 5-FOA plates, growth in the YPD liquid medium. The deletions of some genes may increase the mutation rate, but none of our 21 genes are among those previously shown to increase the mutation rate upon deletion (Huang et al., 2003). The average fitness cost of random mutations is likely to be smaller than that of the mutations we created in the 21 genes because the average fitness cost of 21 gene deletions is 0.36, while the corresponding value of the 4450 (nonessential) gene

deletions surveyed (Qian et al., 2012) is only 0.03. Consistent with this prediction, we found the secondary mutations in the *BFR1*, *EOS1*, and *IES6* knockout strains did not have fitness cost (**Fig. 3-3**).

After characterizing the secondary mutations in the 21 mutant gene pools, we found that the number of secondary mutations vary in different pools. For the genes with non-negligible secondary mutations, we can build mutants and the wild-type control simultaneously and replicate this process multiple time to create biologically independent libraries for fitness quantification. In this way, 1) mutant strains and the wild-type control may all carry the same secondary mutations occurred in the first round of genome editing so the effect of secondary mutations is cancelled out in one biological replicate and 2) a genotype can be linked with different secondary mutations in different biological replicates, but if we take an average of mutant fitness from multiple biological replicates, the effect of secondary mutations will be cancelled out. This experiment is currently ongoing.

The conclusions of Chapter 2 are valid for the genes with no or negligible secondary mutations. Recent years have seen an increasing number of reports of fitness effects of synonymous mutations/polymorphisms from both case studies (Agashe et al., 2013; Frumkin et al., 2018; Kristofich et al., 2018; Lebeuf-Taylor et al., 2019; Walsh et al., 2020) and systematic analyses (Lind et al., 2010; Sane et al., 2022; Sharon et al., 2018; She & Jarosz, 2018). These findings, along with ours (Shen et al., 2022), suggest that many synonymous mutations are strongly non-neutral.

3.5 Materials and Methods

3.5.1 Library construction and genome sequencing

For each strain, genomic DNA was extracted from around 10^7 yeast cells using a MasterPure Yeast DNA Purification Kit (Lucigen; MPY80200). Sequencing libraries were constructed using Nextera DNA Flex Library Prep (Illumina; 20018705). Samples were sequenced using an Illumina HiSeq X with a paired-end 150 strategy.

3.5.2 Mutation identification

Sequencing reads were aligned to the *S. cerevisiae* reference genome (v.R64-2-1) using the Burrows-Wheeler Aligner with default parameters, and duplicated reads were removed using Picard tools (http://broadinstitute.github.io/picard/). SNVs were called using the Genome Analysis Toolkit (GATK) platform (McKenna et al., 2010). By comparing the genomes of knockout strains and the BY4742 progenitor strain, secondary mutations in the first round of editing were identified. By comparing the genomes of mutant strains with those of the respective knockout strains, secondary mutations in the second round of editing were identified.

3.5.3 Construction of *BFR1*, *EOS1*, *IES6* wildtype strains

We respectively amplified the wild-type *BFR1*, *EOS1*, *IES6* gene sequences from the genome of the haploid strain BY4742 by PCR and inserted them into the *ΔBFR11*, *ΔEOS1, ΔIES6* cells using CRISPR/Cas9. One colony was picked and the insertion was confirmed by Sanger sequencing. The cells were then counter-selected on 5-FOA plates to remove the CRISPR/Cas9 plasmid.

We measured the maximum growth rates of the previously used wild-type control strain and *BFR1*, *EOS1*, *IES6* wild-type control strains using Biotek Gen5^{TM} Microplate Reader. The

cells were first grown overnight. About 10,000 cells were added into 0.1 mL YPD in a well of a Costar™ 96-well plate and the culture was in continuous shaking at 30°C. Sixteen replicate growth curves were collected per strain, except that one replicate of *EOS1* was contaminated so was discarded. The maximum growth rate was calculated following a previous protocol (Warringer et al., 2003).

3.6 References

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol*, *30*(3), 549-560. https://doi.org/10.1093/molbev/mss273

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., & Kim, J. S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*, *24*(1), 132-141. https://doi.org/10.1101/gr.162339.113

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci U S A*, *115*(21), E4940-E4949. https://doi.org/10.1073/pnas.1719375115

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., . . . Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*, *31*(9), 827-832. https://doi.org/10.1038/nbt.2647

Huang, M. E., Rio, A. G., Nicolas, A., & Kolodner, R. D. (2003). A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proc Natl Acad Sci U S A*, *100*(20), 11529-11534. https://doi.org/10.1073/pnas.2035018100

Jakociunas, T., Bonde, I., Herrgard, M., Harrison, S. J., Kristensen, M., Pedersen, L. E., . . . Keasling, J. D. (2015). Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metab Eng*, *28*, 213-222. https://doi.org/10.1016/j.ymben.2015.01.008

Kristofich, J., Morgenthaler, A. B., Kinney, W. R., Ebmeier, C. C., Snyder, D. J., Old, W. M., . . . Copley, S. D. (2018). Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genet*, *14*(8), e1007615. https://doi.org/10.1371/journal.pgen.1007615

Kruglyak, L., Beyer, A., Bloom, J. S., Grossbach, J., Lieberman, T. D., Mancuso, C. P., . . . Kaplan, C. D. (2023). Insufficient evidence for non-neutrality of synonymous mutations. *Nature*, *616*(7957), E8-E9. https://doi.org/10.1038/s41586-023-05865-4

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife*, *8*. https://doi.org/10.7554/eLife.45952

Lind, P. A., Berg, O. G., & Andersson, D. I. (2010). Mutational robustness of ribosomal protein genes. *Science*, *330*(6005), 825-827. https://doi.org/330/6005/825 [pii]

10.1126/science.1194617

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Qian, W., Ma, D., Xiao, C., Wang, Z., & Zhang, J. (2012). The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell Rep*, *2*(5), 1399-1410. https://doi.org/10.1016/j.celrep.2012.09.017

Ryan, O. W., Skerker, J. M., Maurer, M. J., Li, X., Tsai, J. C., Poddar, S., . . . Cate, J. H. (2014). Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife*, *3*. https://doi.org/10.7554/eLife.03703

Sane, M., Diwan, G. D., Bhat, B. A., Wahl, L. M., & Agashe, D. (2022). Shifts in mutation spectra enhance access to beneficial mutations. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2020.09.05.284158

Sharon, E., Chen, S. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, *175*(2), 544-557 e516. https://doi.org/10.1016/j.cell.2018.08.057

She, R., & Jarosz, D. F. (2018). Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change. *Cell*, *172*(3), 478-490 e415. https://doi.org/10.1016/j.cell.2017.12.015

Shen, X., Song, S., Li, C., & Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, *606*(7915), 725-731. https://doi.org/10.1038/s41586-022-04823-w

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc Natl Acad Sci U S A*, *117*(7), 3528-3534. https://doi.org/10.1073/pnas.1907126117

Warringer, J., Ericson, E., Fernandez, L., Nerman, O., & Blomberg, A. (2003). High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci U S A*, *100*(26), 15724-15729. https://doi.org/10.1073/pnas.2435976100

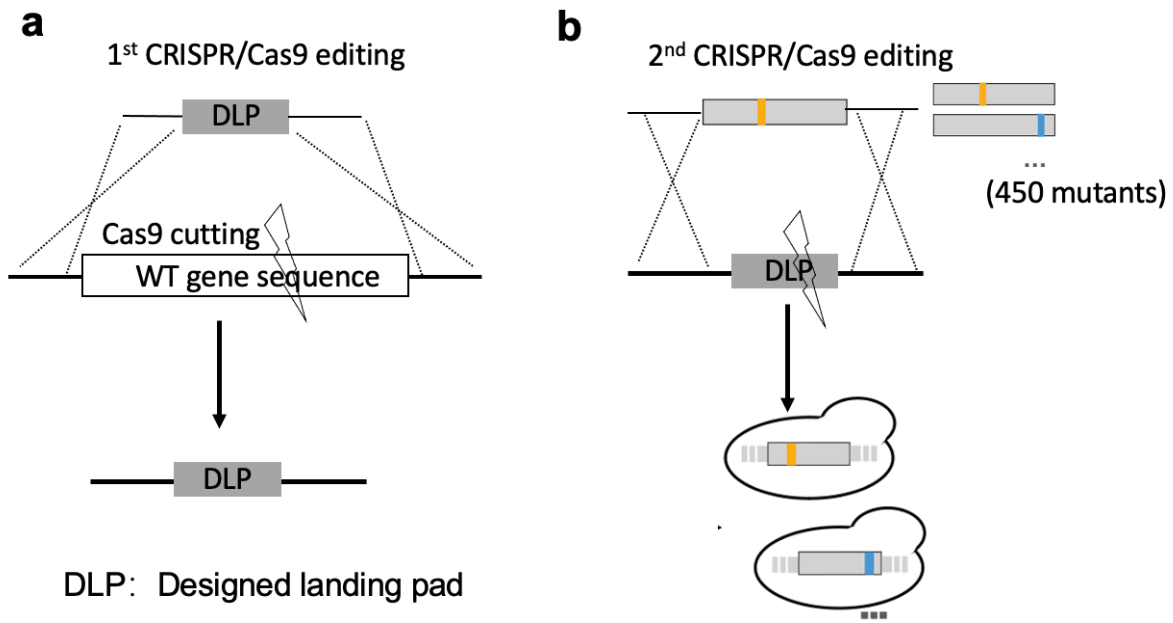

Figure 3-1. Two rounds of CRISPR/Cas9 editing. a, a segment of wild-type sequence was replaced by an artificially designed landing pad (DLP). **b**, the landing pad was replaced with variant sequences.

Sequence	PAM	Score	Gene	Cut Locus
GGTACCTTGGAAGGTCACAA	CGG	100.0	ASC1 (YMR116C)	chrXIII:-500642
GGTACTTTGGATGGTCACAC	CGG	0.5	TIF34 (YMR146C)	chrXIII:-558373
GGTACCATAAAAGGTCAAAA	TGG	0.1	PRP38 (YGR075C)	chrVII:-636664

Figure 3-2. An example of Benchling off-target site prediction (ASC1 deletion gRNA). The first row is the target site in *ASC1*. The second and third rows are two potential off-target sites. The red bases in the first column are the mismatched sequences. The Scores are calculated following a published method (Hsu et al., 2013)

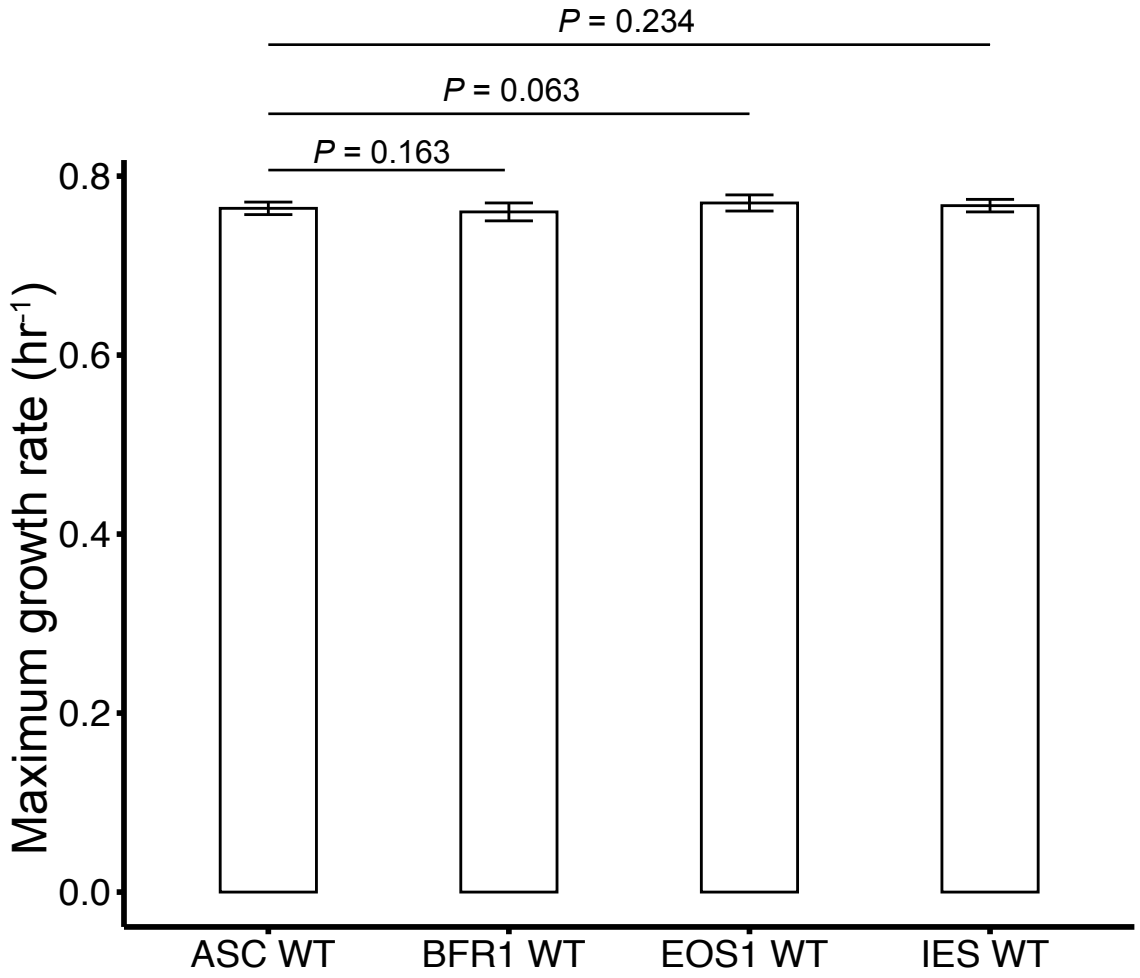

Figure 3-3. The maximum growth rates of four reconstructed wild-type strains. *ASC1 WT* was used as the wild-type control in *en masse* competitions with mutants. Error bar shows the standard error of the mean based on sixteen replicates, except for *EOS1* WT, which had 15 replicates, because one replicate was contaminated and discarded. *P*-values are from *t*-tests.

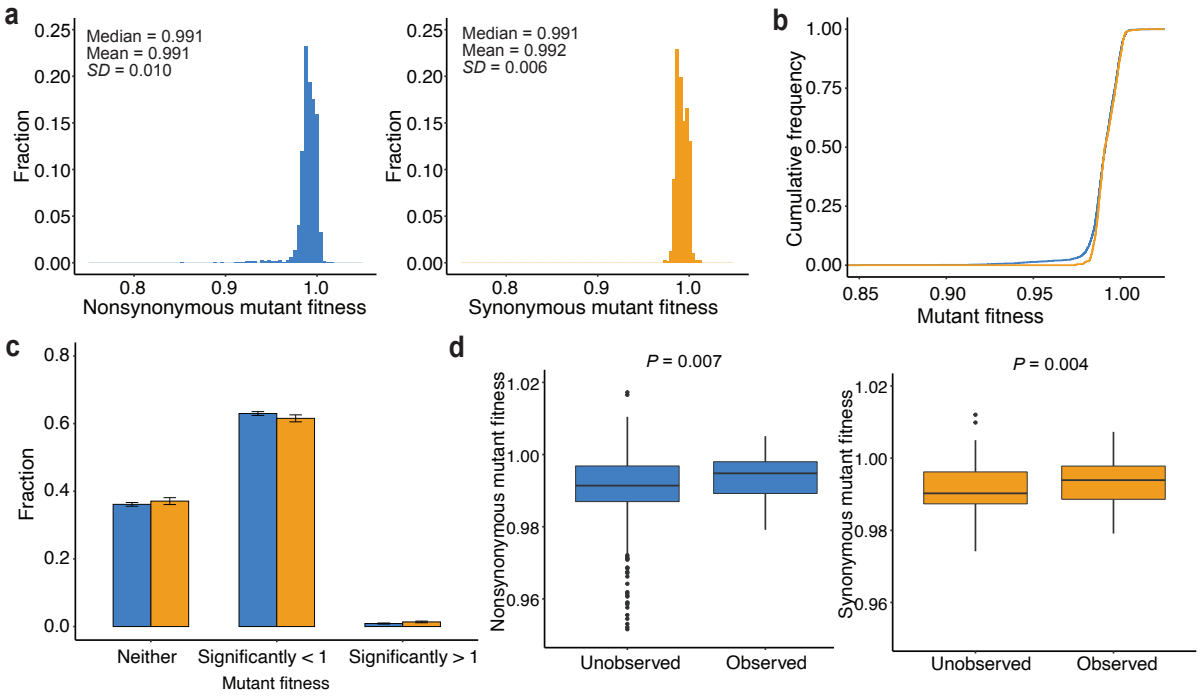

Figure 3-4. Mutant fitness in YPD of the 7 genes. a, Distributions of the fitness of 2,231 nonsynonymous (blue) and 658 synonymous (yellow) mutants. The two distributions are significantly different ($P = 0.049$, two-tailed Wilcoxon rank-sum test; $P = 0.007$, Kolmogorov–Smirnov test). b, Cumulative frequency distributions of fitness of nonsynonymous and synonymous mutants. c, Fractions of nonsynonymous and synonymous mutants with fitness significantly below 1 (nominal $P < 0.05$), significantly above 1, and neither, respectively. Error bars show one standard error. Nonsynonymous and synonymous mutants are not significantly differentially distributed among the three bins (two-tailed Fisher's exact test). d, Mutant fitness is lower when the mutation is not observed than when it is observed in the genomes of five related yeast species. *P*-values are from two-tailed Wilcoxon rank-sum test. Mutants with fitness <0.95 or >1.02 are not shown (see Fig. B-1 for the complete figure). In d, each data point is a mutant. The lower and upper edges of a box represent the first (qu_1) and third (qu_3) quartiles, respectively, the horizontal line inside the box indicates the median (*md*), the whiskers extend to the most extreme values inside inner fences, $md \pm 1.5(qu_3\text{-}qu_1)$, and the dots show outliers.

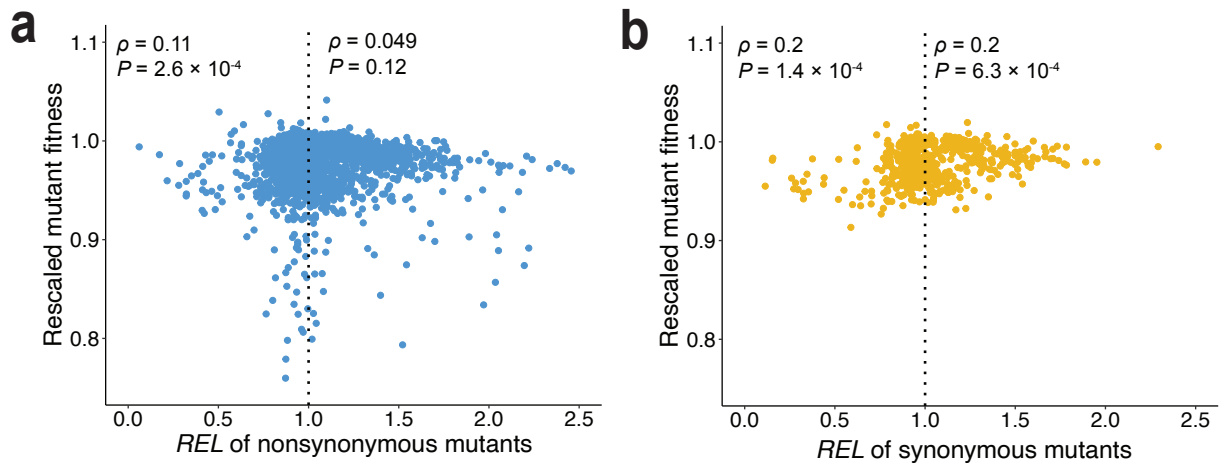

Figure 3-5. The correlations of *REL* and rescaled mutant fitness of the 7 genes. a, Correlation between *REL* and rescaled fitness among nonsynonymous mutants. **b,** Correlation between *REL* and rescaled fitness among synonymous mutants.

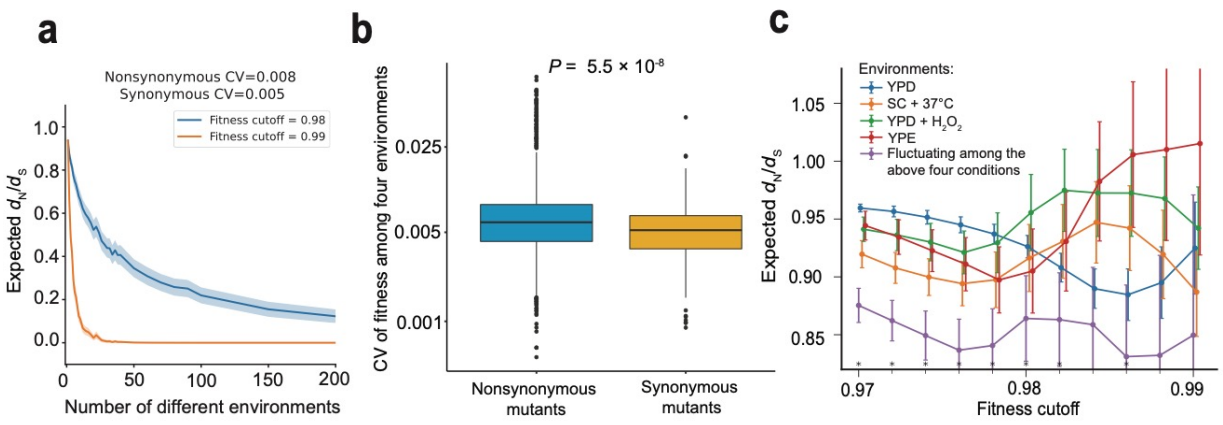

Figure 3-6. A higher fitness CV across environments for nonsynonymous than synonymous mutants in the 4 genes can create dN/dS <<1 despite similar DFEs of synonymous and nonsynonymous mutations in each environment. a, Expected d_N/d_S from 1000 simulations of a population that experiences multiple different environments. A mutant is purged if its fitness is below a preset cutoff such as 0.98 or 0.99 in any environment. Shaded areas represent 95% confidence intervals. **b**, Box plots (blue, nonsynonymous mutants; yellow, synonymous mutants) showing distributions of fitness *CV* across the four environments for 1,247 nonsynonymous and 367 synonymous mutants. The mean *CV* is 0.009 for nonsynonymous and 0.006 for synonymous mutants. The two distributions are significantly different (two-tailed Wilcoxon rank-sum test). **c**, Expected d_N/d_S when the population stays in a constant environment or a changing environment.

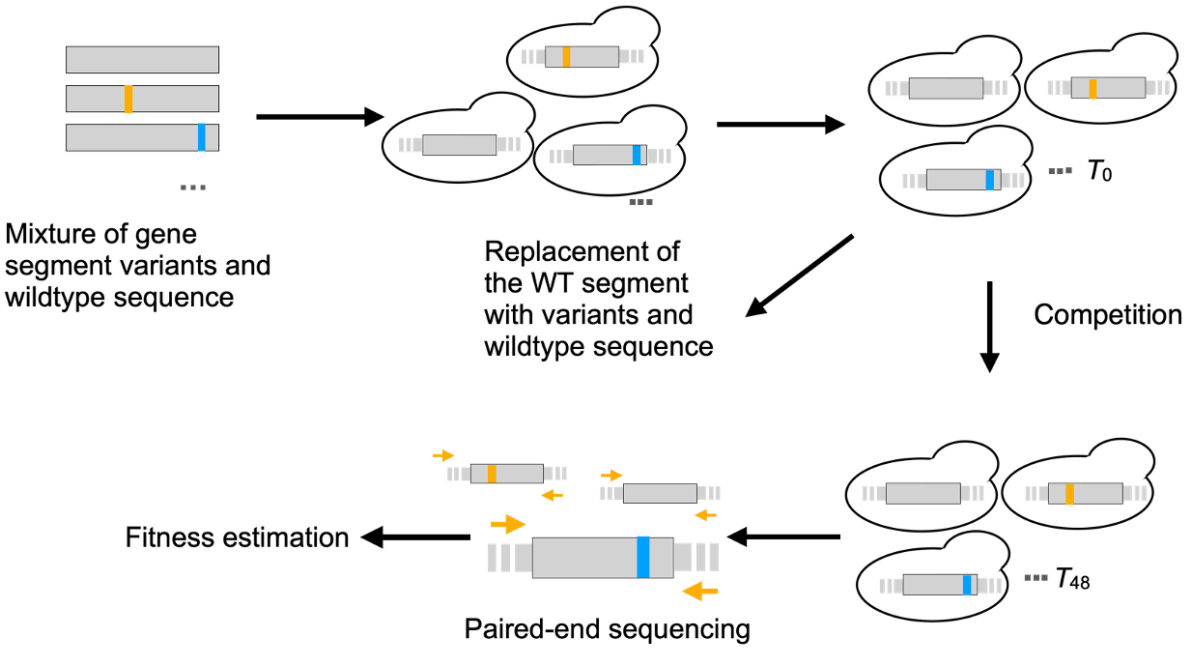

Figure 3-7. Procedure of mutant fitness measurement. A mixture of gene segment variants and wildtype sequence was inserted into the corresponding gene knockout strain simultaneously.

Table 3-1. The gRNA sequences and their off-target scores. Off-target scores range from 0 to 100, and the scores >= 50 are considered to be good guides (https://help.benchling.com/hc/en-us/articles/9684236710413)

gRNA targeted gene	gRNA sequence template	off-target score
ADA2	TGTTCCAAAAATTCATCTTG	50
ASC1	GGTACCTTGGAAGGTCACAA	50
BFR1	AAAGTCTGTTGATGCTGACA	50
BUD23	GGTCTTAGTAGAGAGCTGGA	50
CCW12	GATAACGTCATCGACGGTGA	50
EOS1	TGGCTCGTAAAGCGTTATGC	50
EST1	TGCTCAAACCATTGGAGTGT	50
GET1	TTTGTGAAGATGGGCCTGAA	50
GIM5	TCTGCATCATTATACATCCC	50
IES6	CAAGAAGACACAAATCAGCG	50
LSM 1	TACTGAGCTTACAATAGCAG	50
PAF1	ATATTCGGTGCGTCTCAAGA	50
PRS3	ATATGCAAGACAAGATAGAA	49
RAD6	TCCGTCTTCATATGGAGTAT	50
RPL29	CGCTCACAACCAAACCAGAA	50
RPL39	AAAACAGACCATTGCCACAA	50
RPS7A	TGAGACAAAATCTTGGCTTG	50
SNF6	AGTTCTTCAAAAGGTTGGAT	50
TSR2	ATCCCACTTATAAATAACCA	50
VMA21	GCTGTTTACTGCAGCGATGG	50
VMA7	AGAAGGTAAGACTACTAAGG	50
20-nt Cas9 target sequence inserted in the deletion strain	ATGCGGGTAGAAGATTACGG	100

Table 3-2. Summary of the mutations identified in the gene knockout strains.

Gene/strain name	# of mutations in the knockout strain	# of nonsynonymous mutations	# of synonymous mutations	# of noncoding mutations
ADA2	0	0	0	0
ASC1	0	0	0	0
BFR1	2	1	1	0
BUD23	0	0	0	0
CCW12	1	1	0	0
EOS1	1	1	0	0
EST1	1	1	0	0
GET1	1	0	1	0
GIM5	1	1	0	0
IES6	1	0	1	0
LSM1	0	0	0	0
PAF1	0	0	0	0
PRS3	2	1	1	0
RAD6	3	1	2	0
RPL29	0	0	0	0
RPL39	0	0	0	0
RPS7A	1	1	0	0
SNF6	2	0	1	1
TSR2	0	0	0	0
VMA21	0	0	0	0
VMA7	2	1	1	0
ASC1 wildtype control	0	0	0	0

Table 3-3. Summary of the mutations identified in the mutant and reconstructed wildtype strains.

Gene/strain name	# of mutant strains sequenced	# of mutant strains carrying secondary mutations	# of nonsynonymous mutations	# of synonymous mutations	# of noncoding mutations
ADA2	33	0	0	0	0
ASC1	30	1	0	1	0
BFR1	20	0	0	0	0
BUD23	30	12	5	1	6
CCW12	33	11	9	2	1
EOS1	24	2	1	0	1
EST1	24	11	7	0	4
GET1	24	12	9	0	3
GIM5	24	11	5	1	5
IES6	30	1	1	0	0
LSM1	30	16	12	3	2
PAF1	29	6	6	0	1
PRS3	24	8	3	4	1
RAD6	32	11	12	1	1
RPL29	30	13	9	3	3
RPL39	28	2	0	1	1
RPS7A	24	10	6	2	1
SNF6	24	8	4	0	4
TSR2	32	0	0	0	0
VMA21	30	7	3	1	3
VMA7	24	13	8	3	2
ASC1 wildtype control	1	0	0	0	0

Chapter 4 Epistatic Effects of Synonymous Mutations in Yeast Genes

4.1 Abstract

Interaction between mutations, or epistasis, can be important to the fitness and evolution of organisms. For example, a mutation may have opposite effects in two genetic backgrounds and this difference can lead to different evolutionary trajectories. Intragenic epistasis involving synonymous mutations have not been systematically studied. I recently reported the distribution of fitness effects of more than 8000 single mutations in yeast and found that about three-quarters of synonymous mutation are nonneutral. In that study, I intended to chemically synthesize pools of oligos each carrying one mutation, but unavoidable errors in the synthesis also created oligos carrying multiple mutations. I took advantage of these oligos, the corresponding mutants constructed, and the fitness measurements obtained to estimate intragenic epistasis in four environments. Notably, I found that synonymous mutations genetically interact with synonymous or nonsynonymous mutations, showing 6.4% significantly positive interactions and 5.9% significantly negative interactions. My findings further revealed that 8.5% to 26.1% of epistatic interactions are significantly different between any two of four environments studied. Interestingly, epistasis between nonsynonymous mutations is more variable than that between synonymous mutations across the four environments examined.

4.2 Introduction

To understand the fitness effects of mutations comprehensively, we need to take epistasis into consideration because mutations can have different effects depending on the genetic background (Phillips, 2008). Epistasis can be classified into two categories: positive epistasis, where the outcome is better than expected, and negative epistasis, where the outcome is worse than expected. These epistatic interactions play a crucial role in shaping evolutionary trajectories and make evolution more contingent and less predictable (Domingo et al., 2019; Park et al., 2022; Starr et al., 2018). Epistasis also facilitates speciation if two isolated populations of a species acquire different mutations. Ultimately, the mutations fixed in one population are deleterious in another population (Presgraves, 2007). Notably, epistasis is not a fixed characteristic; rather, it exhibits variation among different environments (G×G×E interaction). This variability in epistasis has been observed across various organisms, including viruses (Lalic & Elena, 2013), bacteria (Remold & Lenski, 2004), yeast (Gerke et al., 2010; Li & Zhang, 2018) and flies (Zhu et al., 2014).

Because of the huge genotype space (4^n genotypes for a gene with n nucleotides), it is difficult to study all possible mutants of a genes. Most systematic studies of genotype-fitness mapping investigated the fitness effects of single mutations (Bank et al., 2015; Flynn et al., 2020; Hietpas et al., 2013; Hietpas et al., 2011; Mavor et al., 2016; Roscoe et al., 2013; Shen et al., 2022). The epistasis between mutations of one gene (Li et al., 2016; Li & Zhang, 2018; Puchta et al., 2016; Sarkisyan et al., 2016) was less studied, not to mention the study of epistasis differences across multiple environments. One genotype-fitness mapping study measured the epistasis in multiple environments but they did not study protein-coding genes (Li & Zhang,

2018). When we study the intragenic epistasis between two mutations, in addition to studying epistasis between nonsynonymous mutations, it is also valuable to study the epistatic interactions between two synonymous mutations and epistasis between a synonymous mutation and a nonsynonymous mutation, because many synonymous mutations are non-neutral (Shen et al., 2022).

Here, we made use of the fitness data of double mutants produced by oligo synthesis errors (Shen et al., 2022) and estimated intragenic epistasis of mutations in multiple genes in each of four environments.

4.3 Results

4.3.1 Characterizing intragenic epistasis in the four environments

The double mutants were produced through oligo synthesis errors (see Methods) so their frequencies are generally lower than the single mutants. To detect more double mutants and estimate their fitness more accurately, we sequenced the competition populations of 13 genes from the YPD environment using NovaSeq so the sequencing depth was about 6.5 times the original sequencing (see Methods). The template switching events in the PCR steps could generate double mutants, but they were very rare (see Methods). Of these 13 genes, we observed that only up to two mutant strains carried secondary mutations in six genes (*ADA2*, *BFR1*, *EOS1*, *IES6*, *RPL39*, *TSR2*) when compared to their respective knockout strains (see Chapter 3). We identified a total of 8,263 double mutant variants with read counts ≥ 50 at T_0, including 4,797 double nonsynonymous mutants (NN mutants), 3,058 nonsynonymous-synonymous double mutants (NS mutants), and 408 double synonymous mutants (SS mutants) in these 6 genes (**Fig. 4-1**). The fitness estimates were moderately correlated between replicates, with a mean Pearson's *r* of 0.43 (**Fig. C-1**). The lower fitness correlations observed in double mutants (compared with single mutants, **Fig. A-5**) can be attributed to their production through synthesis errors, leading to lower genotype frequencies. The mean fitness of the 8,263 double mutants is 0.978 (**Fig. 4-2a**). The corresponding values are 0.977, 0.979 and 0.982 for the NN mutants, NS mutants, and SS mutants (**Fig. 4-2b-d**). These values are lower than both the mean fitness of nonsynonymous mutants and that of synonymous mutans (**Fig. 1-2a**). In addition, the fitness of NN mutants is significantly lower than that of NS mutants ($P = 0.005$, two-tailed Wilcoxon rank-sum test) and that of SS mutants ($P = 0.02$, two-tailed Wilcoxon rank-sum test). We estimated epistasis between mutations using the fitness estimates of single mutants and double mutants (**Fig. 4-2e-**

h). In YPD, we found substantial epistatic interactions in SS mutants (**Fig. 4-2h**). The mean epistasis is -0.009 for SS mutants, with 5.9% of epistatic interactions being significantly negative and 6.4% being significantly positive (**Fig. 4-2h**). We also found prevalent significant epistatic interactions in NN mutants and NS mutants (**Fig. 4-2h**) and the epistasis distributions of NN, NS, SS mutants (**Fig. 4-2f-h**) are not significantly different between any two groups ($P > 0.05$, two-tailed Wilcoxon rank-sum test). The mean epistasis is negative for NN, NS, and SS mutants, but the distributions of epistasis were not strongly negatively biased as previously observed in a tRNA gene (Li & Zhang, 2018).

4.3.2 G×G×E interactions are prevalent

We estimated intragenic epistasis for 4 genes (*ADA2*, *BFR1*, *RPL39*, *TSR2*) in SC + 37°C, YPD + 0.375 mM H_2O_2, and YPE. In addition to being minimally affected by the secondary mutations in the mutants, these four genes do not have secondary mutations in their knockout strains (see Chapter 3). In SC + 37°C and YPD + 0.375 mM H_2O_2, the fractions of negative epistasis and positive epistasis are similar (**Fig. C-2ab**), while in YPE (**Fig. C-2c**), epistasis is positively biased.

We compared the epistasis in any two of the four environments and found that a substantial fraction of epistasis varied between environments (**Fig. 4-3**), revealing prevalent G×G×E. The extent of these variations varies across the pairs of environments, with the most significant fraction observed between YPD and YPE, reaching an impressive 26.1%. Of particular interest is the change of the sign of epistasis between environments. These fractions range from 3.5% to 10.1% among the six environment pairs. On average, 18.1% of NN epistasis varied between two of the four environments surveyed (**Fig. 4-4**). The corresponding values for

the NS mutants and SS mutants are 14.7% and 14.1%, respectively (**Fig. 4-4**). NN mutants showed significantly more prevalent G×G×E interaction than NS and SS mutants (**Fig. 4-4**). In Chapter 2, I discovered that nonsynonymous mutations exhibited greater G×E interaction across environments compared to synonymous mutants. This is likely due to the fact that nonsynonymous mutations not only impact mRNA level and mRNA folding but also lead to changes in protein functions which are environment-dependent, so the nonsynonymous mutations have fitness effects that are more variable across environments. I now demonstrated that epistasis between nonsynonymous mutations also showed more pronounced variability across environments compared to that of synonymous mutations. The epistasis between nonsynonymous mutations involves the interaction between protein sequence changes which synonymous mutations do not cause, so this adds to the complexity and variability of epistasis between nonsynonymous mutations.

4.4 Discussion

In summary, by analyzing thousands of double mutants, I found that intragenic epistasis between single mutations is prevalent. Besides, synonymous mutations can genetically interact with other synonymous mutations or nonsynonymous mutations. In Chapter 2, I found that mutant fitness and the mRNA expression level of the gene with the mutation are positively correlated. One might expect that fitness epistasis and expression epistasis (see Methods) are positively correlated. I tested this hypothesis using the fitness and Relative Expression Level (*REL*) data in YPD but did not find a significant correlation (**Fig. C-3**), probably due to the limited sample size of the 1,070 double mutants of which both the fitness and *REL* were estimated. Future studies should investigate the mechanisms of epistatic effects of synonymous mutations.

I also found widespread G×G×E interaction. Between any two of the four environments examined, 8.5% to 26.5% epistatic interactions are significantly different (**Fig. 3**). Of particular interest is that I found that nonsynonymous mutations have more prevalent G×G×E interaction than synonymous mutations (**Fig. 4**). Nonsynonymous mutations change protein sequence in addition to altering mRNA level and mRNA folding strength. This not only adds to the prevalence of G×E interaction of nonsynonymous mutations but also leads to more pronounced G×G×E interaction within nonsynonymous mutations themselves. Because epistasis is crucial in evolution, its variation across environments could be especially important for natural populations, as most of them experience fluctuating and dynamic environmental conditions.

In this study, I probed epistasis of intragenic mutations in multiple genes across environments. Epistasis of intergenic mutations and its environmental dependence would be an

interesting subject for future endeavors. Intergenic epistasis is thought to play important roles in the evolution of genetic systems, but our systematic knowledge about intergenic epistasis is largely from the data of double gene deletions. Although double gene deletions can provide information on gene-gene interactions, they likely provide a biased view, or at most an incomplete view, because most mutations are not null mutations. Future endeavors towards studying epistasis of intergenic mutations can be highly rewarding.

4.5 Materials and Methods

4.5.1 Double mutant identification and epistasis estimation

The details of variant construction, competitions, fitness calculations can be found in Chapter 2. In the chemical synthesis of oligo pools, the error rate per base is about 0.43% to 0.73% (https://www.genscript.com/gsfiles/techfiles/Oligo-Pools-design-synthesis-and-research-applications-slides.pdf). The mutant region of each gene is 150 nucleotides, so there is about 1 expected error per oligo. This error and the designed mutation form a double mutant genotype.

To calculate the epistasis between two mutations in each environment, we utilized the single mutant fitness and double mutant fitness calculated following the way described in Chapter 2. The epistasis is defined as $\varepsilon = f_{AB} - f_A f_B$, where f_{AB} is the fitness of the double mutant and f_A and f_B are the fitness of the corresponding single mutants.

4.5.2 Test of template switching

Template switching is a process by which two PCR templates combine to form a chimeric product. If two single-mutant templates switched, double mutants could be produced. To quantify the template switching events, I extracted the genomes of three *GET1* single mutants (The mutations were at the 46th, 139th, 143rd base in the 150-bp mutant region). I mixed equal quantities of three genomes. The mixture was used as the template DNA for 25 cycles of PCR (same as the number of PCR cycles used in the library preparations after competitions) to amplify the mutant region. The PCR product was sequenced with 250-nucleotide paired-end sequencing. Only 0.27% of sequences identified were caused by template switching and no template switching between the two mutations at the 139th base and 143rd base was observed.

This is expected, given that the probability of observing a template switch involving two mutations should increase with the distance between the two mutations. Let us assume that the rate of template switch between the 46th site and the 139th (or 143rd) site be a. Hence, ($2a/3 + a/3 + a/3)/3 = 4a/9 = 0.27\%$, where 2/3 indicates that 2/3 of the template switches involving the first mutant (harboring the mutation at the 46th site) is observable, because switches between the molecules of the same genotype are not observed. We estimated that $a = 0.6075\%$. Because the average distance between two sites in our mutant pool is 75 bases, shorter than the distances between the two mutations tested here (93 or 97 bases), a is likely an overestimate of the template switch rate in the actual mutant pool. By contrast, 13% of sequencing reads of *GET1* T_0 sample encode double mutants. So, double mutants caused by template switching is < 0.6075% / 13% = 4.67% of total double mutants.

4.5.3 Deep sequencing of the competition populations from 13 genes in the YPD environment

We resequenced the competition populations of 13 genes which have the most double mutants identified. In Chapter 2, we sequenced the competition populations of 21 genes on HiSeq2500. Resequencing was performed on NovaSeq. NovaSeq can generate 400 million read pairs in one lane while HiSeq2500 can produce 100 million read pairs in one lane. Consequently, the number of sequencing reads per genotype from NovaSeq was about 6.5 times that from HiSeq2500. In Results, we used the data from 6 of these 13 genes.

4.5.4 Expression epistasis estimation

The Relative Expression Level (*REL*) of the double mutants were estimated following the way described in Chapter 2. The expression epistasis is defined by $\lambda = REL_{AB} - REL_A REL_B$,

where REL_{AB} is the relative expression of the double mutant, and REL_{A} and REL_{B} are the relative expression of the corresponding single mutants.

4.6 References

Bank, C., Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2015). A systematic survey of an intragenic epistatic landscape. *Mol Biol Evol*, *32*(1), 229-238. https://doi.org/10.1093/molbev/msu301

Domingo, J., Baeza-Centurion, P., & Lehner, B. (2019). The Causes and Consequences of Genetic Interactions (Epistasis). *Annu Rev Genomics Hum Genet*, *20*, 433-460. https://doi.org/10.1146/annurev-genom-083118-014857

Flynn, J. M., Rossouw, A., Cote-Hammarlof, P., Fragata, I., Mavor, D., Hollins, C., 3rd, . . . Bolon, D. N. (2020). Comprehensive fitness maps of Hsp90 show widespread environmental dependence. *Elife*, *9*. https://doi.org/10.7554/eLife.53810

Gerke, J., Lorenz, K., Ramnarine, S., & Cohen, B. (2010). Gene-environment interactions at nucleotide resolution. *PLoS Genet*, *6*(9), e1001144. https://doi.org/10.1371/journal.pgen.1001144

Hietpas, R. T., Bank, C., Jensen, J. D., & Bolon, D. N. A. (2013). Shifting fitness landscapes in response to altered environments. *Evolution*, *67*(12), 3512-3522. https://doi.org/10.1111/evo.12207

Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2011). Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A*, *108*(19), 7896-7901. https://doi.org/10.1073/pnas.1016024108

Lalic, J., & Elena, S. F. (2013). Epistasis between mutations is host-dependent for an RNA virus. *Biol Lett*, *9*(1), 20120396. https://doi.org/10.1098/rsbl.2012.0396

Li, C., Qian, W., Maclean, C. J., & Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, *352*(6287), 837-840. https://doi.org/10.1126/science.aae0568

Li, C., & Zhang, J. (2018). Multi-environment fitness landscapes of a tRNA gene. *Nat Ecol Evol*, *2*(6), 1025-1032. https://doi.org/10.1038/s41559-018-0549-8

Mavor, D., Barlow, K., Thompson, S., Barad, B. A., Bonny, A. R., Cario, C. L., . . . Fraser, J. S. (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife*, *5*. https://doi.org/10.7554/eLife.15802

Park, Y., Metzger, B. P. H., & Thornton, J. W. (2022). Epistatic drift causes gradual decay of predictability in protein evolution. *Science*, *376*(6595), 823-830. https://doi.org/10.1126/science.abn6895

Phillips, P. C. (2008). Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, *9*(11), 855-867. https://doi.org/10.1038/nrg2452

Presgraves, D. C. (2007). Speciation genetics: epistasis, conflict and the origin of species. *Curr Biol*, *17*(4), R125-127. https://doi.org/10.1016/j.cub.2006.12.030

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., & Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. *Science*, *352*(6287), 840-844. https://doi.org/10.1126/science.aaf0965

Remold, S. K., & Lenski, R. E. (2004). Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nat Genet*, *36*(4), 423-426. https://doi.org/10.1038/ng1324

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., & Bolon, D. N. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol*, *425*(8), 1363-1377. https://doi.org/10.1016/j.jmb.2013.01.032

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., . . . Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, *533*(7603), 397-401. https://doi.org/10.1038/nature17995

Shen, X., Song, S., Li, C., & Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, *606*(7915), 725-731. https://doi.org/10.1038/s41586-022-04823-w

Starr, T. N., Flynn, J. M., Mishra, P., Bolon, D. N. A., & Thornton, J. W. (2018). Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc Natl Acad Sci U S A*, *115*(17), 4453-4458. https://doi.org/10.1073/pnas.1718133115

Zhu, C. T., Ingelmo, P., & Rand, D. M. (2014). GxGxE for lifespan in Drosophila: mitochondrial, nuclear, and dietary interactions that modify longevity. *PLoS Genet*, *10*(5), e1004354. https://doi.org/10.1371/journal.pgen.1004354

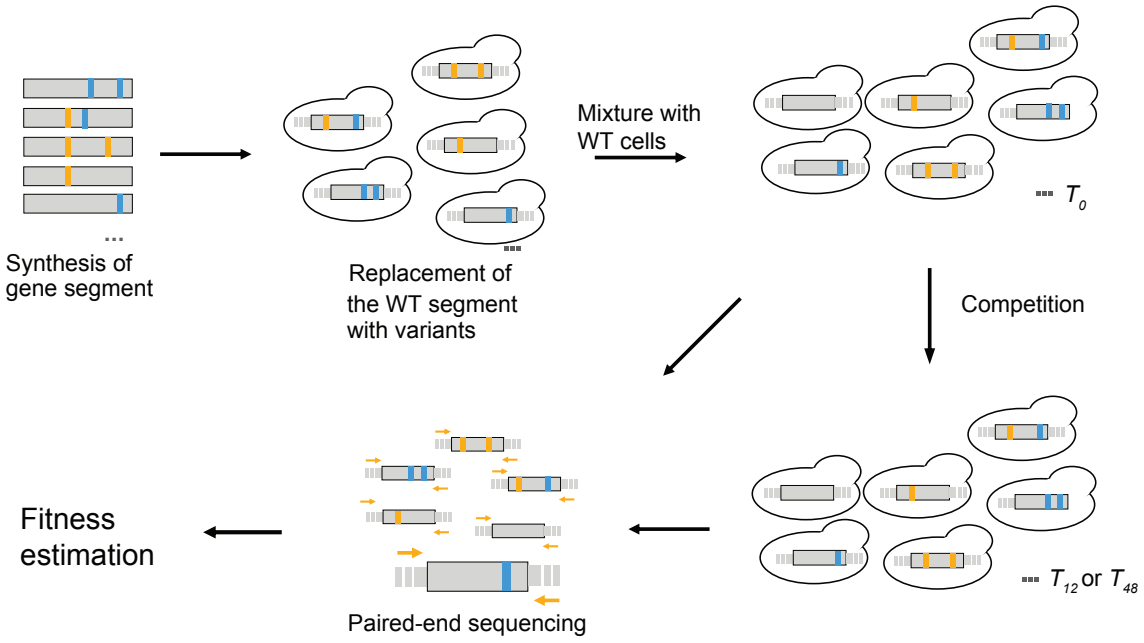

Figure 4-1. Experimental procedure of measuring the fitness of single mutants and double mutants.

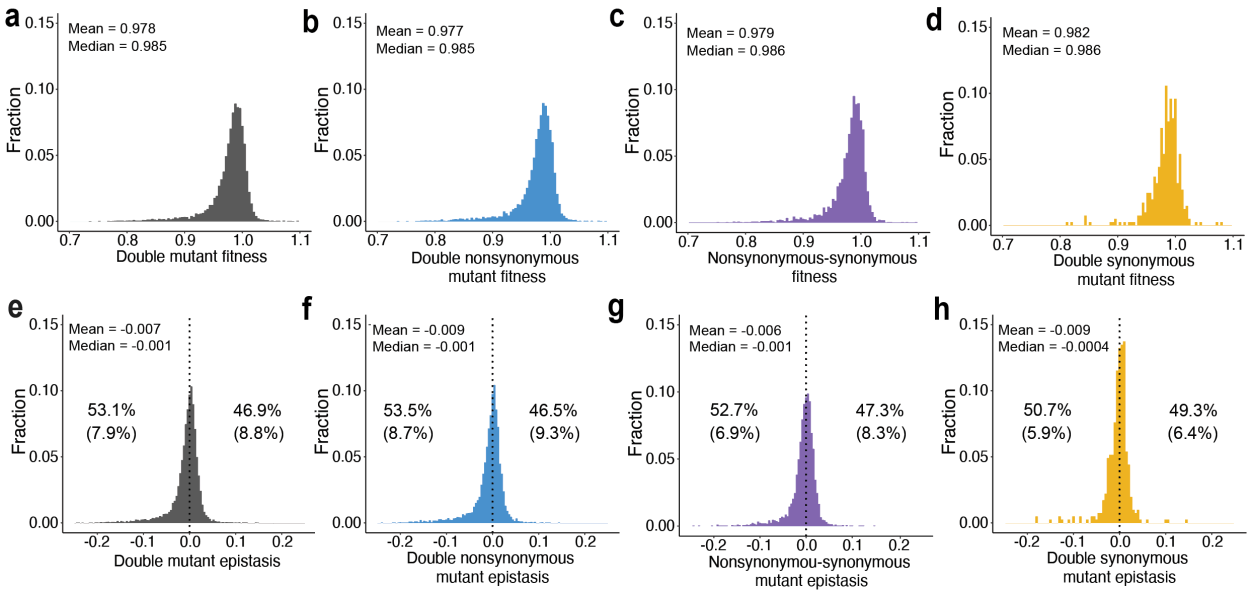

Figure 4-2. Distributions of double mutant fitness and epistasis in YPD for the 6 genes. a-d, Distributions of fitness in YPD. **e-h**, Distributions of epistasis. The percentages not in the parentheses are the fractions of negative or positive epistasis based on the face values. The percentages in the parentheses are the fractions of significant negative or significant positive epistasis (P <0.05, t-test)

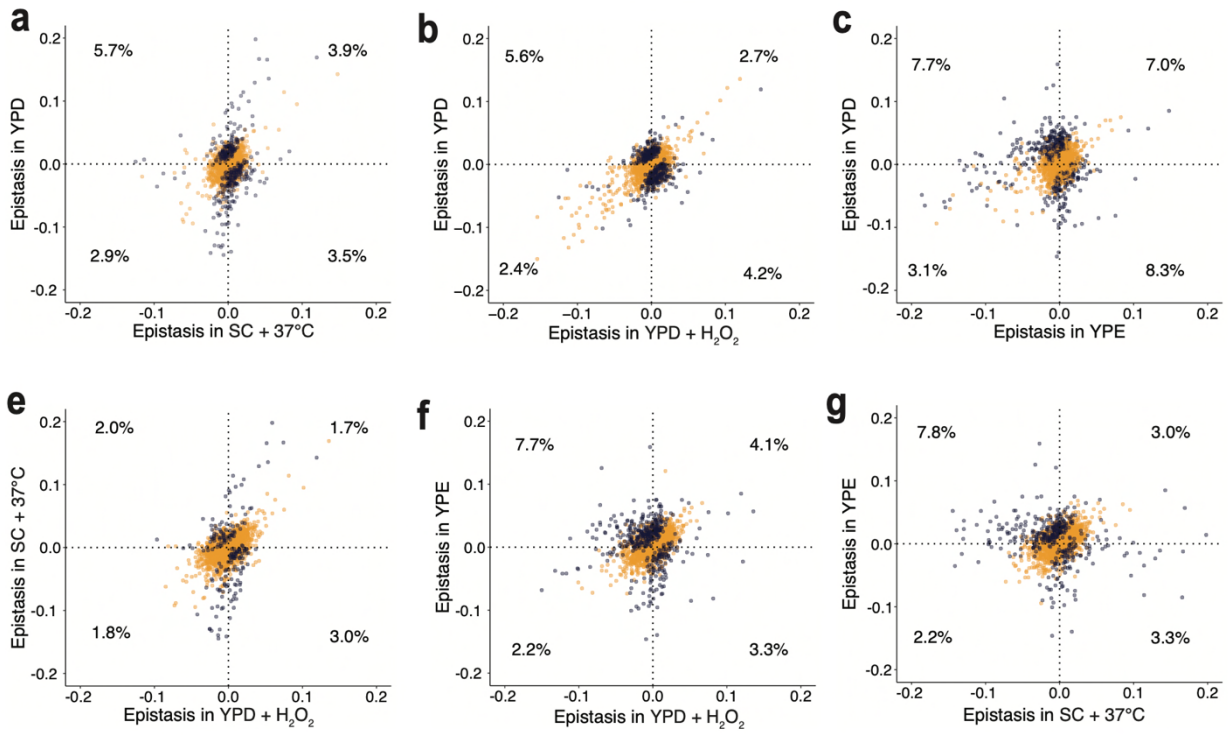

Figure 4-3. Comparison of epistasis between environments. Each dot is a double mutant. Purple dots are the mutants which have significantly different epistasis in the two environments compared (nominal $P < 0.05$, *t*-test), whereas yellow dots are the rest of the double mutants. The percentage in each quadrant shows the number of purple dots in the quadrant divided by the total number of dots in the panel.

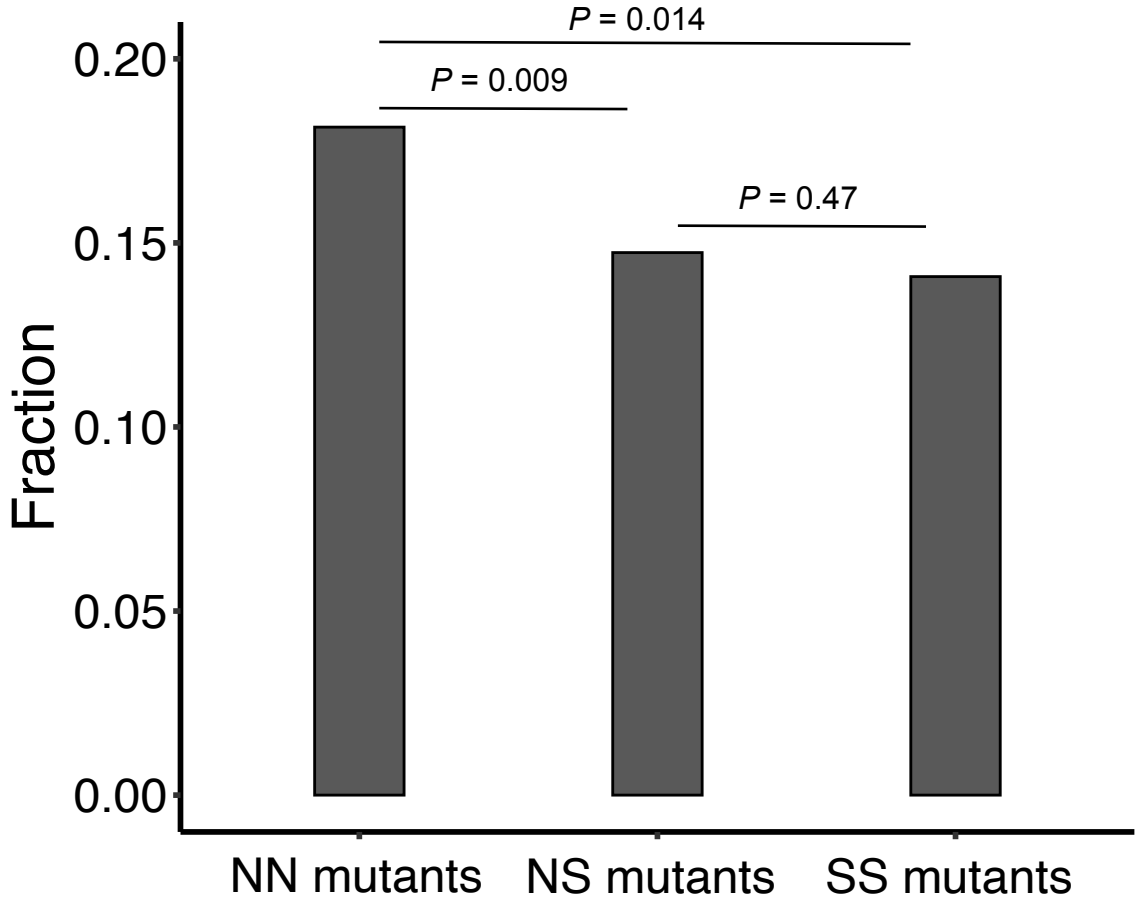

Figure 4-4. Average fraction of mutation pairs that exhibited significant G×G×E for each environment pair of NN, NS and SS mutants (nomial $P < 0.05$, *t*-test).NN mutants have significantly more prevalent G×G×E interactions than NS and SS mutants. *P*-values above the bars are from paired *t*-tests.

Chapter 5 Conclusions

In this concluding chapter, I will summarize my main findings of each chapter and discuss future directions.

In Chapter 1, I discussed the discovery of the non-neutrality of synonymous mutations, biological processes that can be influenced by synonymous mutations, evidence supporting the non-neutrality of synonymous mutations from systematic fitness measurements, impacts of the non-neutrality of synonymous mutations on evolutionary studies, and epistasis. I also introduced the content of Chapters 2, 3, and 4.

In Chapter 2, I created mutant libraries for 21 genes in yeast. In each gene, I randomly chose a 150-nt sequence and chemically synthesized 450 one-mutation-away variants. I replaced the wild-type 150-nt sequence with the variant sequences using two rounds of CRISPR/Cas9 editing and measured the fitness effects of mutations by a sequencing-based method. I estimated the fitness of more than 8000 single mutants, including 1866 synonymous mutants. I found that the distribution of the fitness effects of synonymous mutations is similar to that of nonsynonymous mutations and that most synonymous mutations are nonneutral. I tested whether the fitness effects of synonymous mutations are due to their impacts on translation but did not find a significant signal. I proposed that synonymous mutations change mRNA level and thereby are nonneutral. I assessed this hypothesis by measuring the mRNA level of each mutant gene relative to that in the wild type. I found that mutations (either synonymous or nonsynonymous) decreasing the expression are likely to be deleterious and there is a significant correlation between fitness and expression level. This correlation was not seen in mutations with elevated

expression levels. Further, I found the expression level to be positively correlated with the codon adaptation index (Sharp & Li, 1987), which can be explained by the mutational effect on transcription (Zhou et al., 2016) or mRNA stability (Presnyak et al., 2015). To explain why d_N/d_S <<1 for most genes (Graur et al., 2016; Li, 1997; Nei & Kumar, 2000), I proposed a hypothesis that nonsynonymous mutations have larger across-environment fitness variation than synonymous mutations, making nonsynonymous mutations more likely than synonymous mutations to be purged in fluctuating environments. This hypothesis was supported by a simulation. I then measured the fitness effects of mutations in three additional environments. Indeed, I found that nonsynonymous mutations have higher fitness variations among the four environments than synonymous mutations. A simulation was then conducted to show that, d_N/d_S is lower for a population rotating among the four environments than for a population staying in any one of the four environments.

There is a hypothesis that the fitness effects of synonymous mutations observed in Chapter 2 were due to CRISRP/Cas9 off-target editing or/and secondary mutations. I tested this hypothesis by sequencing the relevant genomes. In Chapter 3, I performed whole-genome sequencing of the BY4742 progenitor strain, wild-type control strain used in the competition, 21 gene knockout strains, and ~28 mutant strains per gene. I did not find any CRISPR/Cas9 off-target edits but found some secondary mutations. I identified seven genes with negligible effects of secondary mutations and found that all results in Chapter 2 hold for this set of seven genes. Future studies on the fitness effects of mutations should be more cautious of potential secondary mutations.

In Chapter 4, I took advantage of the double mutants produced as a result of oligo synthesis errors and estimated intragenic epistasis between mutations in the four environments.

Apart from the epistasis between nonsynonymous mutations, I found substantial epistatic interaction between synonymous mutations. I found that 8.5% to 26.1% of epistatic interactions vary significantly between two environments. Nonsynonymous mutations displayed a higher prevalence of G×G×E interaction compared to synonymous mutations.

Below, I discuss the implications and future directions derived from the results in Chapters 2 to 4.

Synonymous mutations have long been assumed to be neutral markers in population genetic estimation and conservation. However, I found that about three-quarters of synonymous mutations are non-neutral in 21 representative genes and synonymous mutations genetically interact with synonymous or nonsynonymous mutations. The invalidation of the assumption that synonymous mutations are neutral has broad impact on evolution. Effective population size (N_e) is usually calculated by dividing the genetic diversity (π) by 2μ (haploids) or 4μ (diploids), where μ is the neutral mutation rate. Genetic diversity is usually estimated based on the synonymous variants with the presumption that they are neutral in evolution. If many synonymous mutations are non-neutral, effective population size is underestimated. In conservation biology, effective population size is a critical parameter because it provides valuable insights into the genetic health of populations, which in turn affects their ability to adapt and survive in changing environments. If N_e is underestimated, it could necessitate a reevaluation of various conservation decisions. One potential solution to estimating genetic diversity is by considering noncoding variants instead of synonymous ones. However, a systemic yeast study revealed that variants affecting fitness are enriched in the noncoding regulatory regions (Sharon et al., 2018), complicating the use of these regions as neutral markers. Synonymous variants are regarded as neutral polymorphisms or substitutions in many selection tests, such as d_N/d_S test. The presence of non-neutral synonymous

mutations leads to the underestimation of the synonymous mutation rate. However, the d_N/d_S test can still be used to detect selection acting on protein sequences because, in addition to altering the protein sequence, nonsynonymous mutations affect fitness in the same ways as synonymous mutations do.

In Chapter 2, I measured the fitness effects of single mutations in haploid yeast cells. Future studies can assess the mutational effects in heterozygous diploid cells and estimate the dominance of synonymous mutations Dominance is one of the most important genetic phenomena, discovered by Gregor Mendel in his classic garden pea experiments. Several models (Fisher, 1928; Kacser & Burns, 1981; Wright, 1934) have been proposed to explain the cause of dominance. Fisher suggested that dominance arises from natural selection for fitter heterozygotes (Fisher, 1928). Wright, on the other hand, proposed a physiological theory of dominance (Wright, 1934). In support of Wright's theory, Kacser and Burns' metabolic control theory (Kacser & Burns, 1981) predicts a negative correlation between selection coefficient (s) and dominance (h) among mutations of the same enzyme gene. Dominance measured in null mutants (Phadnis & Fry, 2005) of different genes is consistent with this prediction, but the prediction is actually about different mutations of the same (enzyme) gene. Hence, it is unclear whether the prediction is empirically supported. Furthermore, measuring the dominance of coding mutations helps understand the mechanisms of their fitness effects, because deleterious coding mutations are more likely to be recessive if they lower fitness mainly by reducing their physiological functions, but are more likely to be dominant if they reduce fitness mainly by creating cytotoxicity. In Chapter 2, I found that synonymous mutations could reduce the mRNA level and fitness. Because most genes are haplosufficient (Deutschbauer et al., 2005), such mutations are likely to be recessive. By contrast, nonsynonymous mutations can change the protein sequence

and possibly create cytotoxicity. Thus, it is interesting to study whether synonymous mutations are more likely to be recessive than nonsynonymous mutations in heterozygotes.

Our finding that many synonymous mutations are nonneutral in yeast suggests the possibility that synonymous mutations can be disease-causing in humans. A review paper (Sauna & Kimchi-Sarfaty, 2011) published in 2011 listed more than 30 diseases caused by synonymous mutations, and they were found by accident . In a survey (Chen et al., 2010) including the results of 2,113 GWAS studies and in a recent exome-sequencing GWAS study (Karczewski et al., 2022) of nearly 400,000 humans in the UK Biobank, the contribution of synonymous mutations to disease is only slightly smaller than nonsynonymous mutations. I want to finish this paragraph with a case study (Bartoszewski et al., 2010) of cystic fibrosis. This disease is commonly caused by a 3-nt deletion in *CFTR*. This deletion causes the loss of one amino acid and a synonymous change. The loss of the amino acid has long been thought to be responsible for the disease, but the synonymous mutation results in a change in the mRNA structure and a decrease in the gene expression level. It is likely that the disease is actually caused by the synonymous change.

I studied intragenic epistasis between two mutations, but epistasis can occur among more than two mutations (Li et al., 2016; Puchta et al., 2016; Sarkisyan et al., 2016). To study higher-order epistasis, we may use the mutant oligo synthesis method that incorporates alternative bases (Li et al., 2016; Puchta et al., 2016). However, there are two limitations of this method. First, the longest oligos that can be synthesized are 200 nucleotides, but most genes are longer than 200 bases, so it is difficult to study epistatic interactions in the whole gene. Second, the smallest fraction of alternative bases that can be incorporated at a site is 3%. Even if we come up with a method to synthesize a long gene, the expected number of mutations in a 1,000-nt gene is at least

10, which makes the results complicated to analyze. Considering these two limitations, we may first study high-order epistasis in short genes.

References

Bartoszewski, R. A., Jablonsky, M., Bartoszewska, S., Stevenson, L., Dai, Q., Kappes, J., . . . Bebok, Z. (2010). A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J Biol Chem*, *285*(37), 28741-28748. https://doi.org/10.1074/jbc.M110.154575

Chen, R., Davydov, E. V., Sirota, M., & Butte, A. J. (2010). Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One*, *5*(10), e13574. https://doi.org/10.1371/journal.pone.0013574

Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J., Hillenmeyer, M. E., Davis, R. W., . . . Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, *169*(4), 1915-1925. https://doi.org/10.1534/genetics.104.036871

Fisher, R. A. (1928). The possible modification of the response of the wild type to recurrent. mutations. *American Naturalist*, *62*, 115-126. https://doi.org/Doi 10.1086/280193

Graur, D., Sater, A. K., & Cooper, T. F. (2016). *Molecular and genome evolution*. Sinauer Associates, Inc.

Kacser, H., & Burns, J. A. (1981). The molecular basis of dominance. *Genetics*, *97*(3-4), 639-666.

Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., . . . Neale, B. M. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom*, *2*(9), 100168. https://doi.org/10.1016/j.xgen.2022.100168

Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife*, *8*. https://doi.org/10.7554/eLife.45952

Li, C., Qian, W., Maclean, C. J., & Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, *352*(6287), 837-840. https://doi.org/10.1126/science.aae0568

Li, W.-H. (1997). *Molecular evolution*. Sinauer Associates.

Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press.

Phadnis, N., & Fry, J. D. (2005). Widespread correlations between dominance and homozygous effects of mutations: Implications for theories of dominance. *Genetics*, *171*(1), 385-392.

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., . . . Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111-1124. https://doi.org/10.1016/j.cell.2015.02.029

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., & Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. *Science*, *352*(6287), 840-844. https://doi.org/10.1126/science.aaf0965

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., . . . Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, *533*(7603), 397-401. https://doi.org/10.1038/nature17995

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet*, *12*(10), 683-691. https://doi.org/10.1038/nrg3051

Sharon, E., Chen, S.-A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, *175*(2), 544-557.e516. https://doi.org/10.1016/j.cell.2018.08.057

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, *15*(3), 1281-1295. https://doi.org/10.1093/nar/15.3.1281

Wright, S. (1934). Physiological and evolutionary theories of dominance. *American Naturalist*, *68*, 24-53. https://doi.org/Doi 10.1086/280521

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C. H., Fu, J., . . . Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A*, *113*(41), E6117-E6125. https://doi.org/10.1073/pnas.1606724113

Appendices

Appendix A: Supplementary Tables and Figures for Chapter 2

Table A-1. Primers used in the study.

Primer name	Sequence (These primers are used to delete the WT sequences)
ADA2 Confirmation F	TGCAGTTGATTAAAGGCGCA
ADA2 Confirmation R	AAATTCTAATCTGCCCGGCA
ADA2 gRNA F	GATCTGTTCCAAAAATTCATCTTGGTTTTAGAGCTAG
ADA2 gRNA R	CTAGCTCTAAAACCAAGATGAATTTTTGGAACA
	GGCAGGATATTGCTGACCATATAGGCAGCAGAGGCAAA
ADA2 HR F	GAAGAAGTTAAGATGCGGGTAGAAGATTACGGAGG
	GAATTCTGTTTCAAATTCTAATCTGCCCGGCATAAACCC
ADA2 HR R	CTGTACTTCATCCTCCGTAATCTTCTACCCGCAT
ASC1 Confirmation F	ACTGCTCCTTTGGTTTTCCT
ASC1 Confirmation R	TGTCCCAAGAAGCAGACAAA
ASC1 gRNA F	GATCGGTACCTTGGAAGGTCACAAGTTTTAGAGCTAG
ASC1 gRNA R	CTAGCTCTAAAACTTGTGACCTTCCAAGGTACC
	AAAATCCTTATAACACACTAAAGTAAATAAAGTGAAAA
ASC1 HR F	ATGGCATCTAACATGCGGGTAGAAGATTACGGAGG
	GTAAGCACCGTCAGCAGTCAAAGTACAGTCTTGGACAA
ASC1 HR R	TGTGACTGTGACCCTCCGTAATCTTCTACCCGCAT
BFR1 Confirmation F	TGCAGAAGCCAAGCAAAGAA

BFR1 Confirmation R	AAGGCAGCACGTTTGTTGAA
BFR1 gRNA F	GATCAAAGTCTGTTGATGCTGACAGTTTTAGAGCTAG
BFR1 gRNA R	CTAGCTCTAAAACTGTCAGCATCAACAGACTTT
	CTGCAGAAGCCAAGCAAAGAATCAATGAGATTGAAGAG
BFR1 HR F	TCTATTGCCTCTATGCGGGTAGAAGATTACGGAGG
	AGTTTTTGAATGGATATCGTTCAATTTTTGCTGGTTCTCT
BFR1 HR R	TCGAATTGGTCCTCCGTAATCTTCTACCCGCAT
BUD23 Confirmation F	GCGTTGGAGCTTTTGAATCT
BUD23 Confirmation R	AACCTCATCAACCGCTGTTT
BUD23 gRNA F	GATCGGTCTTAGTAGAGAGCTGGAGTTTTAGAGCTAG
BUD23 gRNA R	CTAGCTCTAAAACTCCAGCTCTCTACTAAGACC
	TGGATATCGGGTGCGGGTCCGGACTGTCTGGGGAGATTT
BUD23 HR F	TGACGCAGGAGATGCGGGTAGAAGATTACGGAGG
	TGTGTTGAAAAACCTCATCAACCGCTGTTTAGGATCGTT
BUD23 HR R	GTATGAAGTGTCCTCCGTAATCTTCTACCCGCAT
CCW12 Confirmation F	AACGTTACCACTGCTACTGT
CCW12 Confirmation R	TTACAACAACAAAGCAGCGG
CCW12 gRNA F	GATCGATAACGTCATCGACGGTGAGTTTTAGAGCTAG
CCW12 gRNA R	CTAGCTCTAAAACTCACCGTCGATGACGTTATC
	TCGCCGCTGTCGCTTCTGCCGCTGCTAACGTTACCACTG
CCW12 HR F	CTACTGTCAGCATGCGGGTAGAAGATTACGGAGG
	TGGAGCAGCAGAGGTGGTGTTCTTTGGAGCTTCAGTAGA
CCW12 HR R	GGTAACTGGAGCCTCCGTAATCTTCTACCCGCAT

EOS1 Confirmation F	TGAGCGAGCGACCTTTAAAA
EOS1 Confirmation R	ATTCAGACGCTCTTGCTGTA
EOS1 gRNA F	GATCTGGCTCGTAAAGCGTTATGCGTTTTAGAGCTAG
EOS1 gRNA R	CTAGCTCTAAAACGCATAACGCTTTACGAGCCA
	CCTTATGCAGGGACATATCCCTTTTACCGCCGTTAACTT
EOS1 HR F	ACATCTTCACAATGCGGGTAGAAGATTACGGAGG
	GCAAAGGAGATATTCAGACGCTCTTGCTGTAGTGAGTGC
EOS1 HR R	GGATAATAGGACCTCCGTAATCTTCTACCCGCAT
EST1 Confirmation F	ATCACGTTCAGATGCTTCCT
EST1 Confirmation R	TTTGCGCATAGGTGTTCGAT
EST1 gRNA F	GATCTGCTCAAACCATTGGAGTGTGTTTTAGAGCTAG
EST1 gRNA R	CTAGCTCTAAAACACACTCCAATGGTTTGAGCA
	AAAATATGTACCATAATAACAATTACGAACGCATAAAT
EST1 HR F	GATTCCGTGATAATGCGGGTAGAAGATTACGGAGG
	ATTTGAAATAACGGAATTCATATCGTACTTTGCGCATAG
EST1 HR R	GTGTTCGATGACCTCCGTAATCTTCTACCCGCAT
GET1 Confirmation F	AGGTCAAAGAACGTCACGAA
GET1 Confirmation R	TAACCAGCCTTGAGACCAAA
GET1 gRNA F	GATCTTTGTGAAGATGGGCCTGAAGTTTTAGAGCTAG
GET1 gRNA R	CTAGCTCTAAAACTTCAGGCCCATCTTCACAAA
	GAATTAAAAGAATTCAACAACTCTATCTCCGCGCAGGAT
GET1 HR F	AATTATGCCAAATGCGGGTAGAAGATTACGGAGG

	ACACCACTTACAAACGTAGGGAACAAGGTGCTAGTCGA
GET1 HR R	GGAACTCAACTTCCTCCGTAATCTTCTACCCGCAT
GIM5 Confirmation F	TTCGACCAAGAATTGCAGCA
GIM5 Confirmation R	ATGCTGTCTAATAGCGGCTT
GIM5 gRNA F	GATCTCTGCATCATTATACATCCCGTTTTAGAGCTAG
GIM5 gRNA R	CTAGCTCTAAAACGGGATGTATAATGATGCAGA
	TTCGACCAAGAATTGCAGCATTTCACACAGTCCTTGCAA
GIM5 HR F	GCATTAACCATATGCGGGTAGAAGATTACGGAGG
	TTGTTAAGCTTGTCTACTTTCTTTTGGTAAAATGCGATTG
GIM5 HR R	CTGCTTCAGCCCTCCGTAATCTTCTACCCGCAT
IES6 Confirmation F	TGGGCGAACGCAATGAAATT
IES6 Confirmation R	TGCGTTGTGATACCGAATGT
IES6 gRNA F	GATCCAAGAAGACACAAATCAGCGGTTTTAGAGCTAG
IES6 gRNA R	CTAGCTCTAAAACCGCTGATTTGTGTCTTCTTG
	ACGAGAGATTGCTGTTTCTAAGAAGCGTGGGCGAACGC
IES6 HR F	AATGAAATTGGCATGCGGGTAGAAGATTACGGAGG
	AACATCGCAGTACTTCTTGGCAGGCCTGATAGACGGTGG
IES6 HR R	CGCTTCCACGCCCTCCGTAATCTTCTACCCGCAT
LSM1 Confirmation F	ACCAACATTTGCTCCGCTTT
LSM1 Confirmation R	TCTCCACGCAATCTTGAAGT
LSM1 gRNA F	GATCTACTGAGCTTACAATAGCAGGTTTTAGAGCTAG
LSM1 gRNA R	CTAGCTCTAAAACCTGCTATTGTAAGCTCAGTA

	CAAATAGCAAGGACAGAAATCAGTCCAATCAGGATGCG
LSM1 HR F	AAGCGACAACAGATGCGGGTAGAAGATTACGGAGG
	GTATTTGTTTTCTTCGCTAAAATATATTCTCTCCACGCAA
LSM1 HR R	TCTTGAAGTACCTCCGTAATCTTCTACCCGCAT
PAF1 Confirmation F	AGGTATGCCGGTTGATTTGA
PAF1 Confirmation R	TTCACCGGATGTTGCCATTT
PAF1 gRNA F	GATCATATTCGGTGCGTCTCAAGAGTTTTAGAGCTAG
PAF1 gRNA R	CTAGCTCTAAAACTCTTGAGACGCACCGAATAT
	AACTGCTTTACGGCTTTGATAATGTGAAATTGGACAAAG
PAF1 HR F	ATGATCGAATTATGCGGGTAGAAGATTACGGAGG
	TTGCCATTTGTCCGTCTTATTGAATGTTCCTTCGACCCTA
PAF1 HR R	CTAATTATATCCTCCGTAATCTTCTACCCGCAT
PRS3 Confirmation F	TCATCACGCAAATTGGCTCT
PRS3 Confirmation R	AACTACGCTTGGTTCTGCAT
PRS3 gRNA F	GATCATATGCAAGACAAGATAGAAGTTTTAGAGCTAG
PRS3 gRNA R	CTAGCTCTAAAACTTCTATCTTGTCTTGCATAT
	TCTTTATCATCACGCAAATTGGCTCTGGTGTCGTGAACG
PRS3 HR F	ATCGTGTTCTAATGCGGGTAGAAGATTACGGAGG
	ATCTACTGGGACGTCGAAGAACCCTTGAATTTGGGAAG
PRS3 HR R	CATGCAAATCCACCTCCGTAATCTTCTACCCGCAT
RAD6 Confirmation F	AACGTATGAAGGAAGATGCC
RAD6 Confirmation R	TTGCAGCTTCAACGTTTGCT
RAD6 gRNA F	GATCTCCGTCTTCATATGGAGTATGTTTTAGAGCTAG

RAD6 gRNA R	CTAGCTCTAAAACATACTCCATATGAAGACGGA
	GAAGGTTGATGAGAGATTTTAAACGTATGAAGGAAGAT
RAD6 HR F	GCCCCACCGGGTATGCGGGTAGAAGATTACGGAGG
	TGGAGTCCATCTGTTCTGCAAAATATCCAAACAAATTTC
RAD6 HR R	ACCATTTGCATCCTCCGTAATCTTCTACCCGCAT
RPL29 Confirmation F	TCTGCGTACATTCATCGTCT
RPL29 Confirmation R	TGTGGCACATAAGGAGGAAA
RPL29 gRNA F	GATCCGCTCACAACCAAACCAGAAGTTTTAGAGCTAG
RPL29 gRNA R	CTAGCTCTAAAACTTCTGGTTTGGTTGTGAGCG
	GACCATTCGCAATTTCTGCGTACATTCATCGTCTTCTCCA
RPL29 HR F	GAAAATGGCTATGCGGGTAGAAGATTACGGAGG
	AATCAGACAAAATAATATGTAAATTTTTAACGTATTATA
RPL29 HR R	ATCTTAAAAAGCCTCCGTAATCTTCTACCCGCAT
RPL39 Confirmation F	TTGGATCCGTGAATGCATCA
RPL39 Confirmation R	AGGGAAGGATGGAAGACAAA
RPL39 gRNA F	GATCAAAACAGACCATTGCCACAAGTTTTAGAGCTAG
RPL39 gRNA R	CTAGCTCTAAAACTTGTGGCAATGGTCTGTTTT
	TTTACAATTGTACACTTCGTATGTGCACGATATGTTTCCC
RPL39 HR F	TTTTAATTAGATGCGGGTAGAAGATTACGGAGG
	ATGGAAGACAAATGACAAAAAGTTTGAAGCATAAATAT
RPL39 HR R	GTTCTTCGCTTACCTCCGTAATCTTCTACCCGCAT
RPS7A Confirmation F	ACTGCGTTAGAATCCTGGTA
RPS7A Confirmation R	TCGTCCCGTTCACACTTTTT

RPS7A gRNA F	GATCTGAGACAAAATCTTGGCTTGGTTTTAGAGCTAG
RPS7A gRNA R	CTAGCTCTAAAACCAAGCCAAGATTTTGTCTCA
	ATCTTATTTTAAGAAAGCTGAAAGGAAGAAAGATCATC
RPS7A HR F	ACGAACAACATGATGCGGGTAGAAGATTACGGAGG
	TCTGGTATCGTCCCGTTCACACTTTTTTCCTTTGTTACTC
RPS7A HR R	TCCATTATTCCCTCCGTAATCTTCTACCCGCAT
SNF6 Confirmation F	TTCGCGGAGGAAAACAACTA
SNF6 Confirmation R	AACTCTGCCGCTTGTGTTTT
SNF6 gRNA F	GATCAGTTCTTCAAAAGGTTGGATGTTTTAGAGCTAG
SNF6 gRNA R	CTAGCTCTAAAACATCCAACCTTTTGAAGAACT
	GCTCCAGTGCCGGCATGAATGGCAGATCGCTTACGTACG
SNF6 HR F	CGCAGCAACAGATGCGGGTAGAAGATTACGGAGG
	TTGCATGAGATATCTTATTGTTTTTTCACTGAACAATCTG
SNF6 HR R	GACTTCTCAACCTCCGTAATCTTCTACCCGCAT
TSR2 Confirmation F	TTGCAACAGGAAGAAAGGTG
TSR2 Confirmation R	AAGCGGCGTCAACAACTTTT
TSR2 gRNA F	GATCATCCCACTTATAAATAACCAGTTTTAGAGCTAG
TSR2 gRNA R	CTAGCTCTAAAACTGGTTATTTATAAGTGGGAT
	GAACAATGAGCACACAATATATTGATGAGACAGCATTT
TSR2 HR F	GTTCAGGCTGAGATGCGGGTAGAAGATTACGGAGG
	GAACAATGAGCACACAATATATTGATGAGACAGCATTT
TSR2 HR R	GTTCAGGCTGAGATGCGGGTAGAAGATTACGGAGG
VMA21 Confirmation F	ATGTTCCTCGTGCGGTGATT

VMA21 Confirmation R	GGTGCGTTGGAAAAATCAAC
VMA21 gRNA F	GATCGCTGTTTACTGCAGCGATGGGTTTTAGAGCTAG
VMA21 gRNA R	CTAGCTCTAAAACCCATCGCTGCAGTAAACAGC
	AAAGAATCAAATAATGGCTGTAGATGTTCCTCGTGCGGT
VMA21 HR F	GATTAATAAACGACCTTACAGAGGTAATCGCCGG
	CAATCAGTCTTCCTTTTTATTACCATCAACTTTGTGATCT
VMA21 HR R	TCAGTATCCTCCGGCGATTACCTCTGTAAGGTC
VMA7 Confirmation F	ACCAACGTGAATTGCAAGCA
VMA7 Confirmation R	AAATAGCAGGGAACGCATTG
VMA7 gRNA F	GATCAGAAGGTAAGACTACTAAGGGTTTTAGAGCTAG
VMA7 gRNA R	CTAGCTCTAAAACCCTTAGTAGTCTTACCTTCT
	TGGCTGAGAAACGTACTCTTATAGCTGTGATAGCTGACG
VMA7 HR F	AAGATACTACAATGCGGGTAGAAGATTACGGAGG
	TAAAATAGCAGGGAACGCATTGGTGAAGGAGTCCACTC
VMA7 HR R	TAGCTCTTATGTCCTCCGTAATCTTCTACCCGCAT

(Continued)

Primer name	Sequence (sequences within parentheses were mutated. The unmutated flanking regions were used as the primer binding sites. All oligos were synthesized in a pool.)
ADA2	CAGCAGAGGCAAAGAAGAAGTTAAG(GAACATTACCTAAAATATTATCTGGAAAGCAAATACTATCCAATACCTGATATTACCCAAAATATACATGTCCCACAAGATGAATTTTTGGAACAGCGAAGGCATAGAATCGAGTCCTTCCGGGAGAGGCCGCTAGAGCCTCCAAGAAAG)CCCATGGCATCGGTTCCTAGCTGCC
ASC1	AATAAAGTGAAAAATGGCATCTAAC(GAAGTTTTAGTTTTGAGAGGTACCTTGGAAGGTCACAACGGTTGGGTCACATCTTTGGCTACTTCTGCTGGTCAACCAAACCTATTGTTGTCCGCTTCCCGTGATAAGACTTTGATCTCCTGGAAGTTGACTGGTGACGACCAAAAGTTT)GGTGTCCCAGTTAGATCTTTCAAGG
BFR1	TGAGATTGAAGAGTCTATTGCCTCT(GGTGACCTTTCTTTGGTTCAAGAAAAACTACTAGTCAAAGAAATGCAATCTTTGAACAAATTGATTAAGGACTTAGTTAACATCGAGCCAATCAGAAAGTCTGTTGATGCTGACAAGGCTAAAATCAATCAATTGAAGGAAGAATTGAAC)GGATTGAATCCAAAGGATGTCTCCA
BUD23	GTCTGGGGAGATTTTGACGCAGGAG(GGAGACCATGTGTGGTGTGGTTTGGATATATCGCCCAGCATGCTTGCGACCGGTCTTAGTAGAGAGCTGGAGGGCGACTTGATGTTGCAGGATATGGGCACCGGGATACCGT

	TCCGGGCGGGCTCGTTTGACGCGGCTATTAGTATCAGT)GCGATCCA
	ATGGCTGTGCAATGCGG
	TAACGTTACCACTGCTACTGTCAGC(CAAGAATCTACCACTTTGGTC
	ACCATCACTTCTTGTGAAGACCACGTCTGTTCTGAAACTGTCTCCCC
	AGCTTTGGTTTCCACCGCTACCGTCACCGTCGATGACGTTATCACTC
	AATACACCACCTGGTGCCCATTGACCACTGAAGCC)CCAAAGAACG
CCW12	GTACTTCTACTGCTG
	ACCGCCGTTAACTTACATCTTCACA(TCTCTGCGGAAGGCTTGGAGA
	GTCTCCATGCGCACCAGCATAACGCTTTACGAGCCACAATCACTAA
	GAGATGCGTTCACTTATTTCTGGCAAAAACTCAATAGCGCTTACGA
	CAATAACTCATCATTTGAAGGAGCTTCGCAAAAGGCT)GTGAATGG
EOS1	CGACGGTAAGGATTCAC
	CGAACGCATAAATGATTCCGTGATA(CCATTGGTTCTGAAACTTTTA
	TGGCTTCAAATTCACGAACCTACACTCCAATGGTTTGAGCACTGGT
	TCCATGATATCATGCGACTAAGTAACAGAAGAAAGTTCAGAGTTTT
	TAGAATTTTTCAAAAAAAAATGATTCAATTTTTCAAA)ATTACACAC
EST1	AGGTATTACTATGACA
	TCTCCGCGCAGGATAATTATGCCAA(ATGGACTAAGAACAATAGAA
	AATTGGACTCGTTAGATAAAGAAATAAATAACTTGAAGGACGAAA
	TACAATCAGAAAATAAAGCCTTTCAGGCCCATCTTCACAAACTCAG
	GTTATTGGCATTGACGGTGCCATTTTTTGTGTTTAAGAT)TATGTAC
GET1	GGCAAGACACCAGTTTAC

	CACAGTCCTTGCAAGCATTAACCAT(GGCTAAGGGCAAGTTCACAG
	AATGTATTGATGATATTAAAACAGTCTCCCAAGCAGGAAATGAAG
	GGCAAAAACTACTGGTTCCAGCATCTGCATCATTATACATCCCAGG
	TAAGATTGTAGACAATAAGAAATTCATGGTCGACATTGG)TACAGG
GIM5	ATATTACGTTGAAAAGAGC
	CGTGGGCGAACGCAATGAAATTGGC(TTTCCCTCTAGATTCAAGTCG
	GCGCATTACAAGAAACCGACAAGAAGACACAAATCAGCGAGGCAG
	TTGATCTCGGACGAAAACAAGCGGATCAACGCCTTGTTGACCAAGG
	CTAACAAAGCTGCAGAGAGTTCTACTGCTGCTAGGCGA)CTTGTGC
IES6	CCAAAGCGACGTACTTTA
	CAATCAGGATGCGAAGCGACAACAG(CAGAATTTCCCAAAGAAGAT
	TTCAGAAGGTGAGGCCGATTTATATCTCGACCAGTATAACTTCACT
	ACCACCGCTGCTATTGTAAGCTCAGTAGACCGTAAAATCTTCGTTC
	TTTTGCGTGATGGAAGAATGCTATTCGGTGTACTAAGA)ACCTTTGA
LSM1	CCAATATGCAAATTTGA
	GAAATTGGACAAAGATGATCGAATT(TTACTGAGGGACCCTAGAAT
	AGATAGACTGACCAAGACTGATATATCAAAGGTTACCTTCTTGAGA
	CGCACCGAATATGTCTCCAATACAATTGCAGCCCATGATAACACAT
	CGTTGAAAAGGAAAAGGCGCTTGGATGATGGAGATTCG)GATGATG
PAF1	AAAACCTTGATGTTAATC
	TGGTGTCGTGAACGATCGTGTTCTA(GAACTACTGATCATGATCAAT
	GCTTCGAAGACTGCGTCTGCAAGAAGAATCACTGCTATTATTCCAA
PRS3	ATTTCCCATATGCAAGACAAGATAGAAAGGATAAGTCTCGTGCTCC

	TATTACCGCTAAATTGATGGCCGATATGTTAACAACA)GCCGGGTGT
	GATCATGTTATTACTA
	TATGAAGGAAGATGCCCCACCGGGT(GTATCTGCTTCACCATTACCT
	GATAACGTCATGGTATGGAACGCCATGATTATCGGGCCAGCCGATA
	CTCCATATGAAGACGGAACTTTTAGGTTATTGTTGGAGTTTGATGA
	AGAATATCCCAATAAGCCACCGCATGTCAAATTTTTG)AGTGAAAT
RAD6	GTTTCATCCCAATGTCT
	TCATCGTCTTCTCCAGAAAATGGCT(AAGTCTAAGAACCATACCGCT
	CACAACCAAACCAGAAAGGCTCACAGAAACGGTATCAAGAAGCCA
	AAGACCTACAAGTACCCTTCTTTGAAAGGTGTTGATCCAAAGTTTA
	GAAGAAACCACAAGCATGCCCTACACGGCACTGCTAAG)GCTTTGG
RPL29	CTGCTGCCAAGAAATAAA
	GCACGATATGTTTCCCTTTTAATTAG(GCTCAAAAGTCTTTCAGAAT
	CAAGCAAAAAATGGCTAAGGCTAAGAAGCAAAACAGACCATTGCC
	ACAATGGATCAGATTGAGAACCAACAACACTATCCGTTACAACGCT
	AAGAGAAGAAACTGGAGAAGAACCAAGATGAACATC)TAAGCGAA
RPL39	GAACATATTTATGCTTCAA
	AAAGGAAGAAAGATCATCACGAACAACATG(TCTGCTCCACAAGCC
	AAGATTTTGTCTCAAGCTCCAACTGAATTGGAATTACAAGTTGCTC
	AAGCTTTCGTTGAATTGGAAAATTCTTCTCCAGAATTGAAAGCTGA
	GTTGAGACCTTTGCAATTCAAGTCCATCAGAGAA)GTATGTTATTAA
RPS7A	TTTGAATCTAAACTTAA

	ATCGCTTACGTACGCGCAGCAACAG(CTTAATAAGCAAAGACAGGA
	CTTCGAACGTGTACGACTTAGACCAGAACAGCTCAGCAATATCATA
	CATGACGAGAGCGACACGATATCGTTCCGATCCAACCTTTTGAAGA
	ACTTTATAAGCTCGAACGACGCATTTAACATGCTGAGT)TTGACCAC
SNF6	GGTACCGTGCGACAGAA
	GTTCCTCGTGCGGTGATTAATAAAC(TTATGCTGTTTACTGCAGCGA
	TGGTGGTACTGCCCGTACTCACTTTTTTCATTATTCAGCAATTTACG
	CCAAATACCTTAATTAGTGGAGGTTTAGCTGCTGCAATGGCCAATG
	TTGTTCTAATCGTTTACATTGTTGTAGCGTTCCGCG)AGGATACTGA
VMA21	AGATCACAAAGTTGA
	TGAGACAGCATTTGTTCAGGCTGAG(CAAGGTAAAACCAATCTAAT
	GTTCTCTGACGAAAAGCAACAGGCACGTTTTGAGCTCGGTGTTTCC
	ATGGTTATTTATAAGTGGGATGCGTTGGATGTTGCCGTAGAAAACA
	GTTGGGGTGGTCCAGACTCAGCTGAGAAGAGAGACTGG)ATTACAG
TSR2	GGATTGTAGTAGACCTTT
	TGTGATAGCTGACGAAGATACTACA(ACTGGTTTATTGTTAGCCGGG
	ATTGGACAAATCACTCCTGAAACCCAAGAAAAGAACTTTTTTGTTT
	ACCAAGAAGGTAAGACTACTAAGGAGGAAATCACTGACAAGTTTA
	ATCACTTTACTGAAGAGAGAGACGATATTGCCATCCTT)CTAATCAA
VMA7	CCAACATATCGCGGAAA

(Continued)

Primer name	Sequence (Amp F and Amp R were used to amplify the mutant oligos from the mutant oligo pool.)
ADA2 amp F	GGCAGGATATTGCTGACCATATAGGCAGCAGAGGCAAAGAAGAAGTTAAG
ADA2 amp R	GAATTCTGTTTCAAATTCTAATCTGCCCGGCATAAACCCCTGTACTTCATGGCAGCTAGGAACCGATGCCATGGG
ASC1 amp F	AAAATCCTTATAACACACTAAAGTAAATAAAGTGAAAAATGGCATCTAAC
ASC1 amp R	GTAAGCACCGTCAGCAGTCAAAGTACAGTCTTGGACAATGTGACTGTGACCCTTGAAAGATCTAACTGGGACACC
BFR1 amp F	CTGCAGAAGCCAAGCAAAGAATCAATGAGATTGAAGAGTCTATTGCCTCT
BFR1 amp R	AGTTTTTGAATGGATATCGTTCAATTTTTGCTGGTTCTCTTCGAATTGGTTGGAGACATCCTTTGGATTCAATCC
BUD23 amp F	TGGATATCGGGTGCGGGTCCGGACTGTCTGGGGAGATTTTGACGCAGGAG
BUD23 amp R	TGTGTTGAAAAACCTCATCAACCGCTGTTTAGGATCGTTGTATGAAGTGTCCGCATTGCACAGCCATTGGATCGC
CCW12 amp F	TCGCCGCTGTCGCTTCTGCCGCTGCTAACGTTACCACTGCTACTGTCAGC
CCW12 amp R	TGGAGCAGCAGAGGTGGTGTTCTTTGGAGCTTCAGTAGAGGTAACTGGAGCAGCAGTAGAAGTACCGTTCTTTGG

	CCTTATGCAGGGACATATCCCTTTTACCGCCGTTAACTTACATCTT
EOS1 amp F	CACA
	GCAAAGGAGATATTCAGACGCTCTTGCTGTAGTGAGTGCGGATA
EOS1 amp R	ATAGGAGTGAATCCTTACCGTCGCCATTCAC
	AAAATATGTACCATAATAACAATTACGAACGCATAAATGATTCC
EST1 amp F	GTGATA
	ATTTGAAATAACGGAATTCATATCGTACTTTGCGCATAGGTGTTC
EST1 amp R	GATGATGTCATAGTAATACCTGTGTGTAAT
	GAATTAAAAGAATTCAACAACTCTATCTCCGCGCAGGATAATTAT
GET1 amp F	GCCAA
	ACACCACTTACAAACGTAGGGAACAAGGTGCTAGTCGAGGAACT
GET1 amp R	CAACTTGTAAACTGGTGTCTTGCCGTACATA
	TTCGACCAAGAATTGCAGCATTTCACACAGTCCTTGCAAGCATTA
GIM5 amp F	ACCAT
	TTGTTAAGCTTGTCTACTTTCTTTTGGTAAAATGCGATTGCTGCTT
GIM5 amp R	CAGCGCTCTTTTCAACGTAATATCCTGTA
	ACGAGAGATTGCTGTTTCTAAGAAGCGTGGGCGAACGCAATGAA
IES6 amp F	ATTGGC
	AACATCGCAGTACTTCTTGGCAGGCCTGATAGACGGTGGCGCTTC
IES6 amp R	CACGCTAAAGTACGTCGCTTTGGGCACAAG
	CAAATAGCAAGGACAGAAATCAGTCCAATCAGGATGCGAAGCGA
LSM1 amp F	CAACAG

	GTATTTGTTTTCTTCGCTAAAATATATTCTCTCCACGCAATCTTGA
LSM1 amp R	AGTATCAAATTTGCATATTGGTCAAAGGT
	AACTGCTTTACGGCTTTGATAATGTGAAATTGGACAAAGATGATC
PAF1 amp F	GAATT
	TTGCCATTTGTCCGTCTTATTGAATGTTCCTTCGACCCTACTAATT
PAF1 amp R	ATATGATTAACATCAAGGTTTTCATCATC
	TCTTTATCATCACGCAAATTGGCTCTGGTGTCGTGAACGATCGTG
PRS3 amp F	TTCTA
	ATCTACTGGGACGTCGAAGAACCCTTGAATTTGGGAAGCATGCA
PRS3 amp R	AATCCATAGTAATAACATGATCACACCCGGC
	GAAGGTTGATGAGAGATTTTAAACGTATGAAGGAAGATGCCCCA
RAD6 amp F	CCGGGT
	TGGAGTCCATCTGTTCTGCAAAATATCCAAACAAATTTCACCATT
RAD6 amp R	TGCATAGACATTGGGATGAAACATTTCACT
	GACCATTCGCAATTTCTGCGTACATTCATCGTCTTCTCCAGAAAA
RPL29 amp F	TGGCT
	AATCAGACAAAATAATATGTAAATTTTTAACGTATTATAATCTTA
RPL29 amp R	AAAAGTTTATTTCTTGGCAGCAGCCAAAGC
	TTTACAATTGTACACTTCGTATGTGCACGATATGTTTCCCTTTTAA
RPL39 amp F	TTAG
	TACATATATATTGAGAATAAGGGAAGGATGGAAGACAAATGACA
RPL39 amp R	AAAAGTTTGAAGCATAAATATGTTCTTCGCTTA

	ATCTTATTTTAAGAAAGCTGAAAGGAAGAAAGATCATCACGAAC
RPS7A amp F	AACATG
	TCTGGTATCGTCCCGTTCACACTTTTTTCCTTTGTTACTCTCCATTA
RPS7A amp R	TTCTTAAGTTTAGATTCAAATTAATAACATAC
	GCTCCAGTGCCGGCATGAATGGCAGATCGCTTACGTACGCGCAG
SNF6 amp F	CAACAG
	TTGCATGAGATATCTTATTGTTTTTTCACTGAACAATCTGGACTTC
SNF6 amp R	TCAATTCTGTCGCACGGTACCGTGGTCAA
	GAACAATGAGCACACAATATATTGATGAGACAGCATTTGTTCAG
TSR2 amp F	GCTGAG
	AAGTAACGTTTCTTCGATTAAAGCGGCGTCAACAACTTTTTCATT
TSR2 amp R	TTTGAAAAGGTCTACTACAATCCCTGTAAT
	AAAGAATCAAATAATGGCTGTAGATGTTCCTCGTGCGGTGATTAA
VMA21 amp F	TAAAC
	TTCTCTTCTAGCAACATATACTACTCAATCAGTCTTCCTTTTTATT
VMA21 amp R	ACCATCAACTTTGTGATCTTCAGTATCCT
	TGGCTGAGAAACGTACTCTTATAGCTGTGATAGCTGACGAAGAT
VMA7 amp F	ACTACA
	TAAAATAGCAGGGAACGCATTGGTGAAGGAGTCCACTCTAGCTC
VMA7 amp R	TTATGTTTTCCGCGATATGTTGGTTGATTAG

(Continued)

Primer name	Sequence (these primers were used for gene-specific reverve treanscription.)
ADA2_RT_R	AACCCCTGTACTTCATGGCA
ASC1_RT_R	AATGTGACTGTGACCCTTGA
BFR1_RT_R	GGTTGGAGACATCCTTTGGA
BUD23_RT_R	ATCGTTGTATGAAGTGTCCG
CCW12_RT_R	ACTGGAGCAGCAGTAGAAGT
EOS1_RT_R	TGAGTGCGGATAATAGGAGT
EST1_RT_R	AGGTGTTCGATGATGTCATAG
GET1_RT_R	AGTCGAGGAACTCAACTTGT
GIM5_RT_R	TGCTTCAGCGCTCTTTTCAA
IES6_RT_R	TTCCACGCTAAAGTACGTCG
LSM1_RT_R	ATTCTCTCCACGCAATCTTG
PAF1_RT_R	TGAATGTTCCTTCGACCCTA
PRS3_RT_R	ACCCTTGAATTTGGGAAGCATG
RAD6_RT_R	TGTTGGAGTCCATCTGTTCT
RPL29_RT_R	TTCTTGGCAGCAGCCAAAGC
RPL39_RT_R	GAAGCATAAATATGTTCTTCGCTT
RPS7A_RT_R	TTACCACCAGCAACGTCGAT
SNF6_RT_R	TCACTGAACAATCTGGACTTC
TSR2_RT_R	TCTTCGATTAAAGCGGCGTC
VMA21_RT_R	TACCATCAACTTTGTGATCTTC

VMA7_RT_R	AGGAGTCCACTCTAGCTCTT

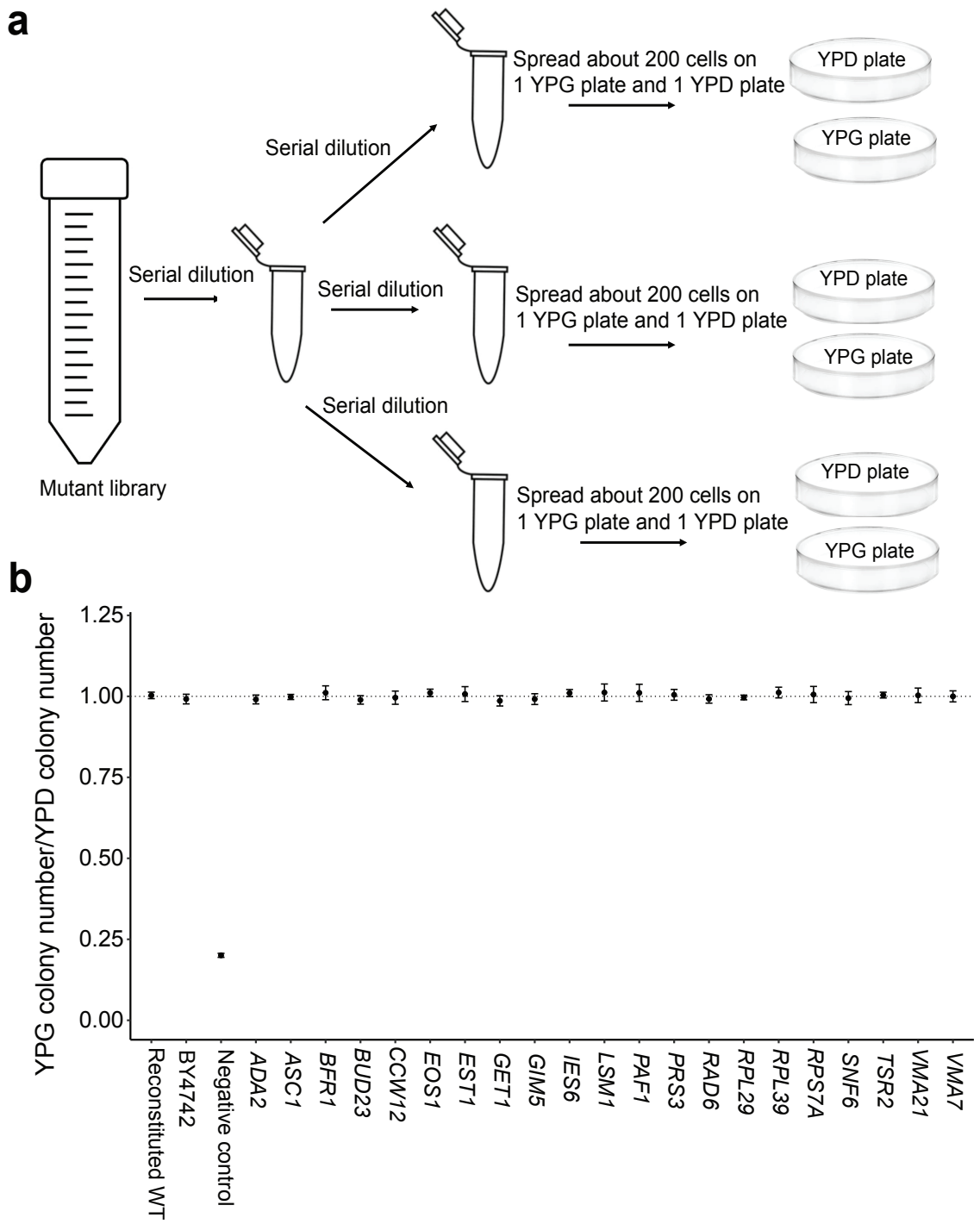

Fig. A-1. Respiration function of mutant cells. a. Experimental procedure for testing cellular respiratory functions. Cells from each of the 21 mutant libraries were spread on YPD and YPG plates, followed by colony counting after growth. Respiration is needed for cell growth on YPG but not on YPD. **b.** Mean ratio of YPD colony number to YPG colony number for each mutant library, based on three replicates per library. Error bars show the standard error of the mean. The negative control is deficient in respiration due to gene deletions (see Methods).

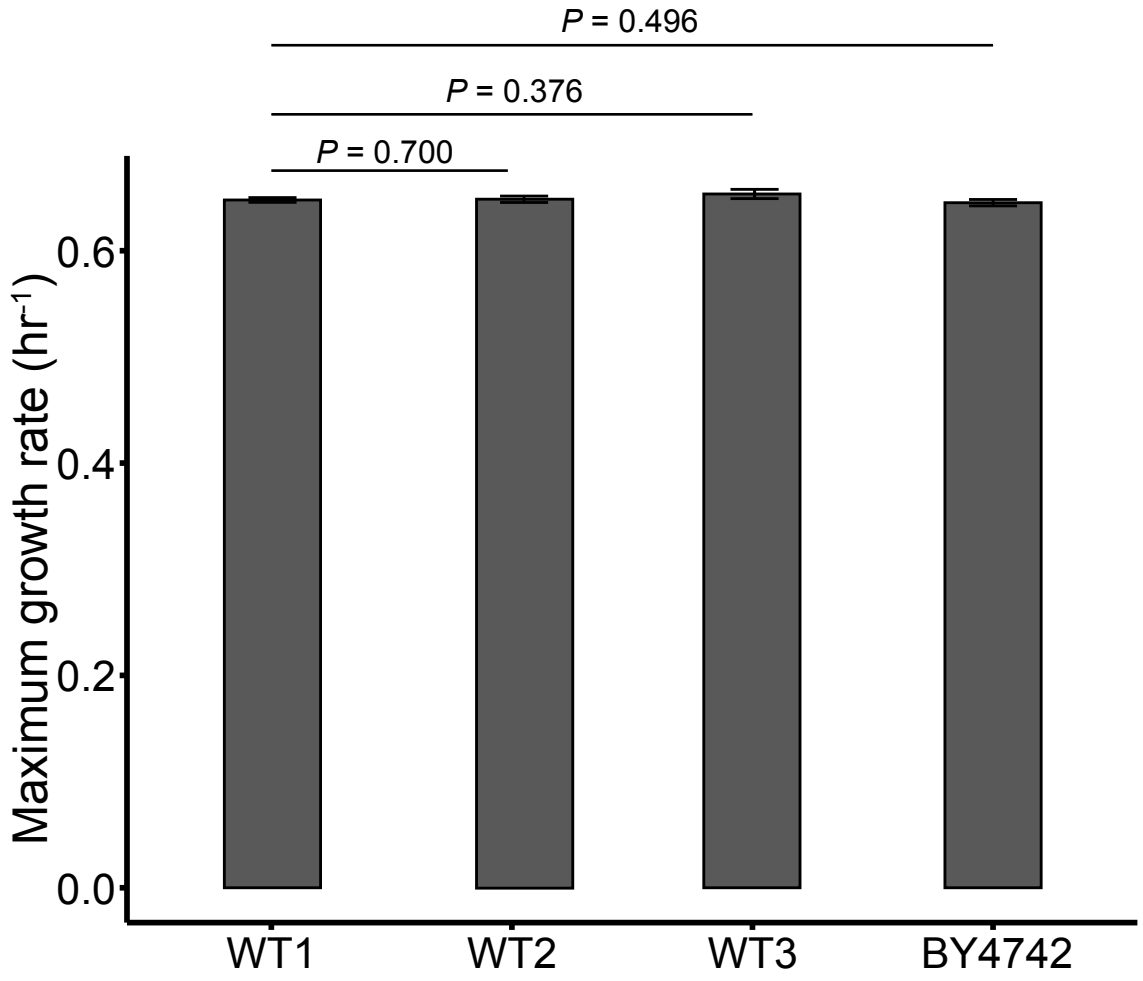

Fig. A- 2. The maximum growth rates of three reconstituted wild-type strains and BY4742. WT1 was used as the wild-type control in *en masse* competitions with mutants. Error bar shows the standard error of the mean based on eight replicates. *P*-values are from *t*-tests. The growth rate is not significantly different among the four strains ($P = 0.58$, one-factor ANOVA test).

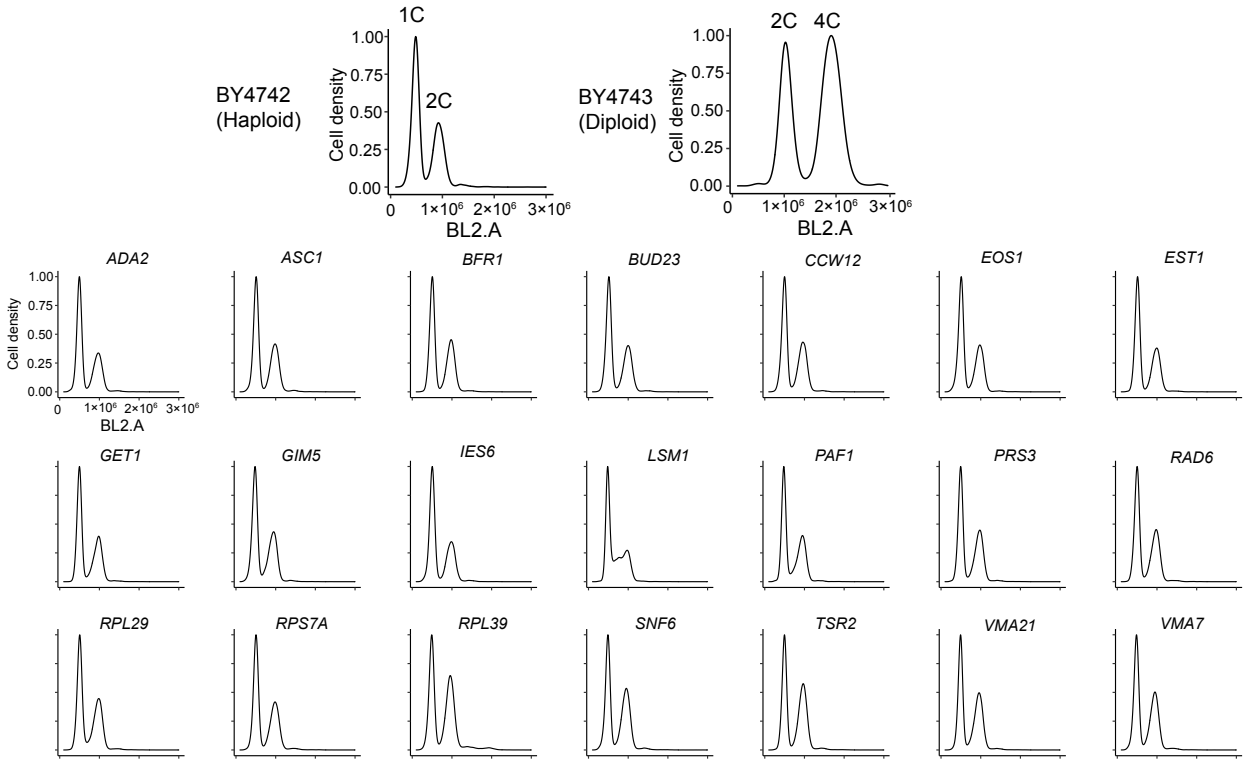

Fig. A- 3.Ploidy of one T48 population per mutant library assessed by flow cytometry. SYTOX Green fluorescence was analyzed using the BL2 detector that measured the output from the 488-nm laser (blue). In control flow cytometry profiles, the two peaks respectively represent cells in the G1 and G2/M cell-cycle stages (1C and 2C DNA content for haploids while 2C and 4C for diploids).

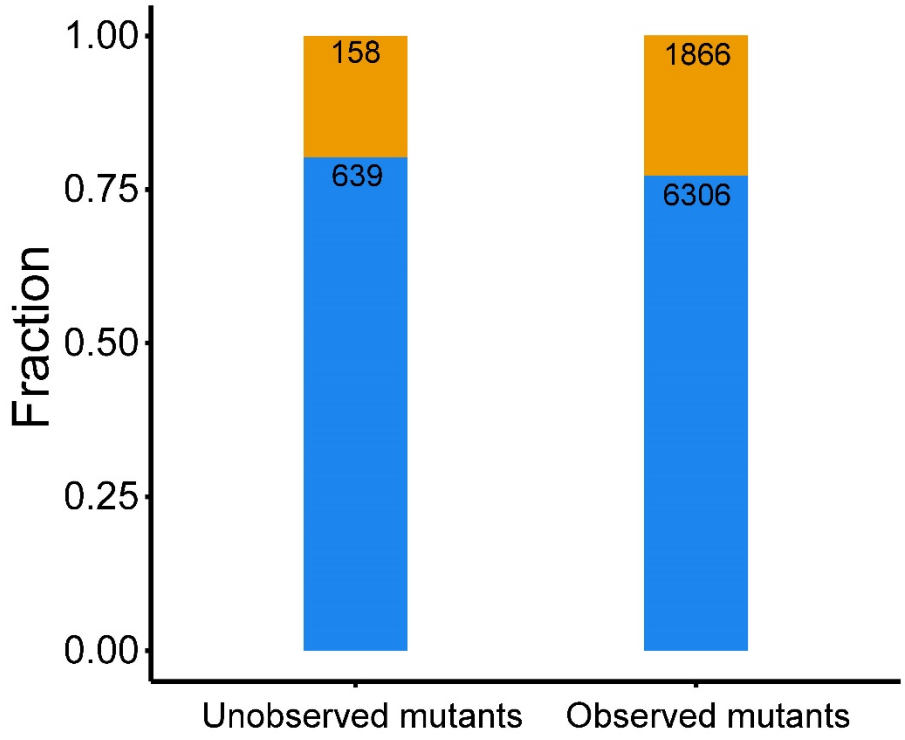

Fig. A-4. Fractions of synonymous (yellow) and nonsynonymous (blue) mutants among designed but unobserved mutants and those among observed mutants. Nonsense mutants are not considered. Numbers in the bars are numbers of mutants. The distributions of synonymous and nonsynonymous mutants among the unobserved and observed mutant groups are not significantly different ($P > 0.05$, Fisher's exact test).

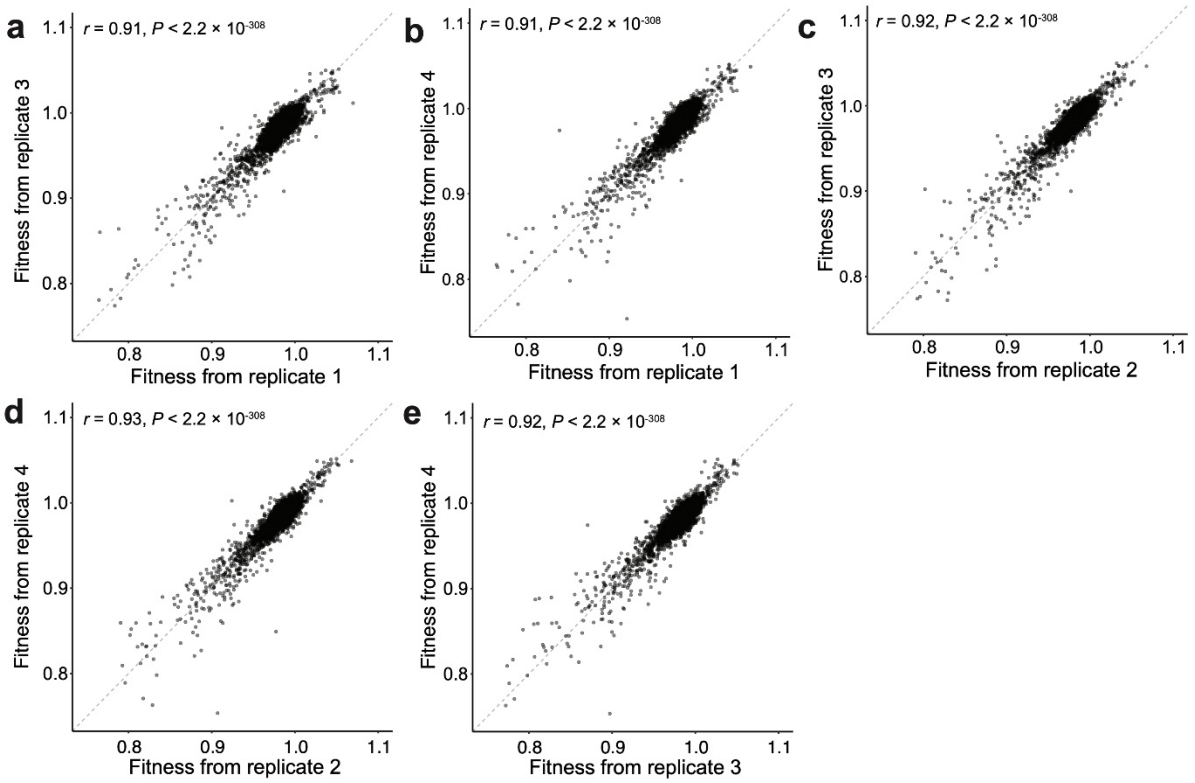

Fig. A-5. Correlation between every two of the four replicates in estimated mutant fitness under YPD at 30°C. The correlation between replicate 1 and replicate 2 is presented in Fig. 1c. Each dot is a mutant and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented. Among-genotype sum of squares explains 93.8% of the total sum of squares (one-factor ANOVA).

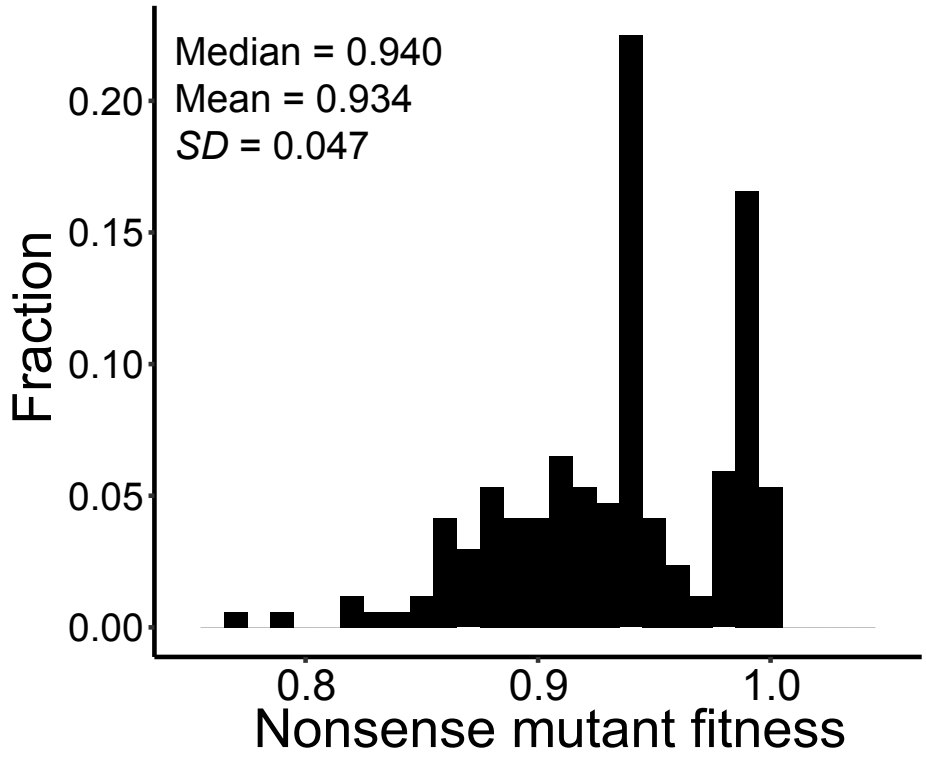

Fig. A-6. Distribution of the fitness of 169 nonsense mutants under YPD at 30°C. The peak around 0.94 is caused by 26 nonsense mutants of *GET1* that all have fitness of about 0.94.

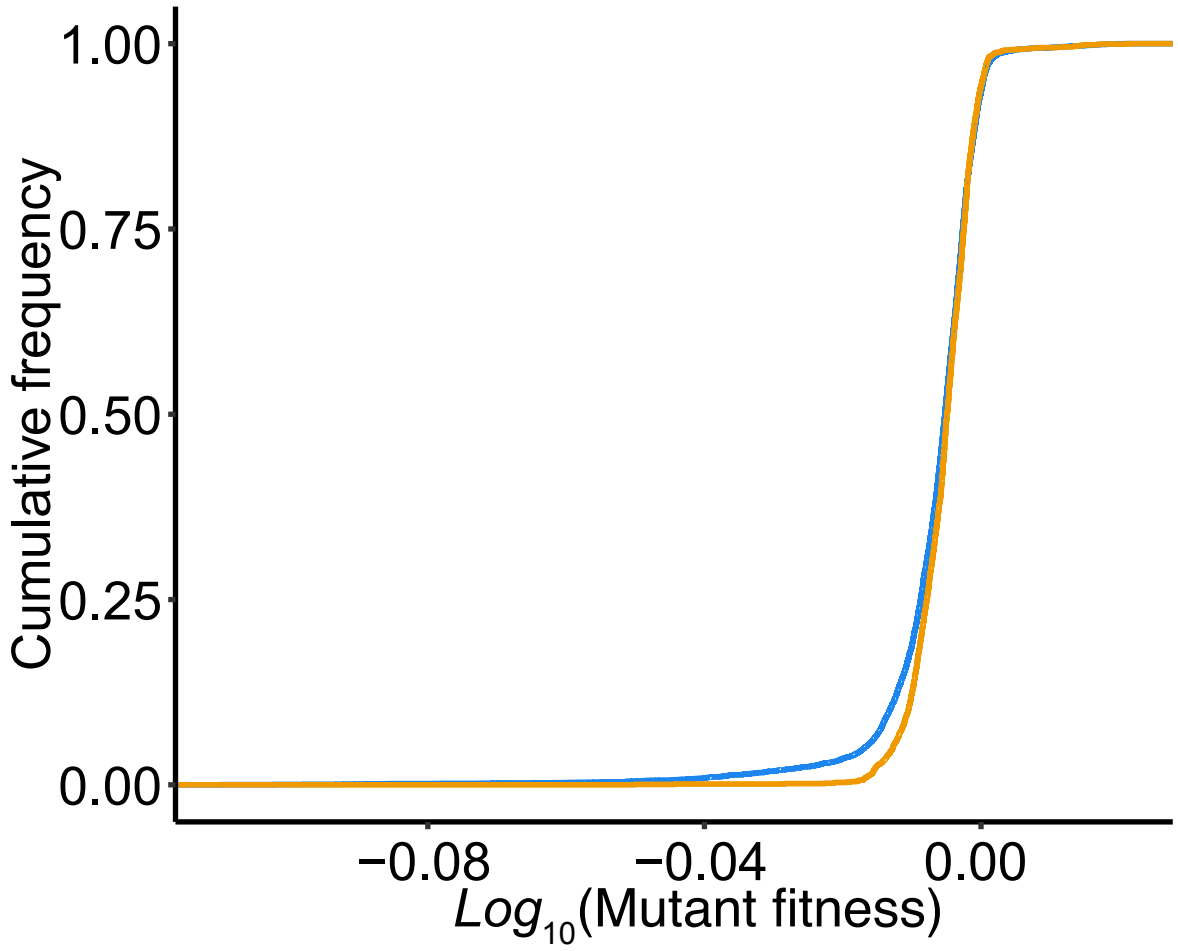

Fig. A-7. Cumulative frequency distributions of log10(mutant fitness) of nonsynonymous (blue) and synonymous (yellow) mutants.

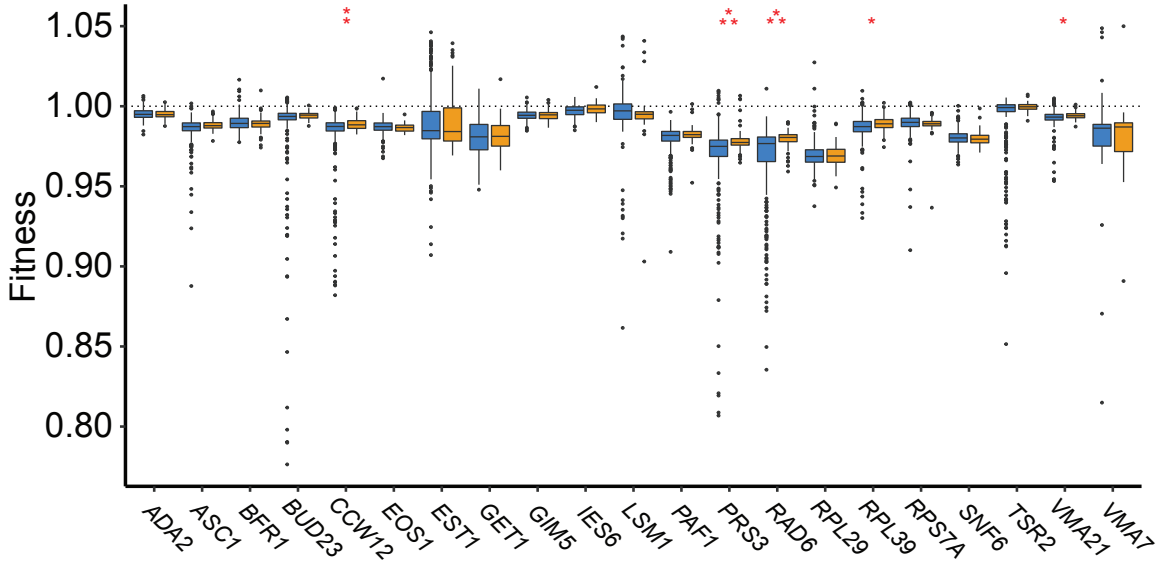

Fig. A- 8. The full figure of Fig. 2c, including low-fitness mutants that are not shown in Fig. 2c.

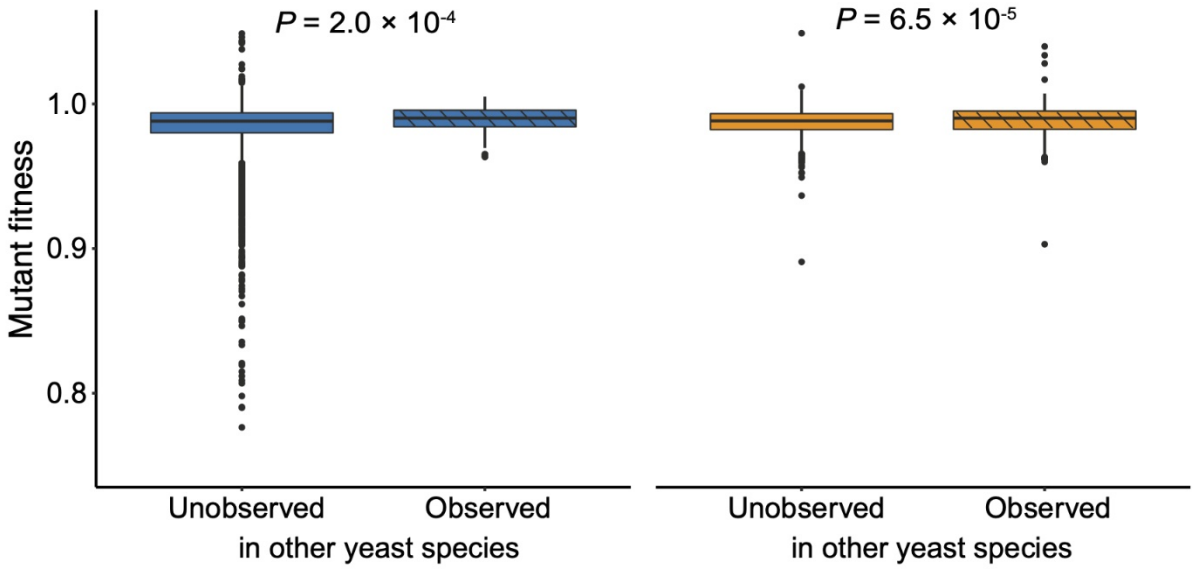

Fig. A-9. The full figure of Fig. 1-2e, including low-fitness and high-fitness mutants that are not shown in Fig. 1-2e.

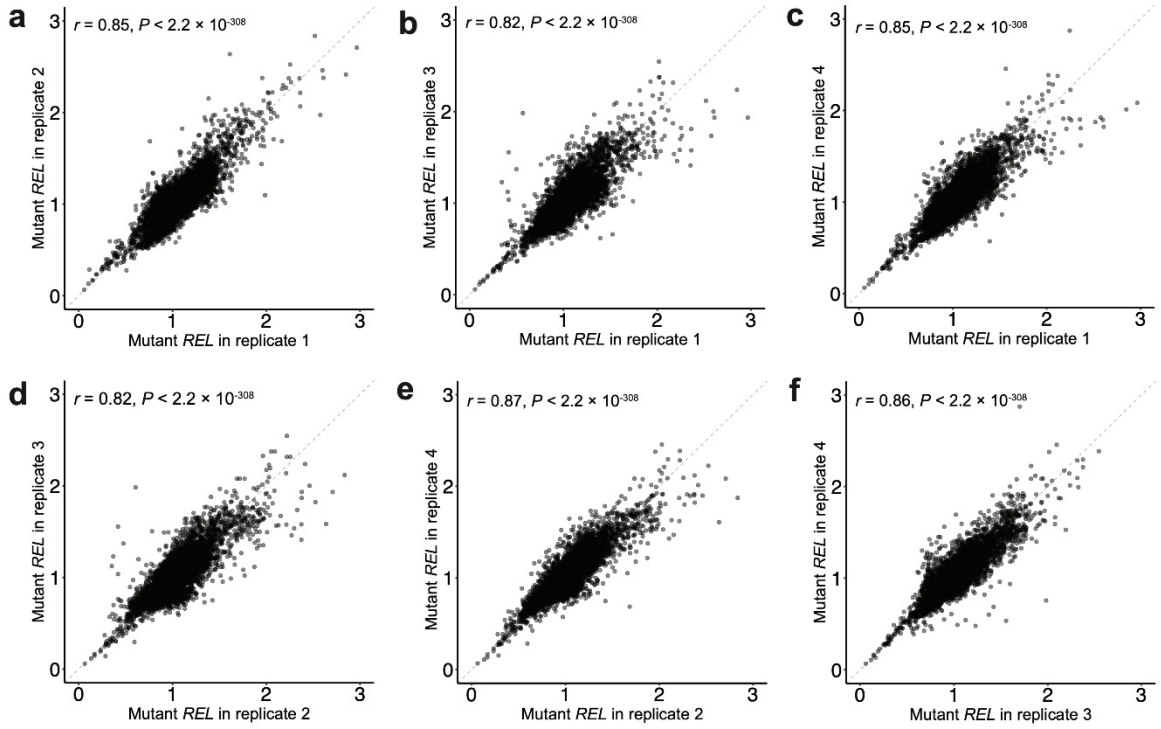

Fig. A-10. Correlation in mutant REL between replicates. Each dot is a mutant, and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented. Among-genotype sum of squares explains 89.7% of total sum of squares (one-factor ANOVA).

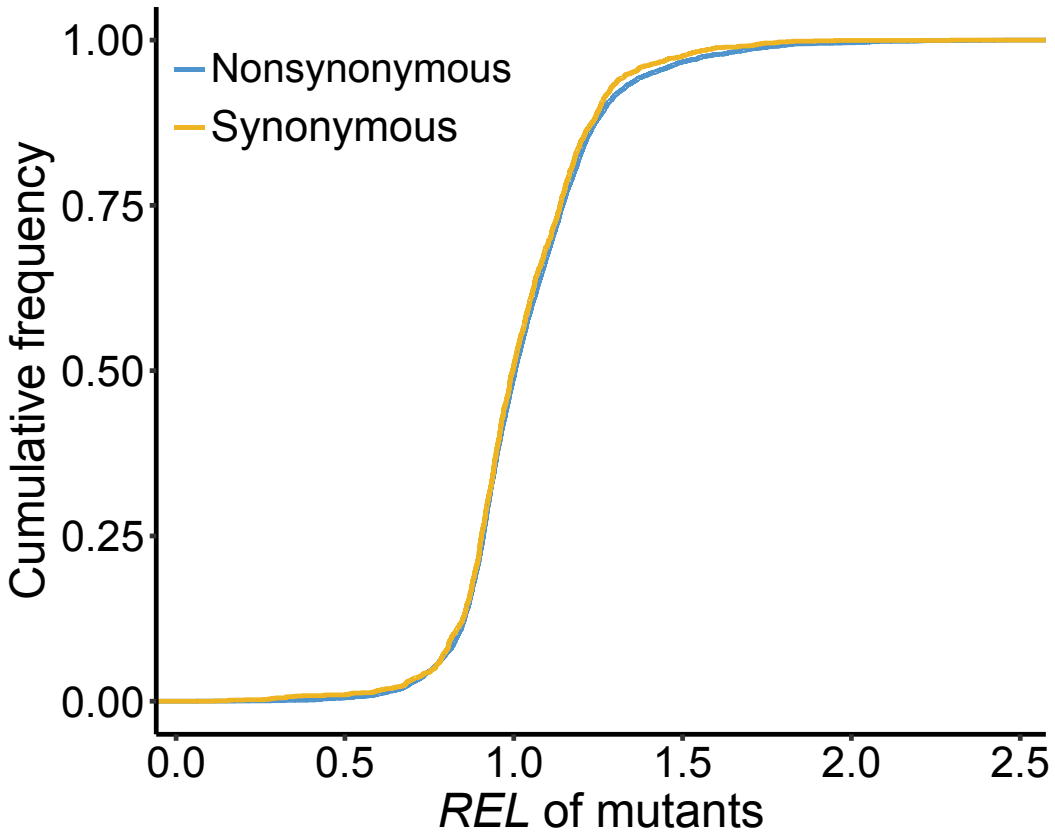

Fig. A-11. Cumulative frequency distributions of REL of nonsynonymous and synonymous mutants.

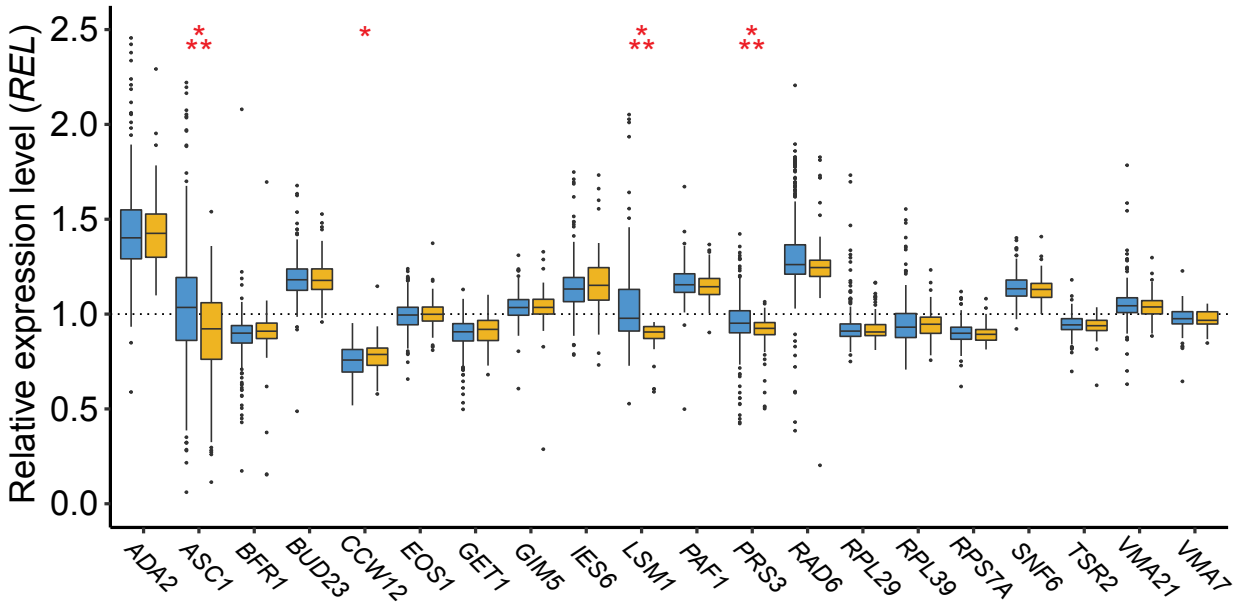

Fig. A-12. Relative expression level (REL) distributions of nonsynonymous (blue) and synonymous (yellow) mutants of 20 individual genes shown by box plots. Nonsynonymous and synonymous distributions of each gene are compared by a Wilcoxon rank-sum test, with FDR-adjusted *P*-values indicated as follows: *, $P < 0.05$; ⁑, $P < 0.01$, ⁂, $P < 0.001$.

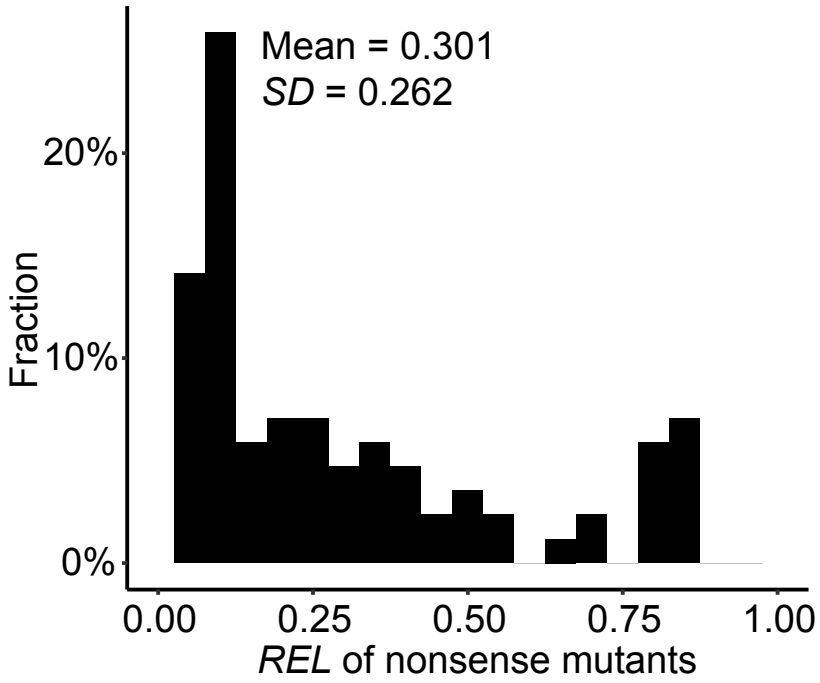

Fig. A-13. Distribution of *REL* of nonsense mutants.

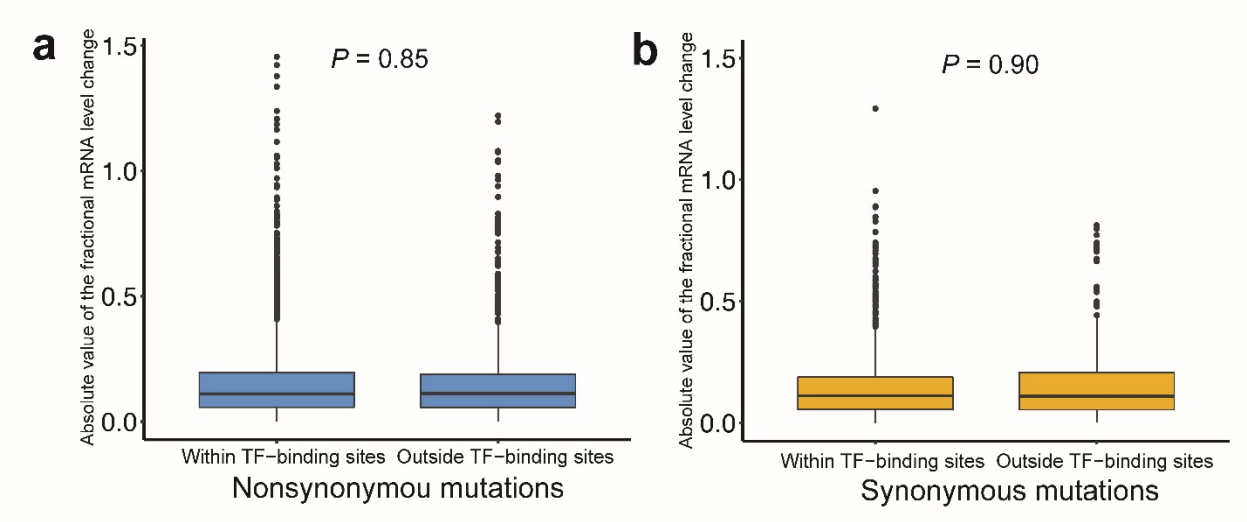

Fig. A-14. Coding mutations within and outside TF-binding sites cause similar absolute fractional changes in the mRNA level shown by box plots. a, Nonsynonymous mutations. **b**, Synonymous mutations. *P*-values are from Wilcoxon rank-sum tests.

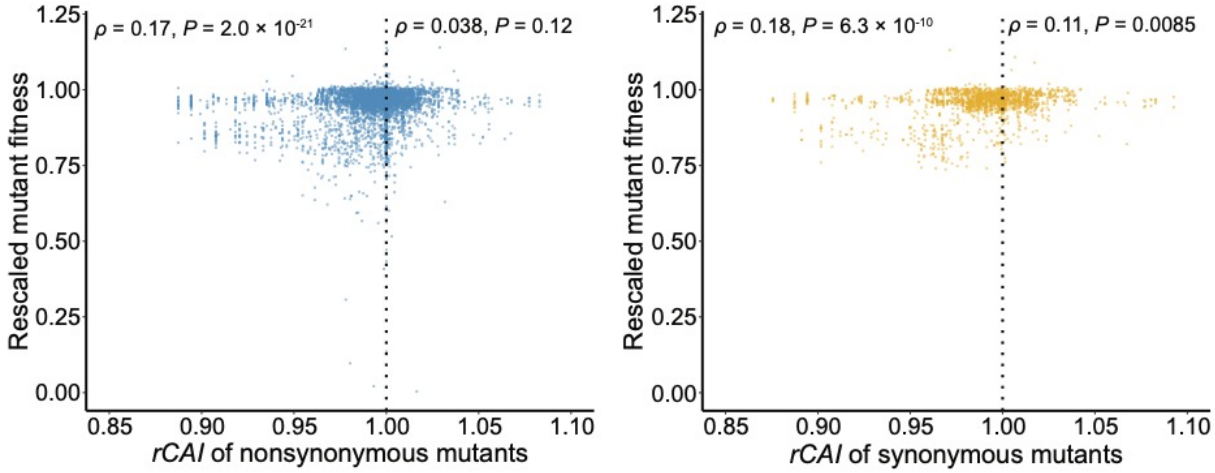

Fig. A-15. Positive correlation between *rCAI* and rescaled fitness among nonsynonymous and synonymous mutants, respectively.

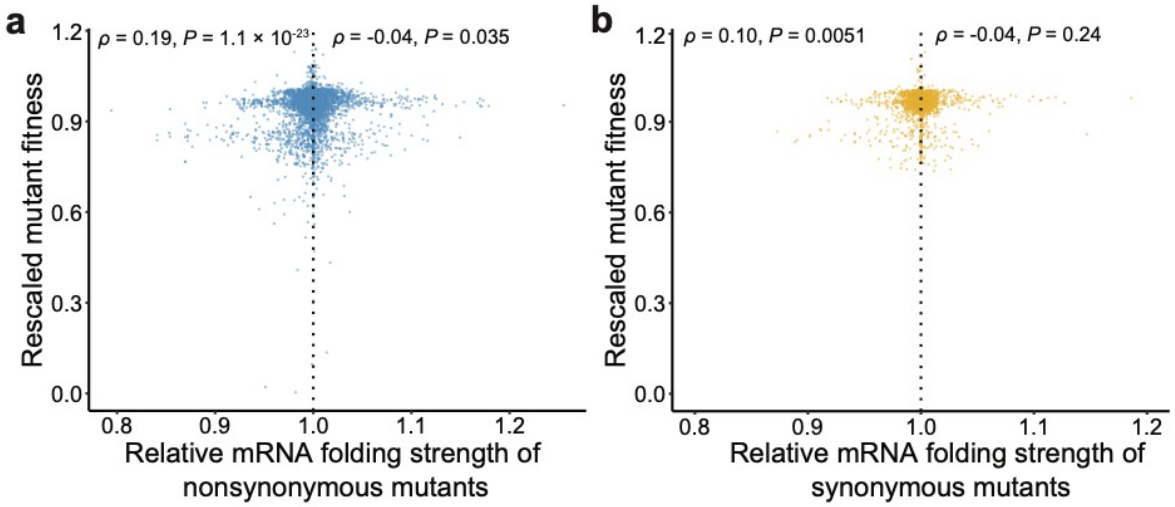

Fig. A-16. Positive correlation between the relative mRNA folding strength (*rMFS*) of a mutant and its rescaled fitness when *rMFS* is below 1. The *rMFS* of a mutant is its mRNA folding strength (i.e., the absolute value of its minimal folding energy) divided by that of the wild-type. In each panel, the correlation is separately computed for mutants with *rMFS* <1 and those with *rMFS* >1. **a**, Nonsynonymous mutants. **b**, Synonymous mutants. Spearman's correlation (ρ) and associated *P*-value are presented.

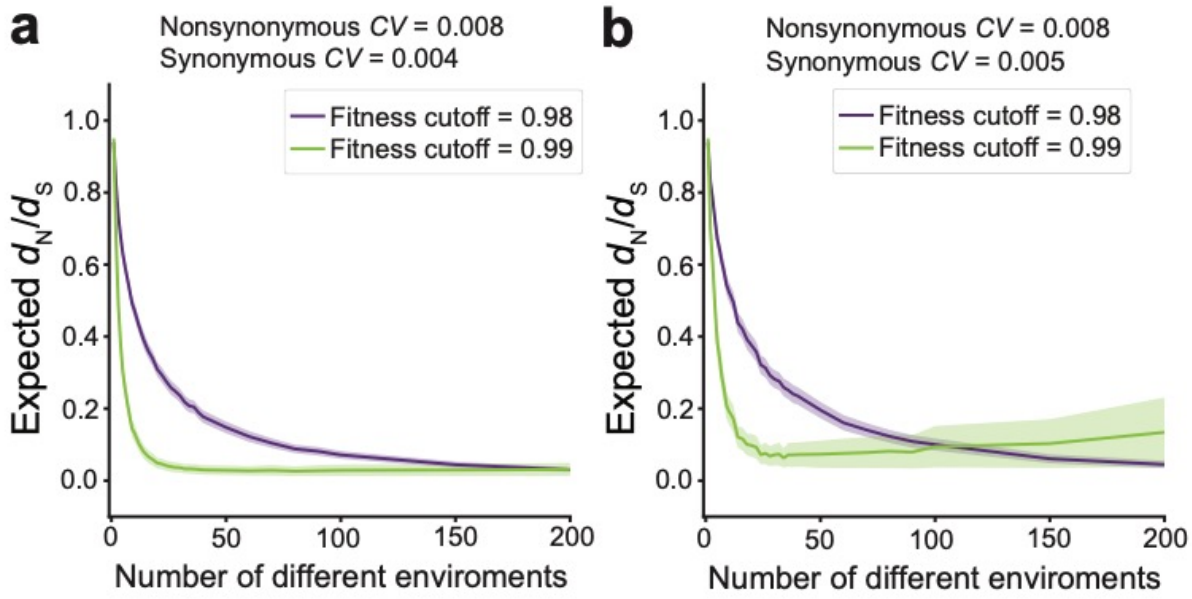

Fig. A-17. Simulation confirms the outcome of d_N/d_S <<1 when the number of environments increases and the across-environment fitness *CV* is higher for nonsynonymous than synonymous mutants. A mutant is purged if its fitness is lower than a preset cutoff such as 0.98 or 0.99 in any environment. The shaded area is the 95% confidence interval. **a**. Results with *CV* = 0.004 for synonymous mutants. **b**, Results with *CV* = 0.005 for synonymous mutants. Note that, under the fitness cutoff of 0.99, d_N/d_S starts to increase with the number (*m*) of environments when *m* is large. Raising *m* reduces the *f*raction of synonymous mutations that are *a*lways *n*eutral (FAN_S) as well as the fraction of nonsynonymous mutations that are always neural (FAN_N). Because the fitness *CV* is larger for nonsynonymous than synonymous mutants in the simulation, FAN_N decreases with *m* more quickly than does FAN_S when *m* is small. When *m* is large, FAN_N is small, making it possible for FAN_S to decrease with *m* more quickly than FAN_N. As a result, d_N/d_S might increase with *m* when *m* is large.

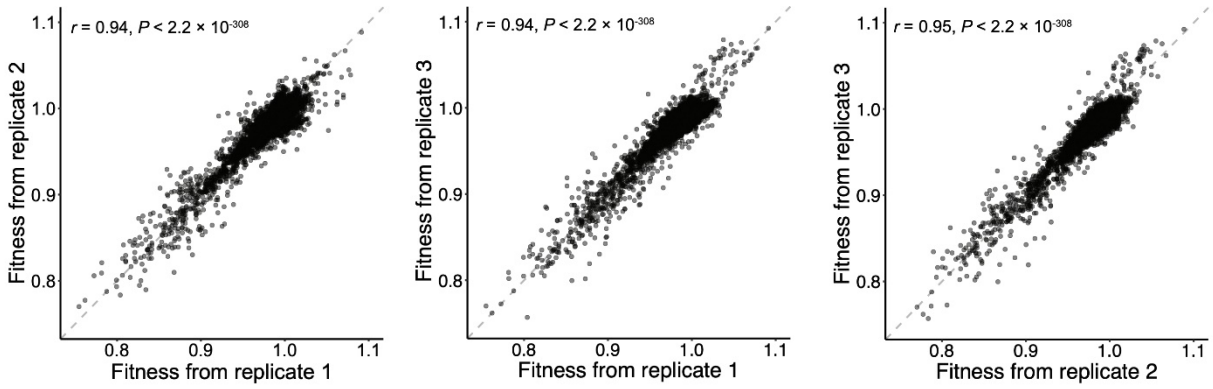

Fig. A-18. Correlation between every two of the three replicates in estimated mutant fitness under SC at 37°C. Each dot is a mutant and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented. Among-genotype sum of squares explains 96.1% of the total sum of squares (one-factor ANOVA).

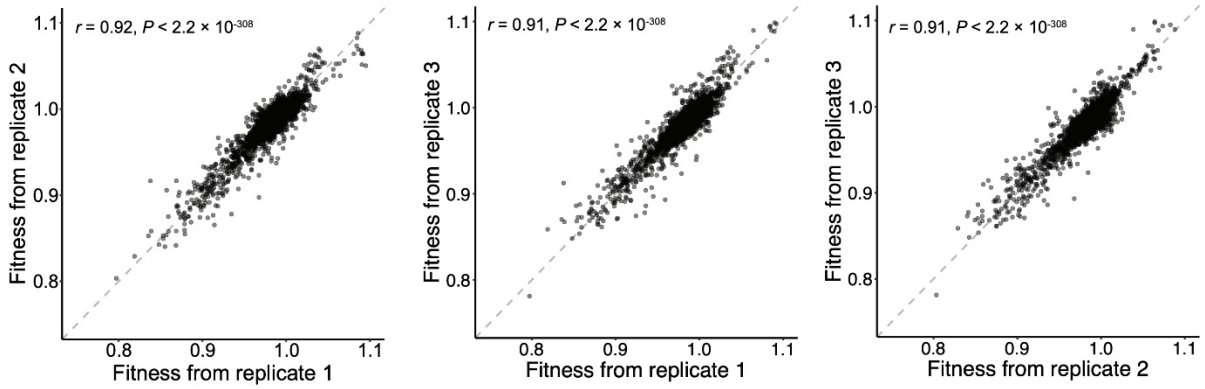

Fig. A-19. Correlation between every two of the three replicates in estimated mutant fitness under YPD + 0.375 mM H_2O_2. Each dot is a mutant and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented. Among-genotype sum of squares explains 94.4% of the total sum of squares (one-factor ANOVA).

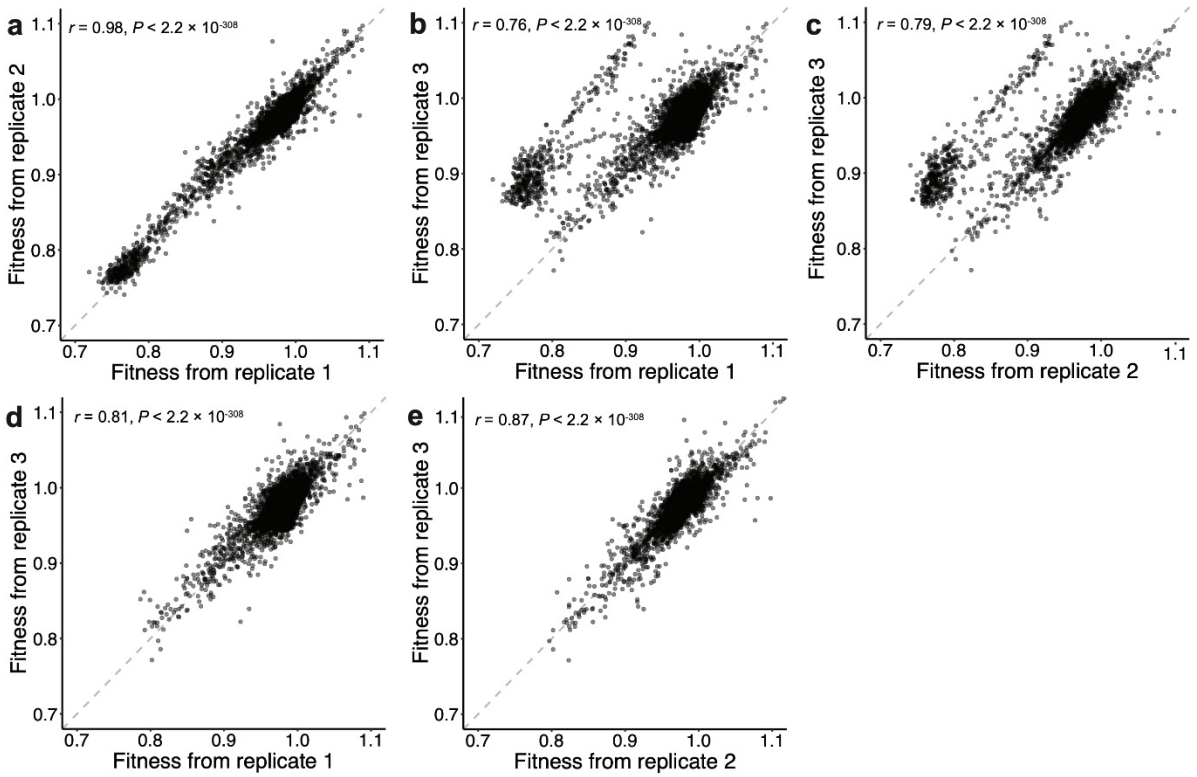

Fig. A-20. Correlation between every two of the three replicates in estimated mutant fitness under YPE. Each dot is a mutant and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented. **a**, Correlation between replicates 1 and 2. **b**, Correlation between replicates 1 and 3. **c**, Correlation between replicates 2 and 3. **d**, Correlation between replicates 1 and 3 after excluding *SNF6* mutants. **e**, Correlation between replicates 2 and 3 after excluding *SNF6* mutants. These data suggest that the fitness estimates of *SNF6* mutants in replicate 3 are unreliable, so are unused in fitness estimation. When *SNF6* is excluded, among-genotype sum of squares explains 91.0% of the total sum of squares (one-factor ANOVA).

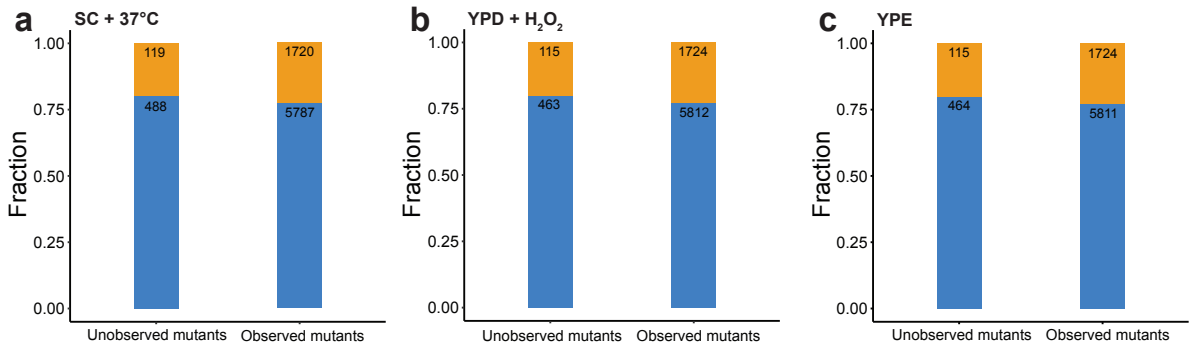

Fig. A-21. Fractions of synonymous (yellow) and nonsynonymous (blue) mutants among designed but unobserved mutants and those among observed mutants in each of the three additional environments tested. Nonsense mutants are not considered. Numbers in the bars are numbers of mutants. The distributions of synonymous and nonsynonymous mutants among the unobserved and observed mutant groups are not significantly different in each environment ($P >$ 0.05, Fisher's exact test).

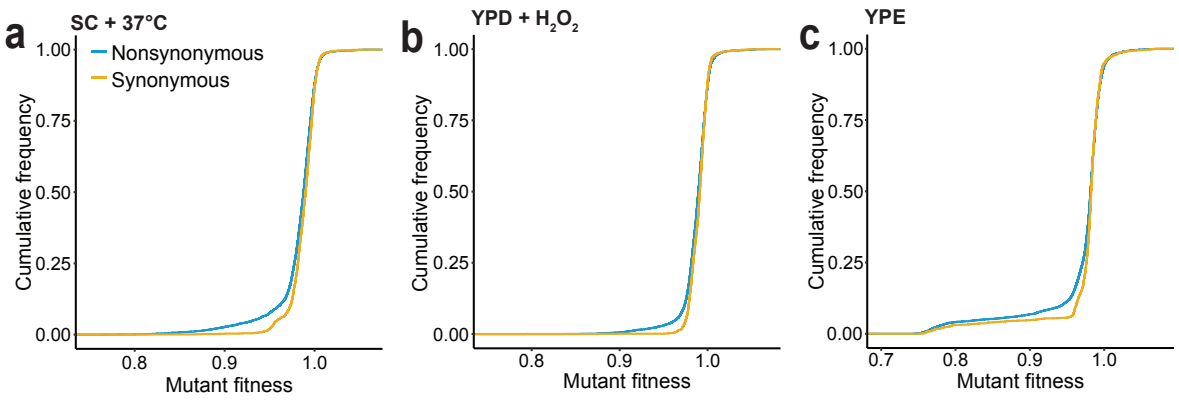

Fig. A-22. Cumulative frequency distributions of fitness of nonsynonymous and synonymous mutants in the three additional environments tested.

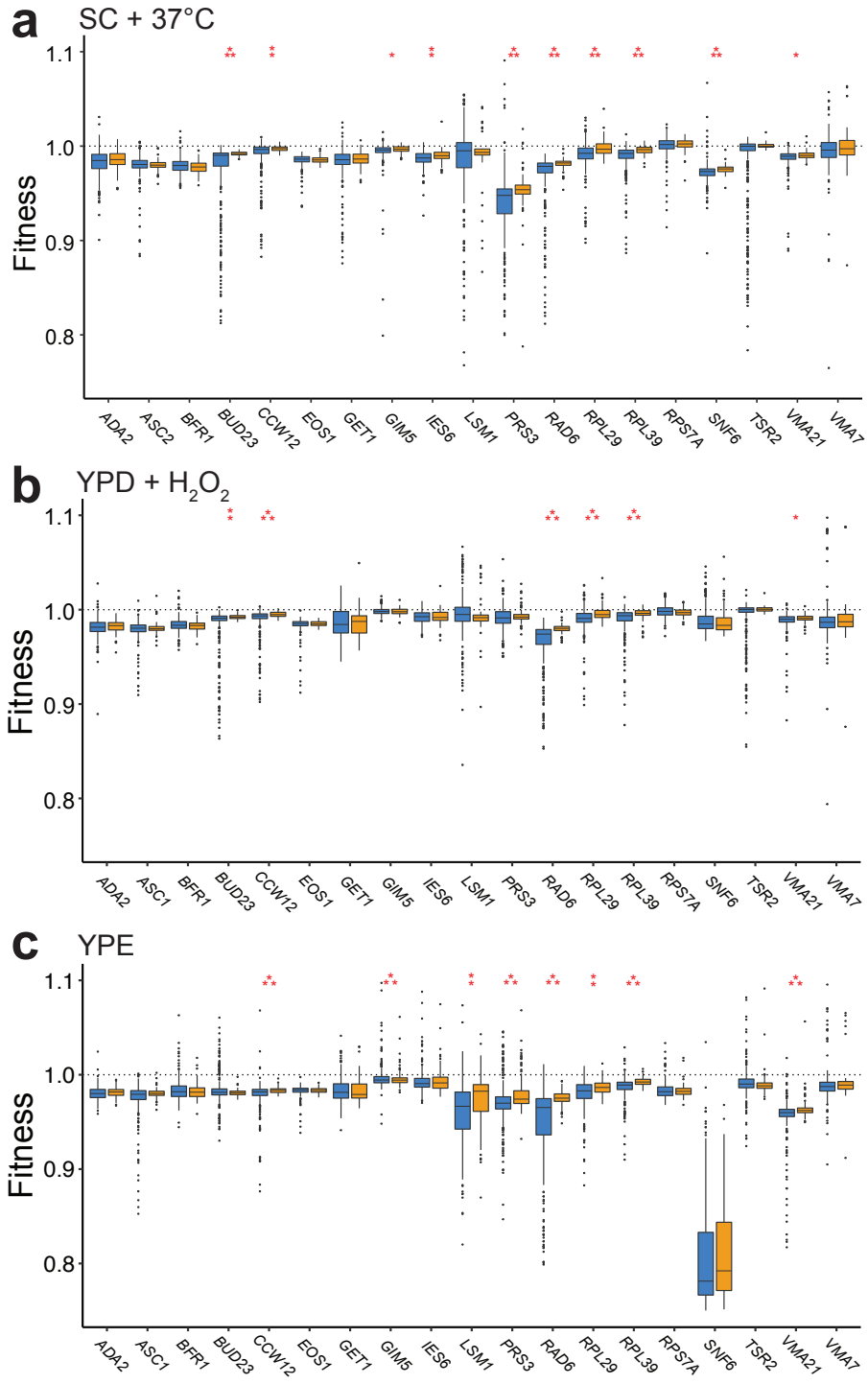

Fig. A-23. Fitness distributions of nonsynonymous (blue) and synonymous (yellow) mutants of 19 individual genes shown by box plots in each of the three additional environments tested. Nonsynonymous and synonymous distributions for each gene are compared by a

Wilcoxon sum-rank test, with the FDR-adjusted *P*-value indicated as follows: *, $P < 0.05$; ⁂, $P < 0.01$, ⁂, $P < 0.001$.

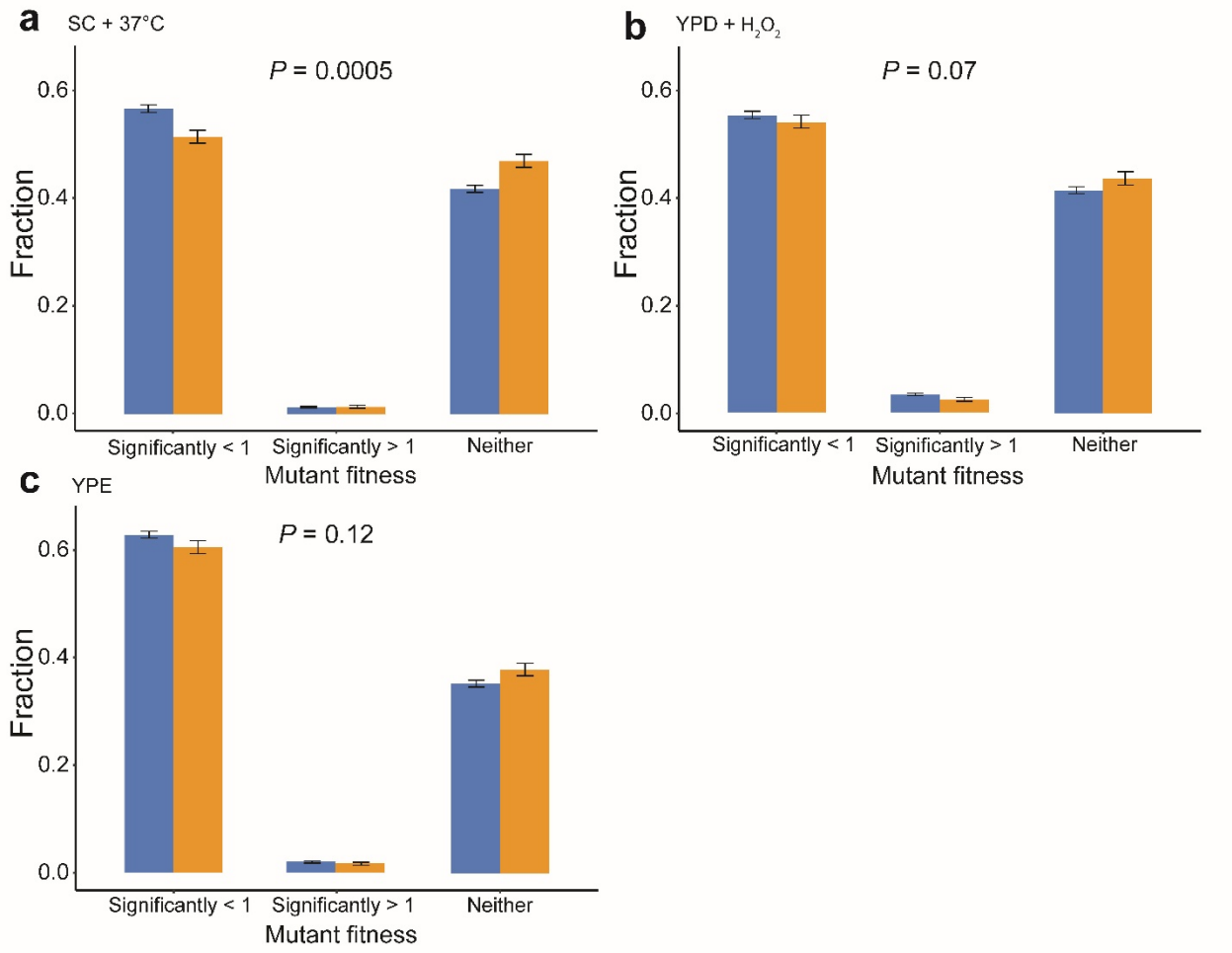

Fig. A-24. Fractions of mutants with fitness significantly below 1 (P <0.05), significantly above 1, and neither, respectively, in the three additional environments tested. The distributional difference between synonymous and nonsynonymous mutants among the three bins is tested by Fisher's exact test, with the *P*-value indicated. At FDR = 0.05, 40.7% and 0.7% of nonsynonymous mutations and 34.8% and 0.5% of synonymous mutations are significantly deleterious and beneficial, respectively, in SC+37°C. These values become 35.5%, 1.7%, 31.9% and 1.6% in YPD+H_2O_2, and 47.6%, 1.4%, 45.6%, and 1.0% in YPE.

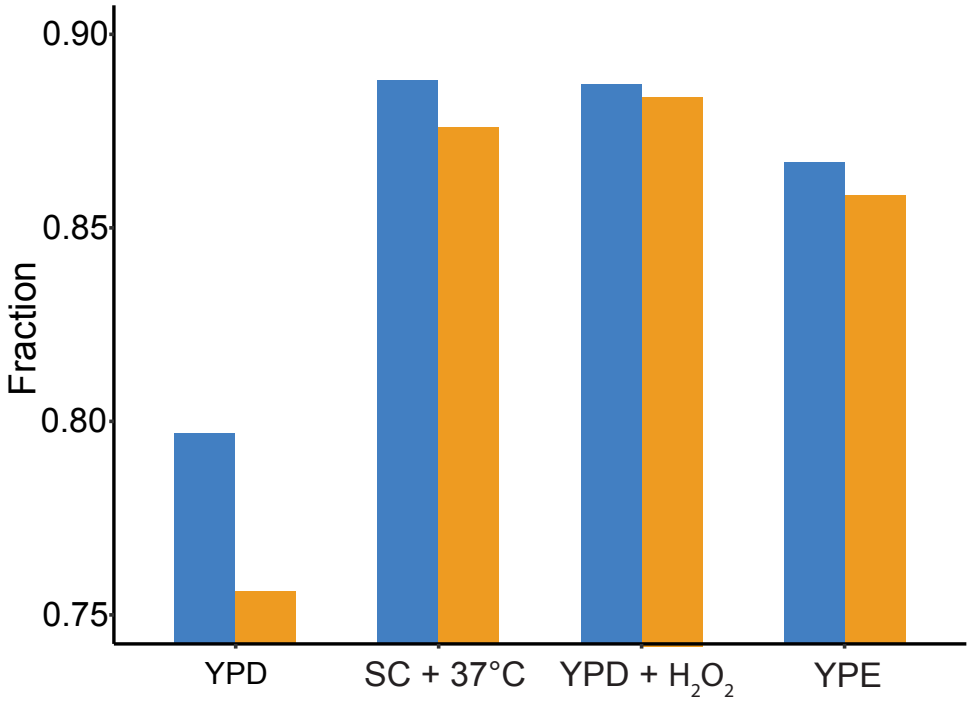

Fig. A-25. Fractions of nonsynonymous (blue) and synonymous (yellow) neutral mutations in one environment (indicated on the X-axis) that become deleterious in any of the other three environments. The fractions are higher for nonsynonymous than synonymous mutations (P <0.05, paired t-test). A mutation is considered deleterious if its fitness is significantly lower than 1 (P <0.05) and neutral if its fitness is not significantly different from 1.

Appendix B: Supplementary Tables and Figures for Chapter 3

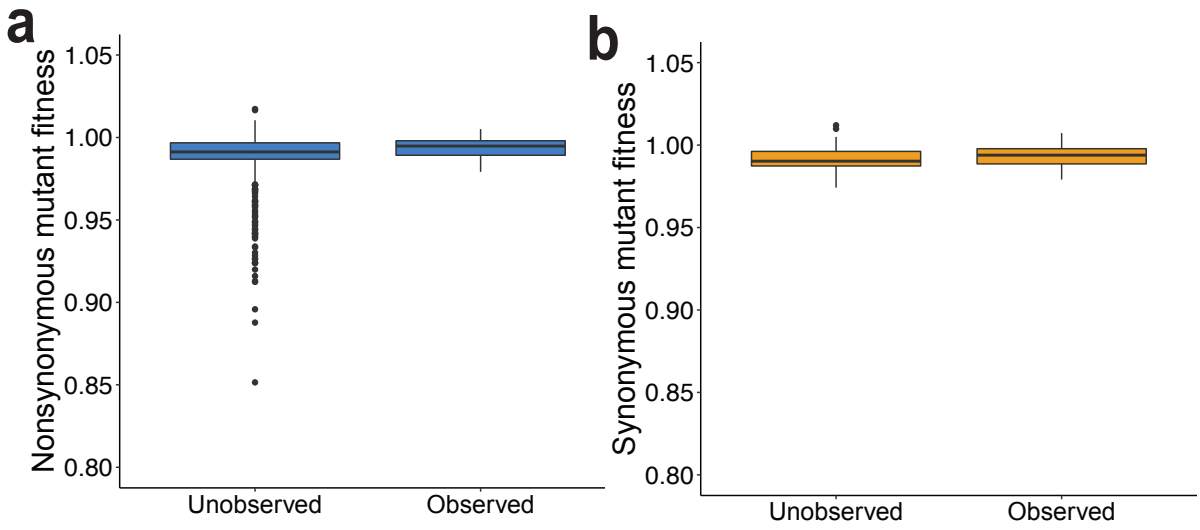

Fig. B-1. The full figure of Fig. 3-3d, including low-fitness and high-fitness mutants that are not shown in Fig. 3-3d.

Appendix C: Supplementary Tables and Figures for Chapter 4

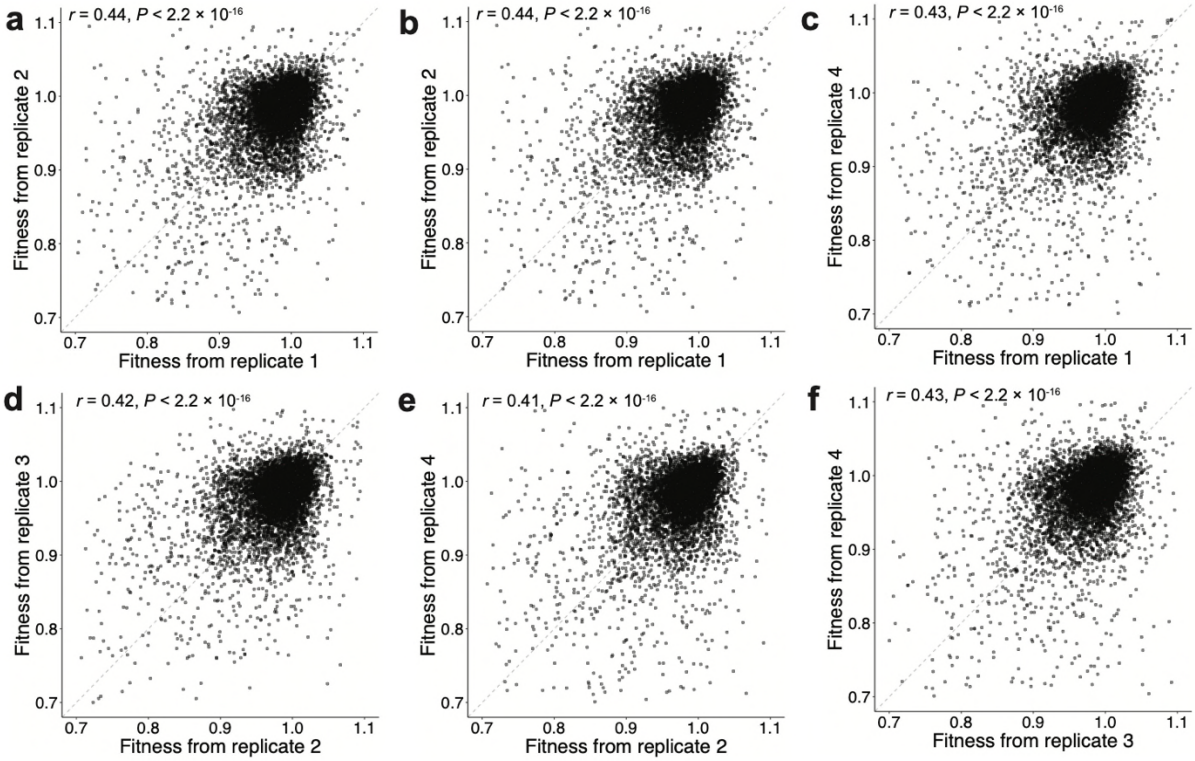

Fig. C-1.Correlation between every two of the four replicates in estimated double mutant fitness under YPD at 30°C. Each dot is a mutant and the dotted line indicates the diagonal. Pearson's correlation *r* and its associated *P*-value are presented.

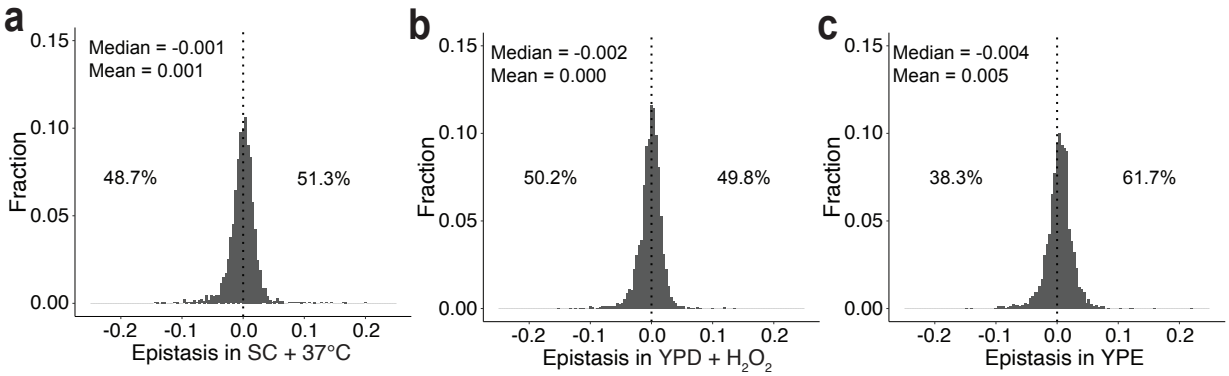

Fig. C-2. Distribution of epistasis in the three environments. a, SC + 37°C. **b,** YPD + 0.375 mM H_2O_2. **C,** YPE. The percentages are the fractions of negative or positive epistasis.

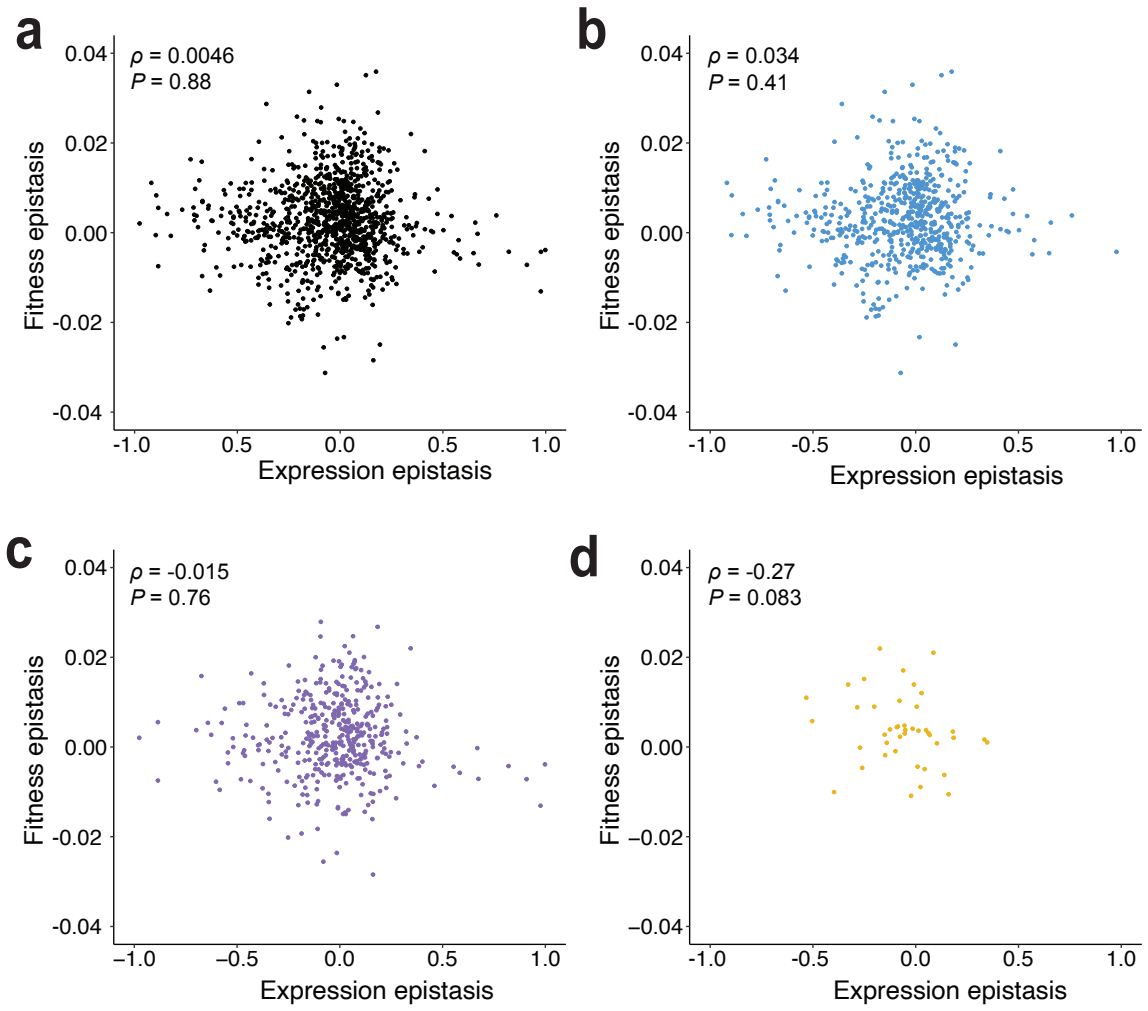

Fig. C-3. Correlations between expression epistasis and fitness epistasis in YPD. a, All double mutants. **b,** Double nonsynonymous mutants. **c,** Double nonsynonymous synonymous mutants. **d,** Double synonymous mutants