nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01578-x

Expression level is a major modifier of the fitness landscape of a protein coding gene

Zhuoxing Wu[1,6], Xiujuan Cai[2,6], Xin Zhang[1], Yao Liu[2], Guo-bao Tian[3,4,5], Jian-Rong Yang[1,4] ✉ and Xiaoshu Chen[2,4] ✉

The phenotypic consequence of a genetic mutation depends on many factors including the expression level of a gene. However, a comprehensive quantification of this expression effect is still lacking, as is a further general mechanistic understanding of the effect. Here, we measured the fitness effect of almost all (>97.5%) single-nucleotide mutations in *GFP*, an exogenous gene with no physiological function, and *URA3*, a conditionally essential gene. Both genes were driven by two promoters whose expression levels differed by around tenfold. The resulting fitness landscapes revealed that the fitness effects of at least 42% of all single-nucleotide mutations within the genes were expression dependent. Although only a small fraction of variation in fitness effects among different mutations can be explained by biophysical properties of the protein and messenger RNA of the gene, our analyses revealed that the avoidance of stochastic molecular errors generally underlies the expression dependency of mutational effects and suggested protein misfolding as the most important type of molecular error among those examined. Our results therefore directly explained the slower evolution of highly expressed genes and highlighted cytotoxicity due to stochastic molecular errors as a non-negligible component for understanding the phenotypic consequence of mutations.

One major task of biological research is to understand how genotypes and phenotypes correspond to each other, which is known as genotype–phenotype mapping (GPM)[1,2]. An incremental approach for GPM is to measure the phenotypic consequences of genetic mutations[3–7]. Mutational effects on phenotypes are generally believed to be dependent on many other genetic factors (that is, epistasis)[4,6–10]. For mutations within genes, the phenotypic effects are expected to be influenced by the expression level of the gene, as mutations are most likely to be neutral when they occur in a 'gene' with zero expression. Indeed, comparisons among different genes have provided indirect evidence of stronger deleterious effects of mutations occurring in highly expressed genes[11,12]. In addition, a more recent study[13] showed that the gene expression level can significantly alter the effects of mutations within a functional domain of a focal gene as well as how they interact. Nevertheless, a full understanding of the influence of the gene expression level on the effect of mutations should be based on the full coding sequence (CDS) instead of a limited domain.

Organismal fitness is arguably the single most important phenotype among all phenotypes, as it is the ultimate target of natural selection and therefore dictates biological evolution[14]. Genetic mutations leading to a significant decrease in fitness (that is, deleterious mutations) tend to have a general association with early-onset, severe genetic disorders[15–18] and will eventually be removed from the genetic pool of evolving populations by negative selection[14]. Given the prevalence of negative selection relative to positive selection[19] (which causes fitness-increasing beneficial mutations to spread and ultimately become fixed in the population), the exclusion of deleterious mutations will slow the evolution of the sequence of a gene. Comparisons among genes within the same genome have revealed a strong anticorrelation between the expression and the evolutionary rate of protein-coding genes (the ER-anticorrelation), which can be found in all three domains of life[12]. Assuming similar mutational inputs for different genes, the ER-anticorrelation suggests that mutations occurring in highly expressed genes are generally more deleterious.

Why are mutations in highly expressed genes more deleterious? Multiple molecular mechanistic models have been proposed to answer this question[12]. The currently prevailing models can be overarchingly summarized as avoidance of molecular stochastic errors, including mistranslation[11,20], protein misfolding[21] and misinteraction[22]. Specifically, these errors have been hypothesized to impose a greater burden on the cell when they occur in highly expressed genes than when they occur in genes with low expression because the former should give rise to greater amounts of erroneous protein molecules (in the case of mistranslation or misfolding) or complexes (in the case of misinteraction). Multiple studies have provided indirect support for these error avoidance hypotheses by comparing different genes within the same genome; for example, by evaluating the preference for less error-prone residues/codons in highly expressed genes or higher error rates in genes with low expression[11,21–23]. The existence of molecular stochastic errors has also been suggested to influence organismal fitness[24]. However, direct evidence supporting these hypotheses (that is, a greater fitness cost of mutations predicted to be deleterious by these error avoidance models in highly expressed genes relative to when the same genes are lowly expressed) is still lacking, as is knowledge of the relative contribution of these errors to the fitness effects of CDS mutations.

In this study, we used doped oligonucleotide synthesis to construct variant libraries of the green fluorescent protein (*GFP*) gene and the URAcil-requiring (*URA3*) gene from *Saccharomyces*

[1]Department of Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. [2]Department of Medical Genetics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. [3]Department of Microbiology, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. [4]Key Laboratory of Tropical Disease Control, Ministry of Education, Sun Yat-sen University, Guangzhou, China. [5]School of Medicine, Xizang Minzu University, Xianyang, China. [6]These authors contributed equally: Zhuoxing Wu, Xiujuan Cai. ✉e-mail: yangjianrong@mail.sysu.edu.cn; chenxshu3@mail.sysu.edu.cn

cerevisiae with full single-nucleotide mutation coverage across their whole CDSs as well as unique barcodes attached to each variant, which were determined by PacBio circular consensus sequencing[25] (PacBio-CCS). The fitness effects of individual mutations (that is, the local fitness landscape) were then evaluated by measuring the relative fitness of each variant in competitive cultures of single variant-containing yeast strains, followed by Illumina high-throughput sequencing of the barcodes. We compared the fitness landscape of the genes when they were individually driven by two promoters with an approximately tenfold activity difference and showed that such expression differences significantly altered the fitness effects of CDS mutations. Contrasting the fitness landscape of *GFP* (an exogenous gene) with that of *URA3* (a functionally required gene) enabled us to distinguish the mutational effects caused by increased stochastic molecular errors from those caused by decreased functional activity. Finally, although only a small fraction of variation in fitness effects among mutations in the same expression level can be explained by the examined hypotheses, we found protein misfolding as the most deleterious type of molecular error among those examined, thereby suggesting greater contribution of misfolding avoidance in constraining the evolution of highly expressed genes compared to the other hypothesized mechanisms. Our results therefore constitute direct experimental validation of the prevailing hypotheses underlying the ER-anticorrelation and, more importantly, highlight the role of molecular stochastic errors in the phenotypic consequences of genetic mutations.

Results

Whole-CDS measurement of the local fitness landscape. We aimed to measure the fitness effect (*s*elective coefficient, *s*) of all single-nucleotide mutations within the CDS of a protein-coding gene. Similar experiments, including those evaluating the functional landscapes of some non-fitness traits, such as fluorescence intensity, have been previously reported for short non-coding genes[6,7,26], short regions within the CDS or regions with low coverage of single-nucleotide variants[5,24]. There are two challenges that must be met to achieve our goal. The first is to obtain variant libraries with full coverage of all single-nucleotide mutations across the whole CDS (~1,000 base pairs (bp), 3,000 single-nucleotide mutations). The random introduction of mutations into the wild-type CDS, for example, via error-prone PCR[5] or totally degenerate mixtures[24], tends to result in multiple mutations and missed single-nucleotide mutations. On the other hand, doped oligonucleotide synthesis provides better control over the fraction of single-nucleotide mutations but is only applicable to regions shorter than 100 bp. To resolve this, we split the focal gene (*GFP* or *URA3*) into non-overlapping subregions of 50 bp and used doped oligonucleotide synthesis with a mutation rate of 3% (1% per non-wild-type nucleotide) per site to get mutation primers for each subregion, which was concatenated by multiple rounds of fusion PCR to obtained the mutant library within an expression cassette (Fig. 1a and Extended Data Fig. 1a–d). Within the expression cassette, the focal gene was also accompanied by the proper promoter (P_{TDH3} or P_{AGP1}), a terminator, a barcode for distinguishing genotypes (see the next paragraph) and a transformation marker (*LEU2*). The mutant libraries for all subregions were then pooled together to form a full mutant library (Fig. 1b) for each combination of promoter and focal gene.

The second challenge is the simultaneous determination of the relative fitness of all variants. We used the 20 random nucleotides added downstream of the terminators during fusion PCR as barcodes for each variant (Fig. 1a, Extended Data Fig. 1b; Methods). The full mutant library was then bulk-transformed into a BY4741-based acceptor strain and integrated into the *HO* locus (Fig. 1b and Extended Data Fig. 1e–g; Methods). The transformants were subsequently subjected to PacBio-CCS (ref. [25]) to determine the correspondence between the barcodes and genotypes (Fig. 1c and Supplementary Table 1). The relative fitness of variants with known barcodes could then be determined by the competitive coculture of the integrated strains (Fig. 1c), followed by variant frequency estimation via HiSeq sequencing (or NovaSeq, collectively referred to as HiSeq hereafter) of the barcodes (Fig. 1c, Supplementary Table 2 and Extended Data Fig. 1h). Following this experimental procedure, we constructed four variant libraries targeting two genes (*GFP* and *URA3*), each of which was driven by two different promoters (P_{TDH3} and P_{AGP1}) whose activities differed by about tenfold[27] (Extended Data Fig. 2a).

To verify the accuracy of our experimental pipeline, the expression cassettes of some randomly chosen colonies were amplified and subjected to Sanger sequencing (Fig. 1c). For example, for the P_{TDH3}-*GFP* library, we found that for all mutants sequenced by both Sanger sequencing and PacBio-CCS, identical sequences were obtained from the two methods (see Supplementary Table 1 for other libraries), indicating highly reliable correspondence between genotypes and barcodes. We captured at least one barcode for a total of 32,376 variants by PacBio-CCS, covering 99.3% of the single-nucleotide mutations and some multiple mutation variants (Fig. 1d and Supplementary Table 1). The frequencies of variants determined by HiSeq-based sequencing (Supplementary Table 2 and Extended Data Fig. 1h) were highly correlated between biological replicates of the same timepoint (Fig. 1e; average Pearson's $R=0.980$) and were less correlated between samples from different timepoints (Pearson's $R=0.947$; Fig. 1e and Extended Data Fig. 1h).

On the basis of the read counts, we estimated the relative fitness per generation by contrasting the read count changes of a genotype during the coculture with that of the wild type (Methods). Multiple measures were applied to ensure reliable fitness estimation. For example, we discarded wild-type barcodes with extreme day 7/day 0 frequency ratios (Extended Data Fig. 2b) to avoid estimation skewed by these outliers. We further required that the barcodes for each variant had at least 100 reads at the initial library, so that fitness could be measured at a reasonable resolution/accuracy. As a result, we obtained the fitness estimates of 29,795 variants in P_{TDH3}-*GFP* (see Supplementary Table 2 for other libraries) covering 99.2% of its single-nucleotide mutations (Fig. 2a and see Fig. 2b–h for the other competitive experiments). We further tested the accuracy of these fitness estimates via assessing the coefficient of variation among barcodes of the same variant (Extended Data Fig. 2c), the between-replicate correlations of the estimates (Extended Data Fig. 2d,e and Supplementary Table 3), consistency between estimates from day 3 and day 7 (Extended Data Fig. 2f), enrichment of deleterious mutations around the activity centre of the protein[28] (Extended Data Fig. 2g), distribution of fitness effect of synonymous versus non-synonymous mutations (Fig. 2i–p) or finer classifications of mutation types (Extended Data Fig. 2h), measuring double time individually for ten least-fit and ten fittest genotypes (Extended Data Fig. 2i). Collectively, these analyses offered general support for the reliability of the fitness estimates (more details in Methods).

The fitness effects of single-nucleotide mutations are expression dependent. When the fitness effect of a given mutation was compared between P_{TDH3}-*GFP* (s_H, fitness effect of the mutation when it occurs on a highly expressed gene) and P_{AGP1}-*GFP* (S_L, fitness effect of the mutation when it occurs on a lowly expressed gene), at least 42% of the mutations (Fig. 3a; Mann–Whitney *U*-test by the biological replicates) displayed expression-dependent fitness effects. Similarly, the 'expression effect' (the difference between S_H and S_L) was evident in the *URA3* libraries (Fig. 3b–d), suggesting that regardless of the functional importance of the gene, the fitness effects of mutations are usually expression dependent. We then asked whether the expression effect tended to be positive (high expression level of a specific mutation is more beneficial/less

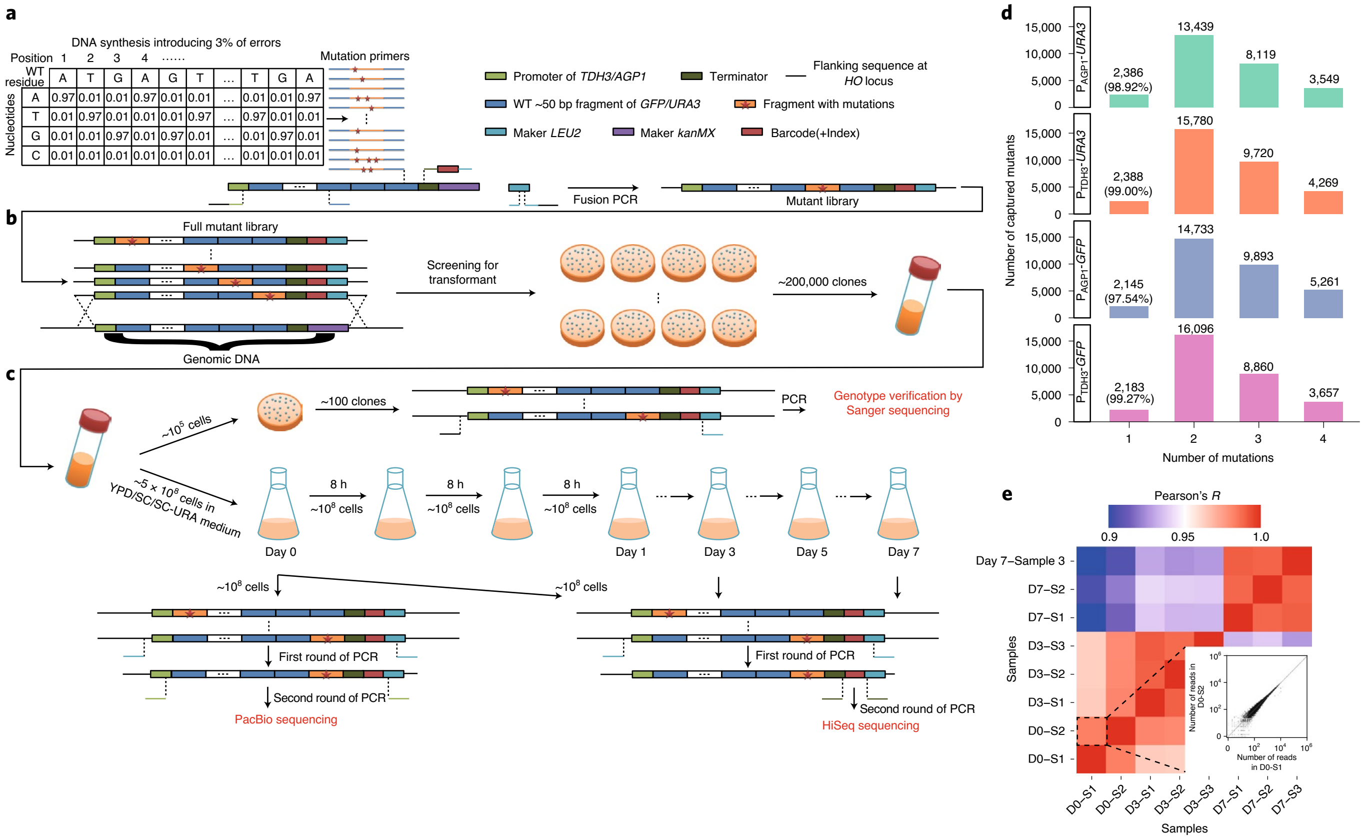


Fig. 1 | Overview of the experiment. a, Construction of the variant libraries by the chemical synthesis of 'mutation primers' and fusion PCR. The mutation primers were chemically synthesized 90-bp oligonucleotides containing a 20-bp invariant sequence, 50 bp of sequences with 3% chance of error and another 20-bp invariant sequence. Other sequence elements, including the promoter, invariant regions of the CDS, terminator, barcode and the *LEU2* marker, were concatenated with the mutation primers via two rounds of fusion PCR (Methods and Extended Data Fig. 1a,b). The overlapping parts of the primers are identically coloured. **b**, Variant libraries for different variable regions of the same CDS driven by the same promoter were combined and transformed into corresponding acceptor strains (Methods and Extended Data Fig. 1e,f), where the expression cassette was homologously recombined with the *HO* locus. Approximately 200,000 clones of the transformants were collected to produce a variant strain pool for each gene and each promoter. **c**, To verify the successful construction of the variant strain pool, ~100 clones from the pool were subjected to Sanger sequencing targeting the expression cassette at the *HO* locus. To determine the relative fitness of the variant strains, the pool was subjected to continuous competitive culture. The starting population of the competition experiment was sequenced by PacBio-CCS to assess the correspondence between the genotype and the barcode (Methods). Populations at different stages of the competition were subjected to barcode sequencing on the Illumina HiSeq platform so that the frequencies of the genotypes could be determined. **d**, Number of variants captured by PacBio-CCS, stratified by the number of mutated nucleotides within the CDS. **e**, For the P_{TDH3}-*GFP* library competing in YPD medium, the frequencies of mutants were highly replicable between biological replicates but were less replicable between samples from different timepoints. Note that all Pearson's *R* values were calculated without the wild type (WT) because it was always the dominant genotype, the inclusion of which would similarly increase all Pearson's *R* values. See also Extended Data Fig. 1.

deleterious than its low expression level) or negative (high expression level of a specific mutation is more deleterious/less beneficial than its low expression level). For *GFP*, elevated expression resulted in significant positive effects for 422 mutations and significant negative effects for 589 mutations (binomial $P<10^{-6}$, Fig. 3a). Similar observations recorded for *URA3* variants competing in yeast peptone dextrose (YPD) medium (Fig. 3b) suggested that when the physiological function of the gene was not essential to the survival of the cell, the expression effect tended to be negative. In contrast, for URA3 strains competing in SC medium, elevated expression tended to result in positive (474) instead of negative (391) effects on the mutations (binomial $P<0.003$, Fig. 3c). This trend became even stronger in synthetic complete media without uracil (SC-URA) (548 versus 267, binomial $P<10^{-22}$, Fig. 3d), indicating that this propensity for a positive expression effect was probably a consequence of the physiological function of the gene. In this case, although uracil was present in SC medium, it was likely in short supply, making the function of URA3 very important for the fast growth of the cells, although not essential.

How can we reconcile the aforementioned contrasting patterns observed for functional (*URA3*) and non-functional (*GFP*) genes? We sought to answer this question by disentangling the expression-dependent and expression-independent components of a mutation's fitness effect. We first confirmed that for most variants, 'mixed' linear models containing both said components tended to outperform an 'expression-dominant' models with no expression-independent components (Extended Data Fig. 3a,b). Then, by fitting the mixed model to each variant, we were able to approximate the expression-dependent and expression-independent components by the slope and intercept of the linear model, respectively (Fig. 3e and Supplementary Table 4). Note that because the traditional intercept (extrapolated y at $x=0$) does not allow sensible biological interpretation, we instead examined the intercept at $x=1.15\times10^4$, which is the average expression level among all coding genes[27]. In other words, this intercept can be interpreted as the expected fitness effect of a mutation if it occurs to an average gene (Fig. 3e). It turns out that this expected fitness

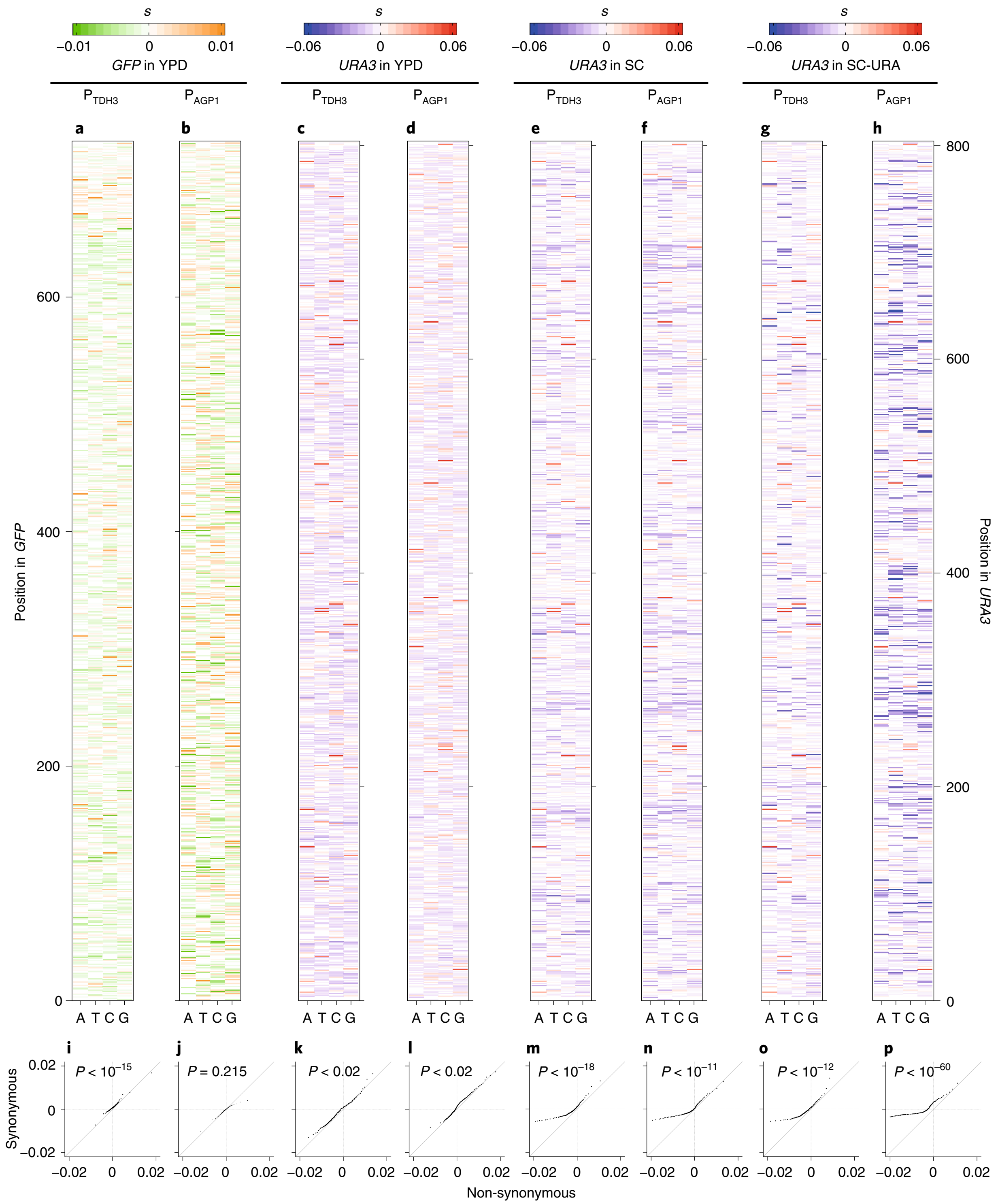


Fig. 2 | Overview of the measured fitness landscapes. a–h, The fitness landscapes were measured for *GFP* in YPD medium (**a** and **b**), *URA3* in YPD (**c** and **d**), SC (**e** and **f**) and SC-URA (**g** and **h**) media. The expression of each gene was driven by either P_{TDH3} (**a**, **c**, **e** and **g**) or P_{AGP1} (**b**, **d**, **f** and **h**). Each tile in the heatmap represents one variant (*x* axis) at a specific position (*y* axis) of the gene. The per-generation fitness effect (selective coefficient, *s*; Methods) is represented by the colour of the tile scaled by the corresponding colour scale bar on top. Variants whose fitness was unavailable are indicated as white. **i–p**, Quantile–quantile plot comparing the fitness effects of synonymous (*y* axis) and non-synonymous (*x* axis) mutations in the fitness landscapes measured for *GFP* in YPD medium (**i** and **j**), *URA3* in YPD (**k** and **l**), SC (**m** and **n**) and SC-URA (**o** and **p**) media, with the expression of each gene driven by either P_{TDH3} (**i**, **k**, **m** and **o**) or P_{AGP1} (**j**, **l**, **n** and **p**). *P* values from two-tailed Mann–Whitney *U*-tests are indicated.

effect of CDS mutations had an equal chance of being deleterious or beneficial (binomial $P=0.32$, Fig. 3f), which is consistent with the non-functional nature of *GFP*. More importantly, the coefficient of the expression-dependent component displayed a significant tendency to be negative (binomial $P<10^{-8}$, Fig. 3g), which suggested that some kind of cellular toxicity (see the next

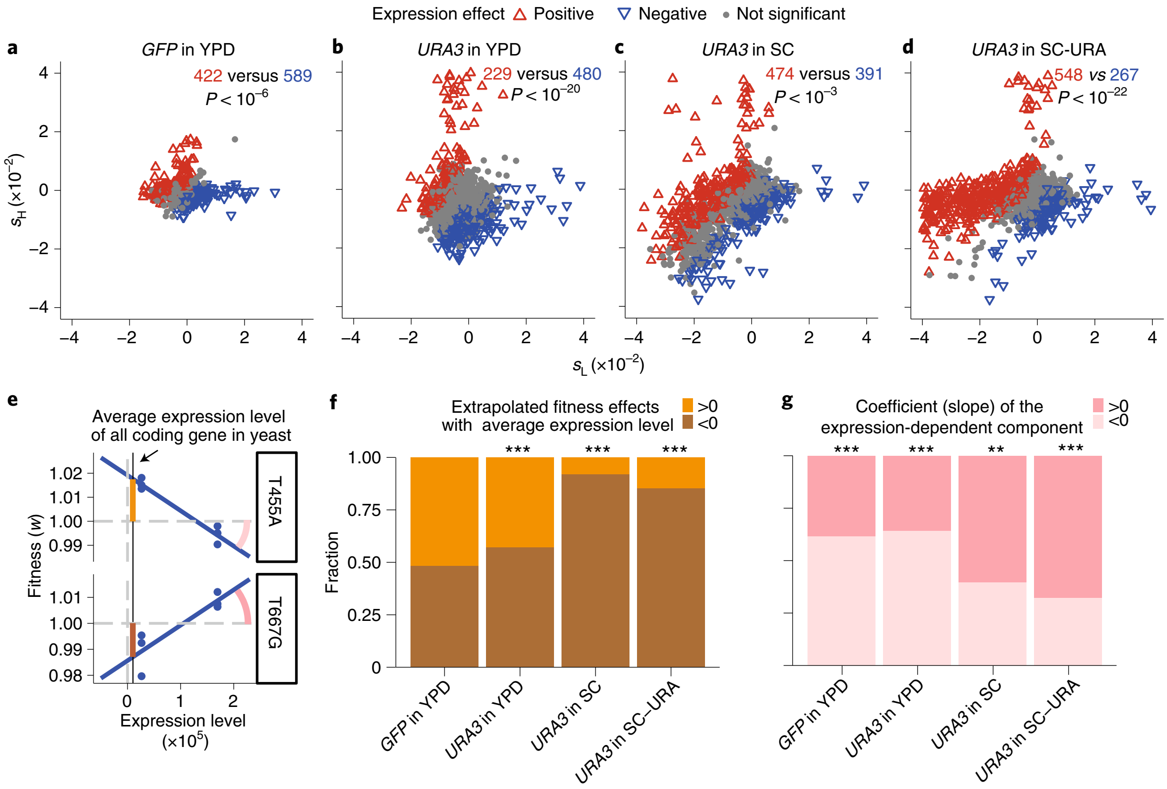


Fig. 3 | The fitness effects of CDS mutations are expression dependent. a–d, For each CDS mutation, the fitness effect in the high-expression context (S_H, *y* axis) was compared with that in the low expression context (S_L, *x* axis) to show the expression effect. Mutations with significant expression effects ($P < 0.1$ in two-tailed Mann–Whitney *U*-test, using the biological replicates of S_H and S_L) are represented with coloured triangles; grey dots are used otherwise. The number of mutations with positive/negative expression effects is indicated in correspondingly coloured fonts. Each panel shows results for one gene in one growth medium, as indicated on top of the panels. Binomial tests against a null expectation of equal chance for positive and negative expression effect were used to calculate the *P* values indicated on each panel. **e**, Two mutations fitted by linear models using the six estimated fitnesses are presented as examples. The vertical thin black line indicates the proteome-wide average expression level of $x = 1.15 \times 10^4$, at which the intercepts (the orange or brown segments) of the fitted linear models represent the expected fitness effect of the mutation if it occurs to an average gene. The slopes (the pink arches to the right) of the fitted linear models represent the direction (positive or negative) of the fitness effect. **f**, For each mutation giving rise to a linear model with a significant intercept component, its expected fitness effect on an average gene was classified as deleterious (brown, exemplified by the brown segment in **e**) or beneficial (orange, exemplified by the orange segment in **e**). The fractions of deleterious/beneficial mutations are shown. **g**, For each mutation giving rise to a linear model with a significant coefficient of the expression-dependent component, the coefficient/slope was classified as positive (pink, exemplified by the pink curve in **e**) or negative (light pink, exemplified by the light pink curve in **e**). For both **f** and **g**, binomial *P* values calculated against the null expectation of equal probabilities for positive or negative values are indicated as $^{**}P < 10^{-3}$ and $^{***}P < 10^{-5}$. See also Extended Data Fig. 3.

section for the molecular nature of such toxicity) triggered by the mutation was increased due to expression elevation. In contrast, mixed models fitted to the *URA3* mutants competing in SC or SC-URA medium tended to give rise to deleterious effects for a gene with an average expression level (Fig. 3f) and positive coefficients for the expression-dependent component (Fig. 3g). These patterns are not unexpected, as *URA3* with low to average expression should be more sensitive to the mutation-triggered functional degeneration than highly expressed *URA3* because the activity of P_{AGP1} was lower than needed (Extended Data Fig. 3c) and less responsive (relative to P_{URA3}) to the uracil-shortage environment (Extended Data Fig. 3d). This pattern is also understandable by the diminishing return of increasing *URA3* expression, in which a mutation-triggered loss of functional molecules will give rise to greater functional reduction when it occurred on high expression level than on low expression level (Extended Data Fig. 3e)

Collectively, the above results suggested two distinct scenarios of the expression dependency of fitness effects of mutations. For a gene with sufficient expression for its function (*GFP* in YPD or *URA3* in YPD), elevated expression tended to increase cellular toxicity due to CDS mutations. On the other hand, for a gene that was underexpressed for its function (*URA3* in SC or SC-URA), elevated expression tended to increase cellular tolerance to functional degeneration caused by CDS mutations. Because the latter scenario heavily depend on the gene function and the difference between the native and optimal (in terms of fitness) expression of the gene[29], and the broadly observed ER-anticorrelation[12] seems to suggest the former scenario as the norm (below), we will focus on the cellular toxicity component hereafter.

Molecular stochastic errors are more deleterious in highly expressed genes. As we observed above, the cellular toxicity caused by CDS mutations became more severe as the expression of the gene increased. Consequently, the same mutation would be more likely to be purged by purifying selection when occurring in a highly expressed gene, which effectively means that the sequences of highly expressed genes are more constrained and therefore evolve more slowly relative to those of the genes with low expression. Indeed, the expression level of a gene has been found to be the most important determinant of its evolutionary rate (the ER-anticorrelation)[12]. Various molecular-level mechanistic models have been proposed to explain why mutations in highly expressed genes are more likely to be deleterious, among which the prevailing models can be overarchingly summarized as avoidance of molecular stochastic errors ('error avoidance' hypothesis). Specifically, mutations that increase the probability of protein mistranslation[23]/misfolding[21]/misinteraction[22] would result in a greater number of erroneous protein molecules/complexes when occurring in highly expressed genes than when occurring in genes with low expression and would therefore impose a more severe burden on the cell[12]. In addition, the secondary structure of messenger RNA has been suggested to act as a suppressor of protein mistranslation[20] and misfolding[30]; therefore, the error avoidance hypothesis also suggests a stronger requirement for mRNA folding among highly expressed genes[31]. Although indirect support for the error avoidance hypothesis has been obtained via comparisons among different genes[21–23] and fitness effects without manipulation of the expression level[24], a direct experimental test of error avoidance is still lacking. The fitness landscape associated with P_{TDH3} and P_{AGP1} offered a unique opportunity to explicitly test the biological relevance of error avoidance to the ER-anticorrelation as well as individual hypotheses regarding different types of stochastic molecular errors.

We aimed to assess the aforementioned hypotheses under a unified scheme. Briefly, the probability of molecular error for each mutant was approximated with commonly used measures (below). Then, the fitness effect (s) of each mutation was compared with the corresponding error probability (Fig. 4a, b). The avoidance of molecular errors would be supported if (1) the correlation between the error probability and the fitness effect of mutations was more negative for the highly expressed gene than the lowly expressed gene (Fig. 4a) and (2) the fraction of deleterious mutations in highly expressed genes relative to that in genes with low expression (that is, the relative strength of the functional constraint) increased with the error probability (Fig. 4b). We followed this scheme to individually test the hypotheses of mistranslation avoidance, misfolding avoidance, misinteraction avoidance and the mRNA folding requirement, as described below. Note that the biological repeats of each variant have been pooled together to obtain a single fitness estimate hereafter (see Methods regarding pooled or separated biological repeats) and we will be presenting in the main text the results based on *GFP* mutant strains growing in YPD, while Extended Data Fig. 4 presents the other results.

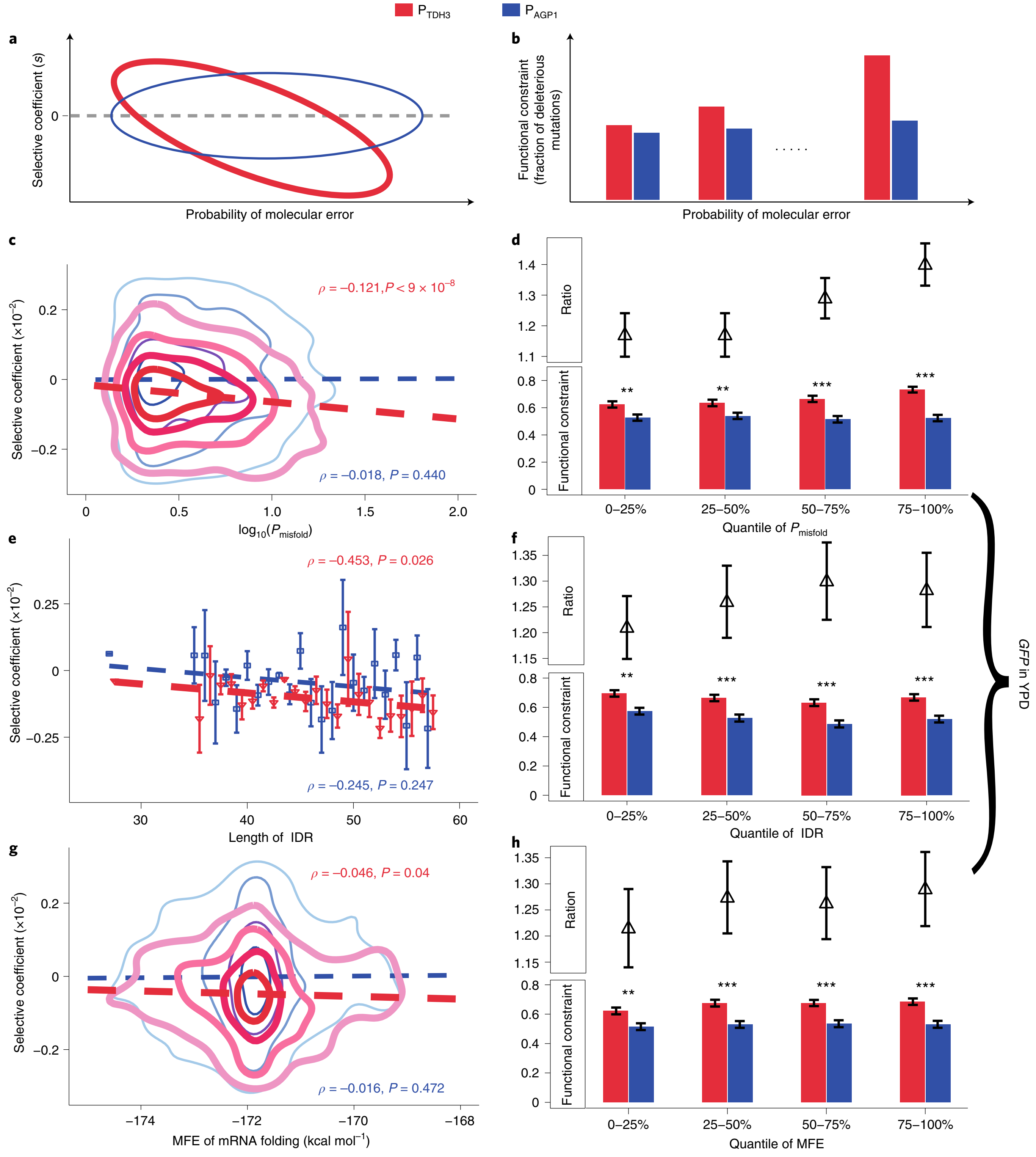
In the case of mistranslation, unpreferred codons are generally considered more prone to translational error because their cognate transfer RNA is in short supply or binds to the codon with lower affinity[23]. We therefore approximated the relative translational error rate with the decrease in the tRNA adaptation index (tAI)[32] due to synonymous mutations but we did not obtain support from the designed tests (Extended Data Fig. 4a,b; Methods). For misfolding, we estimated the relative probability of protein misfolding (P_{misfold}) compared to that for the wild-type sequence based on the increase in protein folding energy ($\Delta\Delta G$) predicted by I-Mutant[33] (Methods) and obtained unanimous support from both tests described above (Fig. 4c,d, corresponding to Fig. 4a,b, respectively). For misinteraction and mRNA secondary structure, we similarly obtained support using the size of intrinsically disordered regions (IDRs), an indicator of the propensity for promiscuous protein–protein interaction[34] (Fig. 4e,f, corresponding to Fig. 4a,b, respectively. See also Extended Data Fig. 4c,d, where we used protein surface hydrophobicity but found no signal) and the minimum free energy (MFE) of the mRNA secondary structure (Fig. 4g,h, corresponding to Fig. 4a,b, respectively). In addition, when the *URA3* fitness landscape measured in YPD was used, most of the analyses described above yielded supportive results (Extended Data Fig. 4e-j). We also conducted the above analyses for the *URA3* fitness landscapes measured in SC and SC-URA media and found that the patterns generally supported the accuracy of the chosen proxies for molecular errors and the increased tolerance of mutation-triggered functional degeneration (more molecular stochastic errors lead to greater functional degeneration) when expression was elevated (Extended Data Fig. 4k–p). Collectively, these results constitute direct experimental validation of the error avoidance hypothesis (specifically for misfolding/misinteraction avoidance and the mRNA folding requirement but not for mistranslation avoidance) and therefore offer an important mechanistic explanation for the expression dependency of the fitness effects of CDS mutations.

Relative importance of misfolding, misinteraction and the mRNA folding requirement. We have demonstrated that protein misfolding/misinteraction and the mRNA folding requirement underlie the expression dependency of mutational effects but their relative importance remains unresolved. We approached this question by fitting a linear model for the P_{TDH3}-associated fitness landscapes using the normalized P_{misfold}, IDR and MFE values of all mutations in the gene (*GFP* or *URA3*). The regressed model gave rise to the most significant coefficient for P_{misfold} but less significant coefficients for misinteraction and mRNA folding (Fig. 5a). These results suggested that, among the three examined hypotheses, protein misfolding is the most deleterious factor for highly expressed genes and therefore contributed most to the constrained evolution of highly expressed genes. Nevertheless, we note that in this linear model, only a small fraction (<6%) of the variation in mutational fitness effect can be explained by the physiological properties captured by the three examined hypotheses (Supplementary Table 5).

To further explore the functional constraint due to misfolding/misinteraction avoidance and the mRNA folding requirement, we conducted additional detailed analyses of functional constraint and mutational sensitivity. Here, functional constraint is the extent to which random mutations, due to their deleterious effects, are purged by natural selection. And mutational sensitivity refers to the

Fig. 4 | Stochastic molecular errors can explain the stronger deleterious effect of mutations in highly expressed genes. a,**b**, Schematic diagrams of the two tests for the expression-dependent cytotoxicity of molecular errors. **a**, The correlation between the probability of molecular error (*x* axis) and selective coefficient (*y* axis) should be more negative when the gene is highly expressed (driven by P_{TDH3}, coloured red) compared to when it is lowly expressed (driven by P_{AGP1}, coloured blue). **b**, Mutations with higher probabilities of molecular errors (*x* axis) should be under greater functional constraint (*y* axis) as estimated by the fraction of deleterious mutations. **c**,**d**, Assessing the misfolding avoidance hypothesis by the two tests using P_{misfold} (Methods). **c**, Distribution of all single-nucleotide mutations in *GFP* driven by P_{TDH3} or P_{AGP} was indicated by contour lines, with darker lines representing higher density. The linear regression using the raw data points is presented as a dashed line and the Spearman's rank correlation coefficient ρ and corresponding P values are also indicated. **d**, The lower half shows the estimated functional constraints for mutations within different ranges of P_{misfold} for P_{TDH3}-driven *GFP* (red) and P_{AGP1}-driven *GFP* (blue), with standard errors assessed by bootstrapping the mutation 1,000 times. The P_{TDH3} versus P_{AGP1} difference within each group was shown as ratio in the upper half panel and tested by Mann–Whitney U-test (*$P < 0.05$, **$P < 10^{-3}$ or ***$P < 10^{-5}$). **e**,**f**, Assessing the misinteraction avoidance hypothesis by the two tests in using the length of the IDR (Methods). These panels are, respectively, similar to **c** and **d**, except that we used the mean and standard error of the fitness effects for each unique *x* value, which are too discrete for contour plot. The Spearman's rank correlation coefficient ρ and the corresponding P values obtained using the mean fitness effects are indicated. **g**,**h**, Assessing the mRNA folding requirement hypothesis by the two tests using MFE (Methods). These two panels are, respectively, similar to **c** and **d**. These panels shown result from *GFP* mutants growing in YPD. Results from other conditions are in Extended Data Fig. 4.

probability that a random mutation will increase the propensity of a certain type of stochastic molecular error. Taking misfolding as an example, the misfolding avoidance model predicts that CDS regions that are mutationally more sensitive to protein misfolding should be more functionally constrained. We therefore estimated a score of mutational sensitivity to protein misfolding for each codon within the CDS on the basis of the number of single-nucleotide mutations in the focal codon that give rise to a higher probability of misfolding (Fig. 5b, the second row on the right). Additionally, we calculated a functional constraint score for each codon, which was defined as the fraction of deleterious single-nucleotide mutations in the focal codon, because the greater the number of deleterious mutations found in a codon, the more likely it is to be functionally constrained (Fig. 5b, the first row on the right). We then examined the segment of *GFP* with a certain level of functional constraint (Fig. 5b, grey bars in the left panel) and checked their average mutational sensitivity to protein misfolding (Fig. 5b, red lines in the left panel). Similar calculations were performed for misinteraction avoidance and



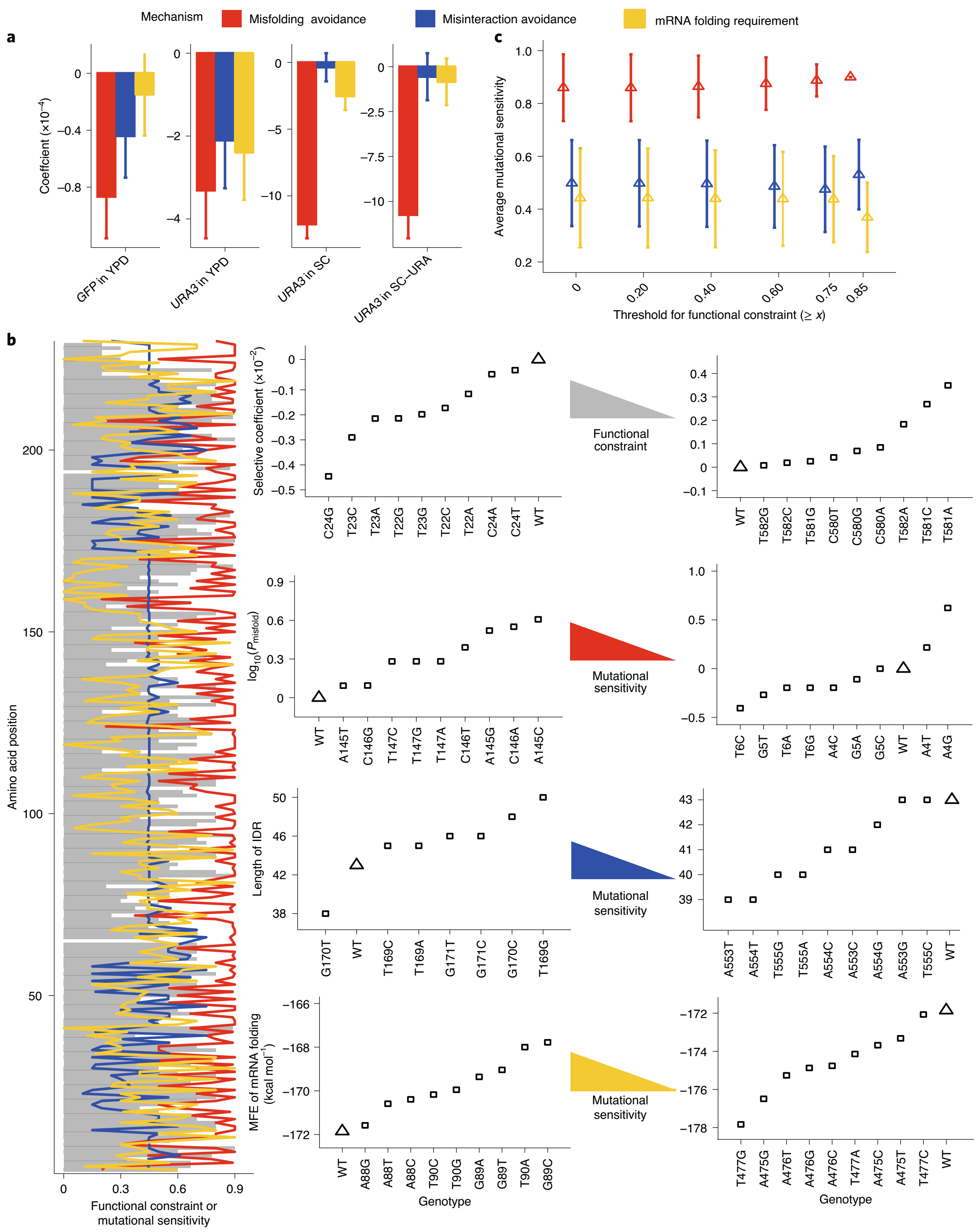


Fig. 5 | Relative contribution of different types of molecular errors. a, A single linear model was regressed against the fitness effects of all single-nucleotide mutations in P_{TDH3}-*GFP* using normalized (*Z*-transformation) $P_{misfold}$ (red), IDR (blue) and MFE (yellow) as the explanatory variables. The estimated coefficient (height of the bar) as well as its standard error (error bar) for each explanatory variable are shown for all measured fitness landscapes (*x* axis). **b**, For each amino acid in GFP, the strength of the functional constraint (grey bar in the left panel) was compared to the mutational sensitivity (coloured lines in the left panel) according to misfolding avoidance (red), misinteraction avoidance (blue) or the mRNA folding requirement (yellow). For each type of value shown in the left panel, one pair of examples showing high and low values is listed on the right, where the *x* axis indicates the genotype (mutants represented with squares and the wild type with triangles) and the *y* axis indicates either the fitness effect or the proxy of molecular errors. **c**, The average mutational sensitivity (*y* axis) for misfolding avoidance (red), misinteraction avoidance (blue) or the mRNA folding requirement (yellow) was calculated for all amino acids within GFP with the strength of the functional constraint exceeding a certain threshold (*x* axis). Error bar represents the standard error.

the mRNA folding requirement (blue and yellow lines in Fig. 5b, respectively). Consistent with the above observations, these analyses again highlighted the importance of protein misfolding, as it always displayed higher mutational sensitivity than the other two models (Fig. 5c). Furthermore, this result was recapitulated using P_{TDH3}-*URA3* libraries from YPD, SC and SC-URA media (Extended Data Fig. 5). In conclusion, although at least a few types of molecular stochastic errors contribute to the expression dependency of the fitness effects of CDS mutations, protein misfolding is probably the single most important one among the examined errors.

Discussion

Using an effective experimental pipeline, we measured the local fitness landscape of a protein-coding gene driven by two different promoters whose activities differed by about tenfold. Comparison between the high-expression and low-expression fitness landscapes suggested the expression dependency of the fitness effects of single-nucleotide mutations. We obtained direct support for several models of the molecular mechanisms underlying such expression dependency, which can be overarchingly summarized as avoidance of stochastic molecular errors. The relative contributions of different types of stochastic errors to the deleterious effects of CDS mutations were also examined, revealing a dominant contribution of the avoidance of protein misfolding. In summary, we have explained with high resolution how gene expression affects the fitness effect of individual CDS mutations.

Investigations of the evolutionary rate of protein-coding genes have resulted in the molecular clock hypothesis[35] and the neutral theory[36], which are both cornerstones in the fields of molecular evolution and comparative genomics. In the genomic era, gene expression has been revealed as a major determinant of the evolutionary rate of protein[12], yet the underlying molecular mechanisms remain elusive, as multiple related hypotheses have been proposed but only indirect support for these hypotheses has been derived from the comparison of different genes[21–23,31]. Our study offers direct evidence that the functional constraint imposed by the stochastic error caused by a given mutation in a given gene are stronger when the gene is highly expressed than when it is lowly expressed. Nevertheless, we want to stress that, according to an ANOVA analysis based on the linear models fitted for the fitness landscapes of P_{TDH3}-driven genes (Fig. 5a), all three hypotheses examined can only explain a small fraction (<6%) of variation in the fitness effects of the measured mutants (Supplementary Table 5). In other words, even though the three hypotheses found support from our experimental data, the vast majority of the variations in mutational fitness effects remained unexplained, either because of the poor measurement/prediction accuracies of molecular errors or existence of additional unknown biological factors.

Besides the ER-anticorrelation, the effect of gene expression level on the organismal fitness has recently been investigated in other contexts. Most notably, ~100 yeast genes were expressed at ~100 distinct expression levels and assayed for the resulting fitness effects in yeast[29]. As a result, it was revealed that the native expression level is sometimes non-optimal in terms of fitness. When the native expression is below the optimal level (as shown for *URA3* libraries in SC and SC-URA media), the protein sequence evolution might favour functional enhancement rather than avoidance of molecular error. Moreover, how gene expressions shift and switch genetic interactions among mutations has also been studied[13] for a functional domain of a protein. Although the functional importance of the mutated gene had not been taken into consideration, the observed expression dependency for genetic interactions was largely compatible with a protein folding-based mechanistic model, which is consistent with the dominating role of protein misfolding in stochastic molecular errors.

In addition to the theoretical value for understanding molecular evolution, the role of molecular stochastic errors in determining the fitness effect of mutations has important biomedical implications. Common sense dictates that functional constraint, which once violated should cause symptoms, should be related to the physiological function of a gene. However, our results suggest that functional constraint could also be a result of cytotoxicity caused by stochastic molecular errors. Such mutation-induced cytotoxicity is clearly distinct from gain-of-function mutations, as it does not rely on the existence of any specific cellular components but the product of the gene itself, whereas gain-of-function mutations usually involve specific interaction partners. Our results therefore strongly support the suggestion that the cytotoxicity caused by stochastic molecular errors is a non-negligible factor in identifying/explaining disease-related mutations.

Methods

Construction of transformation acceptor strains. To measure the fitness effect of all single-nucleotide mutations of a protein-coding gene in a massively parallel manner, we had to collect a large number of transformants with mutated genes. To improve the efficiency of homologous recombination-based yeast transformation, we constructed four receptor strains. First, we amplified two expression cassettes, P_{TDH3}-*GFP-KanMX* and P_{AGP1}-*GFP-KanMX*, from *GFP* strains obtained from a previous study[37] (Supplementary Table 6). These two cassettes were then integrated into the *HO* locus of strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) of *S. cerevisiae*[38] via a standard transformation protocol[39] to construct two *GFP* acceptor strains (Extended Data Fig. 1e). Specifically, the cells of strain BY4741 were cultured at 30 °C in 5 ml of YPD (1% yeast extract, 2% peptone, 2% glucose) overnight until saturation. Then, the cells were diluted to an optical density OD_{660} of 0.2 and grown for ~4 h until the OD_{660} reached 0.7. Thereafter, the culture was harvested by centrifugation at 2,000*g* for 5 min and competent cells were obtained using 0.1 M lithium acetate (LiAc, Sigma). Subsequently, 240 µl of polyethylene glycol (50% w/v, Sigma), 30 µl of LiAc (1 M), 30 µl of water, 10 µl of salmon sperm vector DNA (10 mg ml^{-1}, Sigma), 5 µg of the DNA product of each cassette and 50 µl of competent cells were added to a microcentrifuge tube, briefly vortexed and subjected to heat shock at 42 °C for 30 min. After washing once in water, the mixture was spread on YPD plates with kanamycin (Sigma) and cultured for 2–3 d. Transformants of individual colonies were selected and confirmed by PCR for the correct replacement of the *HO* locus. Specifically, a pair of primers targeting the flanking sequence of the *HO* locus was used to amplify the expression cassette from transformants. Then, the amplification product was subjected to Sanger sequencing using another pair of primers, one of which matched the sequence of the promoter (P_{TDH3} or P_{AGP1}), while the other matched the sequence of the terminator, such that correct recombination could be verified (Supplementary Table 6).

The two *GFP* acceptor strains constructed as described above were then used in another round of transformation in which the CDS of *GFP* was replaced with the CDS of *URA3* from the pGSKU plasmid[40], thereby creating two *URA3* acceptor strains (Extended Data Fig. 1f and Supplementary Table 6). Yeast transformation followed the same protocol described above except that SC-URA was used for screening the transformants.

Construction of variant libraries. We aimed to acquire variant libraries for protein-coding genes (*GFP* or *URA3*, driven by either the P_{AGP1} or P_{TDH3} promoter) with 100% coverage for all possible single-nucleotide mutations. For this purpose, we first divided the full CDS of the gene into non-overlapping regions of 50 bp in length. For each 50-bp region, we ordered chemically synthesized oligonucleotides corresponding to the 90-bp sequence centred on the focal 50-bp region from IDT (https://www.idtdna.com/). The oligonucleotides consisted of (in exact order) 20 invariant sites, 50 variable sites and 20 invariant sites. The synthesis of each variable site was performed by mixing 97% wild-type (the acceptor strain constructed above) nucleotides with 1% of each of the other three nucleotides (Fig. 1a). According to theoretical and previous empirical data[6], it was expected that ~25% of the synthesized oligonucleotides were wild type, while 25% contained single-nucleotide mutations and the others contained multiple mutations. Thus, CDS mutations were introduced during the synthesis of variable regions and the invariable regions could be used as conventional PCR primers. A total of 15 and 16 such mutant primers were synthesized to cover the CDSs of *GFP* (733 bp) and *URA3* (804 bp), respectively (Extended Data Fig. 1a and Supplementary Table 6).

The synthesized mutant primer targeting each 50-bp region was individually concatenated with other flanking elements through two rounds of standard PCR and two rounds of fusion PCR (Fig. 1a, Supplementary Fig. 1b and Supplementary Table 6) performed as follows. First, PCR fragment 1 was amplified using a mutant primer and a terminator primer targeting the 3′ end of the *GFP* terminator (Extended Data Fig. 1b,c and Supplementary Table 6). Second, PCR fragment 2 was amplified using a promoter primer targeting the 5′ end of the promoter and a fragment primer targeting the immediate upstream region of the focal 50-bp variable region (Extended Data Fig. 1b,d and Supplementary Table 6). Thereafter, the first round of fusion PCR was conducted using the promoter primer and a terminator primer linked with a 20-bp region of fully degenerated (randomized) nucleotides serving as a barcode (below) and a 20-bp region identical to the 5′ end of the marker gene *LEU2*. As a result, PCR fragment 1 and PCR fragment 2 were fused into fusion PCR fragment 1. We then constructed a *LEU2* marker with the *LEU2* CDS (from yeast strain S288c; ref. [41]) and the *KanMX* promoter and terminator, such that the sequence similarity with the corresponding region in acceptor strains was higher and the rate of transformation was therefore also higher. Finally, this *LEU2* marker was fused with fusion PCR fragment 1 using the promoter primer and another primer targeting the 3′ end of *LEU2*, giving rise to fusion PCR fragment 2 (Supplementary Table 6). Collectively, the fusion PCR fragment 2 contained (in the following order, 5′ to 3′), the promoter of choice (P_{TDH3} or P_{AGP1}), the wild-type CDS fragments upstream of the focal 50-bp region, the 50-bp mutated region, wild-type CDS fragments downstream of the focal 50-bp region, terminator, a 20-bp region of fully degenerated (randomized) nucleotides serving as a barcode (see also the next paragraph) and the *LEU2* marker gene. Depending on the number of mutant primers, the above process was repeated for multiple rounds. That is, for each promoter, 15 different types of fusion PCR segment 2 were constructed for *GFP* and 16 different types of fusion PCR segment 2 were constructed for *URA3*. All fusion PCR fragment 2 sequences of each gene driven by each promoter were combined to form a library (that is, a total of four libraries) that covered all single-nucleotide mutations of the gene.

Notably, the 20-bp barcode (Supplementary Table 6) incorporated into fusion PCR fragment 1 can be used as a genotype-specific barcode to facilitate massive genotype frequency estimation via high-throughput sequencing on Illumina platforms, whose sequencing length is too short to obtain the whole CDS. There are $4^{20} \approx 10^{12}$ possible sequence combinations for a 20-bp barcode, which is 50× the number of template DNA molecules used for PCR. To monitor possible template switching during PCR[42,43], we included an additional 5-bp index to mark different fusion PCR fragments (Supplementary Table 6). To minimize the occurrence of template switching, we used high-fidelity DNA polymerase (AccuPrime Pfx, Invitrogen) according to the user manual and minimized the number of fusion PCR cycles (the first round of fusion PCR consisted of 20 cycles and the second round of fusion PCR consisted of 28 cycles). In addition, the *LEU2* primer and the promoter primers contained the genomic upstream and downstream flanking sequences of *HO* in the acceptor strain so that fusion fragment 2 could be used for transformation (Extended Data Fig. 1b and Supplementary Table 6).

Another issue worthy of discussion here is the choice of promoter. We used only two promoters in our experiment, among which P_{TDH3} is one of the strongest promoters in the *S. cerevisiae* genome, whereas the activity of P_{AGP1} is approximately tenfold weaker[27] (Extended Data Fig. 2a). This was a compromise between using strong promoters to ensure the penetrance of the mutational effect and achieving a large enough activity difference between the promoters to reveal expression dependency. The deleterious effect of stochastic molecular errors will most likely be undetectable if an even weaker promoter is used unless methods with higher sensitivity to fitness differences are applied. It might therefore be desirable to measure more fitness landscapes under a wider range of expression levels when more sensitive methods for massively parallel fitness measurements become available.

Construction of the variant strain pool. Using the yeast transformation procedure described above, the four variant libraries were integrated into the *HO* locus of the corresponding acceptor strain with the corresponding promoter and target gene to build four strain pools (Fig. 1b). Transformants were selected on leucine-free synthetic complete plates (SC-LEU) and cultured at 30 °C for 2.5 d. To obtain an average of >20 clones for each mutation type at each site of each gene with a certain promoter, we collected at least 200,000 clones from the selection medium for each strain pool. Because each homology arm in the acceptor strains was longer than 400 bp, the rate of homologous recombination increased to $\sim 10^{-4}$ (Extended Data Fig. 1g). We started the transformation with $\sim 10^{10}$ cells, which made this step labour intensive. After washing the collected clones with YPD medium, 20-ml aliquots were stored in 30% sterile glycerol at −80 °C.

Competitive growth assay. We conducted competition experiments to estimate the fitness of different variants relative to that of the strain containing wild-type *GFP* or *URA3* (Fig. 1c). To ensure the presence of all variants (especially those with single-nucleotide mutations) in the competition experiment, $\sim 5 \times 10^8$ cells were used for downstream experiments, corresponding to an average of 2,500 cells per clone. Specifically, frozen samples of the variant strain pool were removed from −80 °C storage and placed at room temperature. For the *GFP* strains, samples were inoculated into 700 ml of YPD. For the *URA3* strains, we inoculated samples into 400 ml of SC and 400 ml of SC-URA media. The YPD and SC media were chosen to ensure fast growth and therefore a relatively short competition time. The SC-URA medium was chosen so that the physiological function of *URA3* (orotidine 5-phosphate decarboxylase) would become essential for the strain, thereby allowing the assessment of the fitness effect of mutations dependent on the functional importance of the gene. Three biological replicates were assayed in parallel for all libraries, except that the P_{AGP1}-*URA3* libraries growing the SC and SC-URA media were only replicated twice. The strains were then cultivated at 250 r.p.m. and 30 °C.

We used OD_{660} readings to estimate the concentration of cultured cells every 8 h to monitor cellular growth so that the population remained in the exponential growth phase, thus ensuring continuous growth competition among different mutants. At the same time, we transferred $\sim 1.5 \times 10^8$ cells (>15× the effective population size of yeast) to fresh medium[44], thereby minimizing the potential influence of genetic drift[45]. In addition, we sampled the population in competition every 24 h by freezing aliquots in 30% glycerol at −80 °C. The competitive culture experiment lasted 7 d. Because the genome size of yeast is 1.2×10^7 bp, the single-nucleotide mutation rate is $\sim 1.67 \times 10^{-10}$ per base per generation[46] and the growth rate in YPD medium is ~90 min per generation. Therefore, during the 7-d cultivation period, the number of point mutations was ~0.2 per genome. Yeast growth in SC/SC-URA medium (~135 min per generation) is slower than that in YPD; therefore, the number of point mutations should not exceed that in YPD. Thus, artifacts due to secondary mutations on other loci of the yeast genome should be negligible.

High-throughput sequencing. To determine the correspondence between the genotypes and barcodes in the variant strain pool, we used the PacBio Sequel platform (Nextomics Biosciences) to sequence the expression cassettes (Fig. 1c). First, phenol-chloroform and glass beads were used to extract genomic DNA[47] from aliquots of cells after melting. Specifically, $\sim 1 \times 10^8$ cells (~500 cells per variant clone) were centrifuged at 4,000 r.p.m. in a 1.5-ml tube for 2 min and then rinsed three times with 1 ml of sterile deionized water. We then added 100 µl of acid-washed glass beads (0.5 mm, Sigma), 200 µl of cell lysis buffer (2% Trion X-100, 1% SDS, 0.1 M NaCl, 1% 100× Tris-EDTA buffer solution, Sigma) and 200 µl of DNA extraction solution (WEST GENE) to the tube, followed by 15 min of vortexing. After centrifugation at 13,500 r.p.m. for 5 min, 200 µl of the supernatant was transferred to a new 1.5-ml tube. Then, 1 ml of cold 100% ethanol was added, after which the contents of the tube were mixed thoroughly and centrifuged at 13,500 r.p.m. for 2 min before the supernatant was discarded. The precipitate was treated with RNase cocktail enzyme (Invitrogen) at 37 °C for 1 h and 1 ml of cold 100% ethanol was added, followed by mixing by inversion and centrifugation at 13,500 r.p.m. for 2 min before the supernatant was discarded. After drying, the precipitate was added to 100 µl of sterile deionized water and stored at −20 °C and 1 µg of genomic DNA was used for the following PCR, which corresponding to ~300 copies of genomic DNA per variants. To minimize the introduction of base substitution errors and template switching during PCR, we used high-fidelity DNA polymerase and minimized the number of PCR thermal cycles. The number of cycles in the first round of PCR was 16 and the number of cycles in the second round of PCR was 13. The purified PCR product from each variant strain pool was sequenced in three PacBio Sequel or Sequel II SMRT Cells. To evaluate the accuracy of PacBio Sequel sequencing, we also performed Sanger sequencing on >100 randomly selected clones (Fig. 1c) and the results were compared with the PacBio-CCS results (Supplementary Table 1).

We also sequenced the barcode region using the Illumina HiSeq platform to estimate the change in genotype frequency before and after competition (Fig. 1c). The DNA extraction and PCR procedures were the same as sample preparation procedures for the PacBio Sequel platform, except that after the region between *HO* and the *LEU2* upstream flanking region was PCR amplified, only the barcode region was amplified by PCR (Fig. 1c and Supplementary Table 6). For each variant strain pool, two technical repeat experiments were performed on the sample from day 0, and one experiment was performed on each of the three biological replicates from day 3 and day 7. We ultimately used the data from day 7 in the downstream analyses because a longer cultivation time gave rise to a better resolution for fitness.

Mapping barcodes to variant genotypes based on PacBio Sequel results. We obtained raw CCS reads from the PacBio Sequel platform (Supplementary Table 1). Because the single-stranded DNA molecules of different alleles in the same library were highly similar, they could form heteroduplexes and cause base detection errors, thereby increasing the misreading of mutations. To avoid this problem, we used BLASR (ref. [48]) to map all the raw CCS reads to the wild-type sequence of the gene and divided them into positive-strand and negative-strand reads. Then, we separately called the consensus sequences (HiFi reads) from the positive strand and the negative strand. To further reduce the error rate in base detection, we also required that each HiFi read be generated from at least five raw CCS reads (number of passes ≥5). Finally, we extracted the genotype and the barcode from each HiFi read without any insertion or deletion. Only barcodes supported by at least two HiFi reads were considered. If one barcode was paired with multiple genotypes, we used the genotype that was supported by the majority (>50%) of the reads of the barcode; if no genotype reached majority, the barcode was discarded. Finally, we obtained 669,526, 632,997, 547,952 and 552,530 barcode–genotype pairs for the P_{TDH3}-*GFP*, P_{AGP1}-*GFP*, P_{TDH3}-*URA3* and P_{AGP1}-*URA3* libraries, respectively, which contained 99.27, 97.54, 99.00 and 98.92% of all single-nucleotide mutations in the corresponding gene (Supplementary Table 1).

We would like to further emphasize that accurate fitness estimation for all CDS mutations was critically dependent on a reliable correspondence between the genotype and barcode. We strove to reduce related artifacts by using relatively long barcodes and PCR reagents/conditions with high fidelity, minimizing the number of PCR cycles (to avoid template switching) and performing sequencing via the highly accurate PacBio-CCS approach. As a result, the erroneous correspondence between genotype and barcode appeared negligible according to the Sanger-sequenced mutants (Supplementary Table 1). In the future application of similar experimental procedures, case-by-case optimized conditions for such correct genotype-to-barcode mapping should be carefully designed.

Estimation of the relative fitness of genotypes by HiSeq sequencing. We sequenced the barcode region via 150-bp paired-end sequencing on the Illumina HiSeq platform (Supplementary Table 6). We required that the barcodes sequenced from both ends be totally consistent with each other without any mismatches. The extracted barcodes were then mapped to their corresponding genotypes. Because the number of reads for wild-type barcodes affects the fitness estimation of all variants and its accuracy is thus of particular importance, we calculated the changes in frequency for each wild-type barcodes that have at least 100 reads at day 0 and at least five reads at day 7 (that is, ratio between frequency at day 7 and frequency at day 0 or day 7/day 0 ratio). Then the barcodes whose day 7/day 0 ratios deviated more than one standard error from the average day 7/day 0 ratios were discarded. In estimating the selective coefficient, the number of reads for all barcodes mapped to the same genotype were combined. We also required that the total number of reads for a variant be at least 100 at day 0 to ensure accuracy/resolution of fitness estimation. The Wrightian fitness (marked as w) of a specific genotype relative to the wild type could then be calculated as follows:

$$w = \left(\frac{f_t/f_0}{F_t/F_0} \right)^{\frac{1}{g}} \tag{1}$$

Here, f_t is the frequency (number of reads) of the genotype in a culture environment after a time period of t in the competitive culture; f_0 is the frequency of genotypes on day 0 of (that is, before) competitive culture; F_t and F_0 are the f_t and f_0 of the wild-type CDS, respectively; g is the number of generations of cell growth during this time period (t) estimated from the OD_{660} obtained at each passage, measured every 8 h during competition, according to $g = \log_2\left(\mathrm{OD}_{660} \times 10^7 \times v/n\right)$, where v is the culture volume in litres (0.7 for YPD and 0.4 for SC/SC-URA) and n is the number of cells transferred at the last passage. Finally, all genotypes with fitness values <0.5 were discarded. For experiments in YPD or SC media, the exclusion of these genotypes should not substantially alter our conclusion because they constitute only <5% of all single-nucleotide mutations. As for *URA3* mutants growing in SC-URA, there are <300 genotypes with fitness <0.5, which is presumably caused by lack of functional *URA3* protein and therefore should not alter our implication about stochastic molecular errors.

We conducted two to three biological replicates for each competitive growth assay. Read/genotype frequencies at time t from these replicates can be used separately for fitness estimates via equation (1), giving rise to two to three replicate-separated fitness estimates for a single mutant that can be used to gauge the measurement errors (for example, Fig. 3). When one single fitness estimate was required for each mutant (for example, Figs. 2, 4 and 5), we followed a previous study[6] to pool all the reads from all replicates and used the resulting read counts as F_t and f_t. With minimal variation in g in equation (1) (number of generations, of which the difference between the largest and smallest never exceeds 1 among biological replicates of each competitive growth assay), the mathematical relationship between the pooled and replicate-separated fitness estimates can be shown by

$$w^g = \frac{f_t/f_0}{F_t/F_0} = \sum_{i=1}^{r} \left(\frac{F_{t_i}}{F_t} \times \frac{f_{t_i}/f_0}{F_{t_i}/F_0} \right) = \sum_{i=1}^{r} \left(\frac{F_{t_i}}{F_t} \times w_i^g \right) \tag{2}$$

Here, r is the number of replicates and i is the index for individual replicates. In other words, the pooled fitness estimate is a weighted average of the replicate-separated fitness estimates. The weight of replicate-separated fitness (F_{t_i}/F_t) is essentially the confidence level of the fitness estimate because an experiment in which the frequency of wild type is lower should have larger error in fitness estimation and therefore should be given lower weight when averaging.

We then sought to assess the accuracy of our fitness estimates. The fitness estimates generally have low coefficient of variation among replicates of the same variant that is comparable to previously reported fitness landscape[6] (Extended Data Fig. 2c). The between-replicate correlations of mutational fitness effects were high (Pearson's $R > 0.6$) for the majority (Extended Data Fig. 2d,e) but low for some other samples (Supplementary Table 3), which was suspected to be caused by large measurement errors of mutations with small fitness effect. Indeed, the between-replicate correlations of fitness were much stronger when only the mutations with absolute fitness effect >1% were considered (Supplementary Table 3). The fitness estimates derived for day 3 were highly consistent with that for day 7 (Extended Data Fig. 2f), the latter of which will be used across this study. We mapped the fitness landscape of $\mathrm{P}_{\mathrm{TDH3}}$-*URA3* measured in the SC-URA medium onto the three-dimensional structure of the orotidine 5-phosphate decarboxylase (the protein encoded by *URA3*) and found a region enriched with deleterious mutations and corresponding to the activity centre of the protein[28] (Extended Data Fig. 2g). We also compared the distribution of fitness effects between non-synonymous and synonymous mutations (Fig. 2i–p) and found that the difference between the two distributions was stronger when the mutated gene was functionally required (Fig. 2m–p, *URA3* in media with slight or complete shortage of uracil), relative to when it was not (Fig. 2i–l, *GFP* or *URA3* in YPD medium, a rich medium in which uracil is not a limiting factor for yeast growth). Furthermore, the average fitness effects of single synonymous mutations were closest to 0, whereas that of single missense mutations, multiple missense mutations and nonsense mutations were, in this exact order, increasingly more deleterious (Extended Data Fig. 2h). In addition, the growth curve of ten least-fit and ten fittest genotypes from the $\mathrm{P}_{\mathrm{TDH3}}$-*GFP* library growing in YPD individually measured by a spectrophotometer suggested that all but two genotypes displayed a doubling time consistent with their fitness estimates (Extended Data Fig. 2i). Collectively, these results offered general support for the reliability of the fitness estimates.

We would also like to note here that our experimental protocol combining Illumina sequencing and PacBio-CCS presents several advantages in determining the fitness landscapes of the whole CDSs of protein-coding genes. First, PacBio-CCS is not only more accurate than Illumina sequencing but can also be applied to much longer sequences and is therefore more reliable for determining the variant sequences of protein-coding genes, whose length dictates that each variant should be rare in the whole library. Second, the targeted sequencing of the barcode instead of the CDS region avoided the weakness regarding sequencing length on the Illumina platform and made use of its strength of a high throughput in terms of the number of reads. Thus, our method took advantage of the strengths of both techniques and represents an improvement over the state-of-the-art method for the deep mutational scanning of a whole protein-coding gene[5].

RT–qPCR for promoter activities. We used quantitative PCR with reverse transcription (RT–qPCR) to assess the activities of $\mathrm{P}_{\mathrm{AGP1}}$ and $\mathrm{P}_{\mathrm{URA3}}$ in YPD, SC and SC-URA media. Two clones of S288c respectively cultured in the three media were harvested during log-phase growth and RNA was extracted with RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Then, 1 μg of total RNA was reverse transcribed to complementary DNA using the PrimeScript RT reagent Kit with gDNA Eraser (TAKARA) according to the manufacturer's instructions. Gene expression was assayed in triplicates by RT–qPCR, amplified using iTaq Universal SYBR Green Supermix (Bio-Rad) on a LightCycler 96 Real-Time PCR System (Roche). RT–qPCR primers (Supplementary Table 4) were designed by NCBI Primer BLAST. The cycling parameters for amplification were 95 °C for 30 s and 40 cycles of 95 °C for 5 s and 60 °C for 30 s.

Protein expression level by flow cytometry. The protein expression level driven by the $\mathrm{P}_{\mathrm{TDH3}}$ or $\mathrm{P}_{\mathrm{AGP1}}$ was estimated by flow cytometry of the respective *GFP* acceptor strain. We cultured cells (three biological replicates each strain) in YPD medium at 30 °C and 250 r.p.m. overnight. The saturated culture was then diluted to $\mathrm{OD}_{660} = 0.2$ in 4 ml of YPD and growth continued at 30 °C for 4 h to reach exponential growth phase. Each promoter has three biological replicates and three technological repeats (3 × 3). Equal numbers of cells (10,000 cells) were collected by Attune N×T flow cytometer (Life Technologies) with a 533/30 nm optical filter from each sample to measure the abundance of GFP protein.

Doubling time measurement for individual genotypes. The $\mathrm{P}_{\mathrm{TDH3}}$-*GFP* single mutants with top ten and bottom ten measured fitness in YPD medium were selected and individually measured by a spectrophotometer to verify the accuracy of competition-based fitness. Briefly, the selected single mutants were individually constructed by the same method as described in sections Construction of the variant libraries and Construction of the variant strain pool above, except that the mutation primer was replaced by a regular primer carrying the specific mutation. Cells were cultured in YPD at 30 °C and 250 r.p.m. overnight and then diluted to OD = 0.1~0.2 and transferred to 96-well plates. Growth was monitored on Epoch2 microplate reader (BioTek) at 30 °C for 12 h by measuring absorbance (OD_{600}) every 10 min. The doubling time (DT) was calculated by contrasting the starting and ending OD of each of these 10-min periods via the following equation[49]

$$\mathrm{DT} = \frac{\ln(2)}{\left(\frac{\ln(\mathrm{OD}_2) - \ln(\mathrm{OD}_1)}{t_2 - t_1} \right)}$$

Here OD_1 and OD_2 are, respectively, the starting and ending OD of the 10-min period. All estimated DTs for the growth range $0.2 < \mathrm{OD} < 0.6$ are averaged and used as the DT estimate of a specific well. Three biological replicates each with three technological repeats (3 × 3) were measured per variant. A wild-type strain was included for each plate to control batch effect. The relative doubling time (Extended Data Fig. 2i) of a variant is the mean doubling time among all repeats (3 × 3) minus the doubling time of wild type.

Extrapolation of the fitness effect of single-nucleotide mutations by linear models. To extrapolate the expression effect on the fitness consequence of single-nucleotide mutations, we fit two lineage models (expression-dominant models, $y = ax$ and mixed models, $y = ax + b$) towards the observed data of each single-nucleotide mutation. Here, x represents the protein abundance and equals to 21,148 for low expression (driven by P_{AGP1}) and 168,876 for high expression (driven by P_{TDH3})[27]. And y represents the fitness effect for each individual biological replicate of the focal mutation. In other words, the mixed models assumed an intrinsic fitness consequence (the 'b' term) of the mutation that is independent of the expression, whereas the expression-dominant models assumed no such expression-independent component. The intercept or slope with $P<0.05$ was considered significant in the fitted models.

Genomic and comparative genomic data. For the assessment of mistranslation, the tRNA gene copy number in the *S. cerevisiae* genome was downloaded from the Genomic tRNA Database[50]. To calculate tAI on the basis of the tRNA gene copy number, we first calculated the absolute adaptiveness value W_i for each codon i:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij})\text{tGCN}_{ij},$$

where n_i is the number of tRNA isoacceptors that recognize codon i, tGCN_{ij} is the gene copy number of tRNA j that recognizes the ith codon and s_{ij} is a selective constraint on the efficiency of the codon–anticodon coupling. The s_{ij} values for eukaryotes were used as indicated in a previous study[51]. Then, we calculated the tAI of a codon as $W_i/W_{\max}$, where $W_{\max}$ is the maximum W_i value.

For the assessment of protein misfolding, we estimated the misfolding probability of a mutant relative to that of the wild-type CDS following our previously published method[21]. Briefly, I-Mutant 2.0 was used to predict the changes in free energy ($\Delta\Delta G$) resulting from amino acid substitutions[33] on the basis of three-dimensional protein structure models downloaded from PDB (ref. [52]) (PDB ID is 1GFL for GFP (ref.[53]) and 3GDL for URA3; ref. [54]). The relative misfolding probability of a mutant compared to that of the wild-type CDS was then calculated as follows

$$P_{\text{misfold}} = \frac{q\prime + \sum_i \left[h_i' e^{-\frac{\Delta\Delta G_{\text{i}}'}{kT}}\right]}{q + \sum_i \left[h_i e^{-\frac{\Delta\Delta G_i}{kT}}\right]} e^{-\left(\frac{\Delta\Delta G}{kT}\right)}$$

Here, $\Delta\Delta G_i$ and $\Delta\Delta G_i'$ are the increases in the unfolding energy caused by the ith translational error in the wild-type and mutant CDSs, respectively. The parameter $\Delta\Delta G_{\text{mt}}$ is the increase in unfolding energy caused by an amino acid mutation in the wild-type CDS. In addition, q and q' are the probabilities that a wild-type and a mutant protein molecule contain no translational errors, respectively. Parameters h_i and h_i' are the probabilities of the ith possible translational error in the wild-type and mutant proteins, respectively. Both q/q' and h_i/h_i' were calculated from the previously estimated probability of mistranslation[21].

For the assessment of protein misinteraction, we used two proxies to estimate the probability of protein misinteraction: the hydrophobicity of surface residues and the IDR. We first used DSSP (ref. [55]) to calculate the solvent accessibility for each amino acid on the basis of the aforementioned protein structure models downloaded from PDB, which was then transformed to the relative solvent accessibility by dividing accessibility value by the surface area of the amino acid[56]. The level of hydrophobicity of each amino acid was retrieved from a previous publication[57]. We used GlobPlot2 (ref. [58]) (command-line parameter 'GlobPlot.py 10 15 74 4 5') to predict the IDRs. For the assessment of mRNA folding requirement, we used RNAfold to predict MFE of mRNA[59].

The above choice of proxies for various types of stochastic molecular errors is certainly worthy of discussion. In particular, although they individually appeared informative in distinguishing strongly error-prone mutations from other mutations in a highly expressed gene (Fig. 4), their difference in quantitative accuracy might severely hinder our assessment for the relative importance of the three hypotheses. It is possible that our conclusion of the superiority of misfolding relative to misinteraction and mRNA folding requirement (Fig. 5) is due to the higher prediction accuracy for misfolding. Nevertheless, from previous benchmarking of these methods, their accuracy appeared comparable[33,60,61], such that the reversion of the superiority of misfolding is unlikely, albeit not impossible.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw data from high-throughput sequencing were deposited to NCBI BioProjects under accession number PRJNA681990.

Code availability

Custom R and python codes were used in data analysis, which are available on Github (https://github.com/woson2020/Experror).

Received: 22 February 2021; Accepted: 1 October 2021;
Published online: 18 November 2021

References

1. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12**, 204–213 (2011).
2. Mackay, T. F., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* **10**, 565–577 (2009).
3. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* **11**, 572–582 (2010).
4. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
5. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
6. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
7. Puchta, O. et al. Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).
8. Taylor, M. B. & Ehrenreich, I. M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* **31**, 34–40 (2015).
9. Mackay, T. F. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).
10. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
11. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
12. Zhang, J. & Yang, J. R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).
13. Li, X., Lalic, J., Baeza-Centurion, P., Dhar, R. & Lehner, B. Changes in gene expression predictably shift and switch genetic interactions. *Nat. Commun.* **10**, 3886 (2019).
14. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).
15. Huang, Y. F. & Siepel, A. Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Res.* **29**, 1310–1321 (2019).
16. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
17. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
18. Huang, Y. F. Unified inference of missense variant effects and gene constraints in the human genome. *PLoS Genet.* **16**, e1008922 (2020).
19. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
20. Yang, J. R., Chen, X. & Zhang, J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.* **12**, e1001910 (2014).
21. Yang, J. R., Zhuang, S. M. & Zhang, J. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* **6**, 421 (2010).
22. Yang, J. R., Liao, B. Y., Zhuang, S. M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl Acad. Sci. USA* **109**, E831–E840 (2012).
23. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
24. Mehlhoff, J. D. et al. Collateral fitness effects of mutations. *Proc. Natl Acad. Sci. USA* **117**, 11597–11607 (2020).
25. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
26. Li, C. & Zhang, J. Multi-environment fitness landscapes of a tRNA gene. *Nat. Ecol. Evol.* **2**, 1025–1032 (2018).
27. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
28. Miller, B. G., Hassell, A. M., Wolfenden, R., Milburn, M. V. & Short, S. A. Anatomy of a proficient enzyme: the structure of orotidine 5′-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc. Natl Acad. Sci. USA* **97**, 2011–2016 (2000).
29. Keren, L. et al. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282–1294 (2016).
30. Faure, G., Ogurtsov, A. Y., Shabalina, S. A. & Koonin, E. V. Role of mRNA structure in the control of protein folding. *Nucleic Acids Res.* **44**, 10898–10911 (2016).
31. Park, C., Chen, X., Yang, J. R. & Zhang, J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **110**, E678–E686 (2013).

32. Sabi, R., Volvovitch Daniel, R. & Tuller, T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* **33**, 589–591 (2017).
33. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, W306–W310 (2005).
34. Protter, D. S. W. et al. Intrinsically disordered regions can contribute promiscuous interactions to RNP granule assembly. *Cell Rep.* **22**, 1401–1412 (2018).
35. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965).
36. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
37. Chen, X. & Zhang, J. The genomic landscape of position effects on protein expression level and noise in yeast. *Cell Syst.* **2**, 347–354 (2016).
38. Brachmann, C. B. et al. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).
39. Gietz, R. D. & Schiestl, R. H. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 35–37 (2007).
40. Storici, F. & Resnick, M. A. The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods Enzymol.* **409**, 329–345 (2006).
41. Mortimer, R. K. & Johnston, J. R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**, 35–43 (1986).
42. Qiu, C. & Kaplan, C. D. Functional assays for transcription mechanisms in high-throughput. *Methods* **159–160**, 115–123 (2019).
43. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143 (2015).
44. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl Acad. Sci. USA* **105**, 4957–4962 (2008).
45. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
46. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl Acad. Sci. USA* **111**, E2310–E2318 (2014).
47. Hoffman, C. S. & Winston, F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* **57**, 267–272 (1987).
48. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinf.* **13**, 238 (2012).
49. Murakami, C. & Kaeberlein, M. Quantifying yeast chronological life span by outgrowth of aged cells. *J. Vis. Exp.* **6**, 1156 (2009).
50. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93–D97 (2008).
51. Chen, F. et al. Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat. Ecol. Evol.* **4**, 589–600 (2020).
52. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
53. Yang, F., Moss, L. G. & Phillips, G. N. Jr. The molecular structure of green fluorescent protein. *Nat. Biotechnol.* **14**, 1246–1251 (1996).
54. Chan, K. K. et al. Mechanism of the orotidine 5′-monophosphate decarboxylase-catalyzed reaction: evidence for substrate destabilization. *Biochemistry* **48**, 5518–5531 (2009).
55. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
56. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**, e80635 (2013).
57. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
58. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).
59. Hofacker, I. L. et al. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188 (1994).
60. Linding, R. et al. Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
61. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).

Acknowledgements

We thank J. Zhang, X. He and W. Qian for comments on the manuscript. This work was supported by the National Natural Science Foundation of China (grant nos. 31771406 to X.C., 81830103 to G.B.T. and J.-R.Y. and 31671320 and 31871320 to J.-R.Y.) and the National Key R&D Program of China (grant nos. 2017YFA0103504 to X.C. and 2018ZX10301402 to J.-R.Y.) and the National Special Research Program of China for Important Infectious Diseases (grant no. 2018ZX10302103 to X.C.) and the start-up grant from '100 Top Talents Program' of Sun Yat-sen University (grant nos. 50000-18821112 to X.C. and 50000-18821117 to J.-R.Y.).

Author contributions

J.-R.Y. and X. Chen conceived the idea and designed and supervised the study. Z.W., X. Cai and Y.L. conducted experiments and acquired data. X. Cai, G.-B.T., J.-R.Y. and X. Chen contributed new reagent and analytical tools. Z.W., X.Z., X. Chen and J.-R.Y. analysed data. Z.W., X. Chen and J.-R.Y. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01578-x.

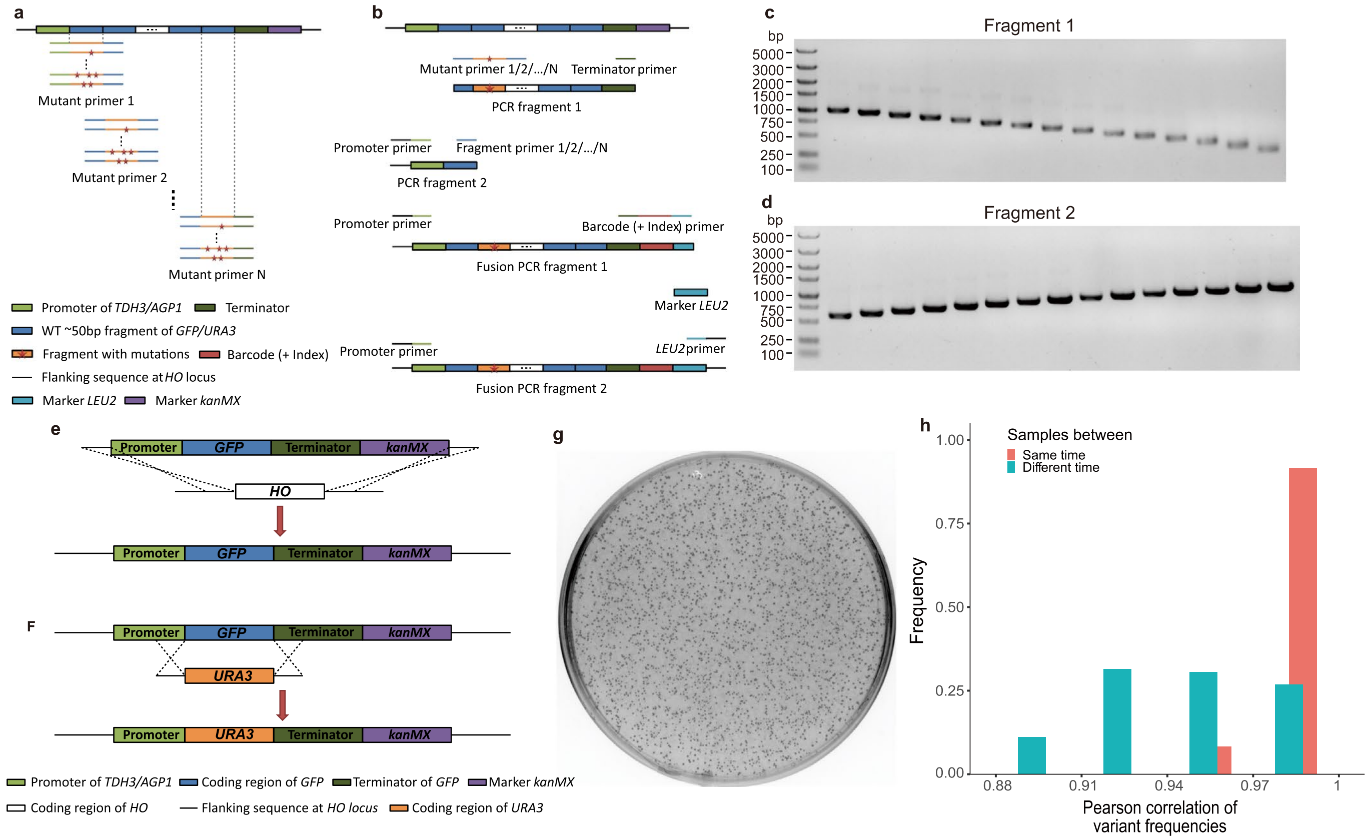
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01578-x.

Correspondence and requests for materials should be addressed to Jian-Rong Yang or Xiaoshu Chen.

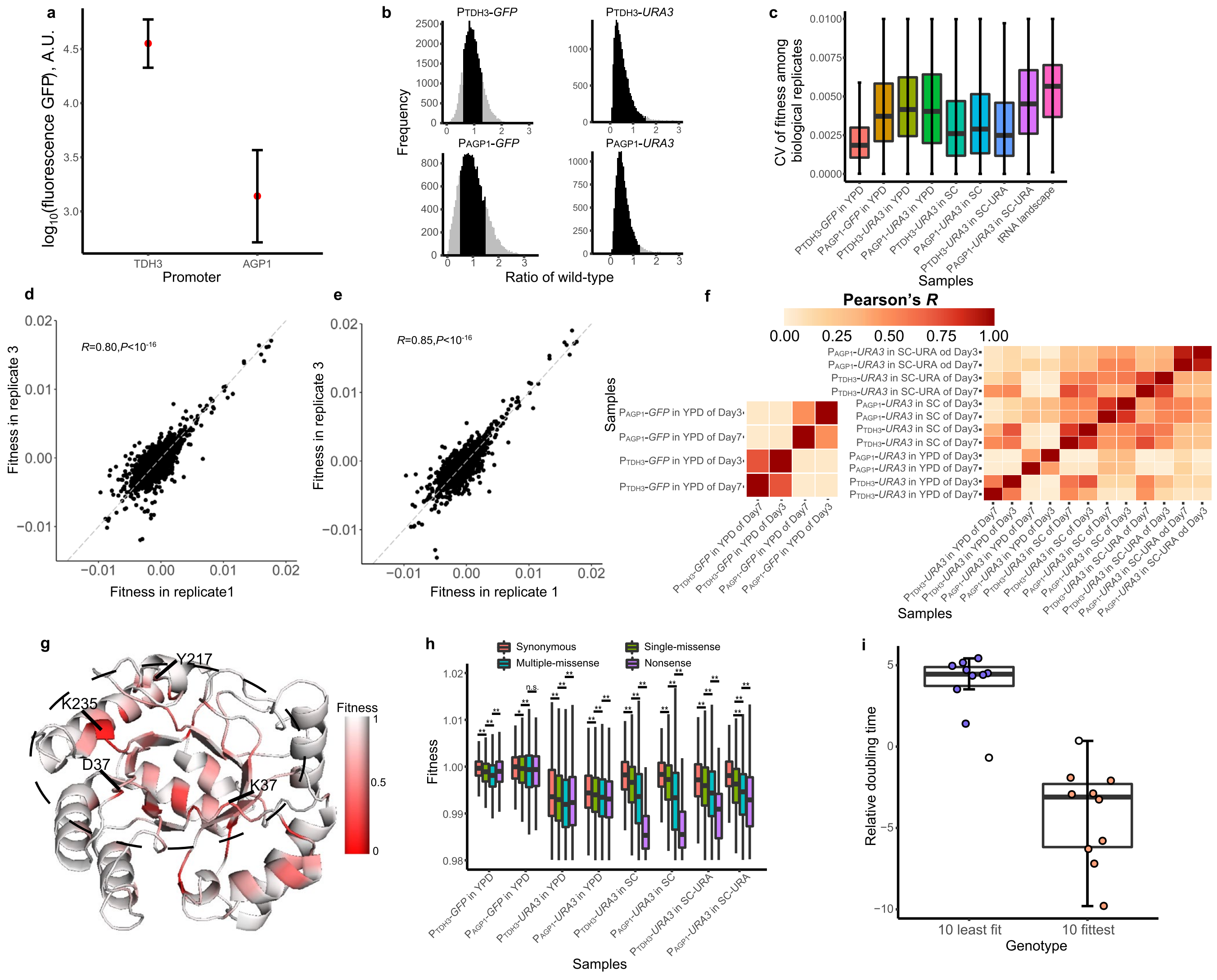
Peer review information *Nature Nature Ecology & Evolution* thanks Yuping Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

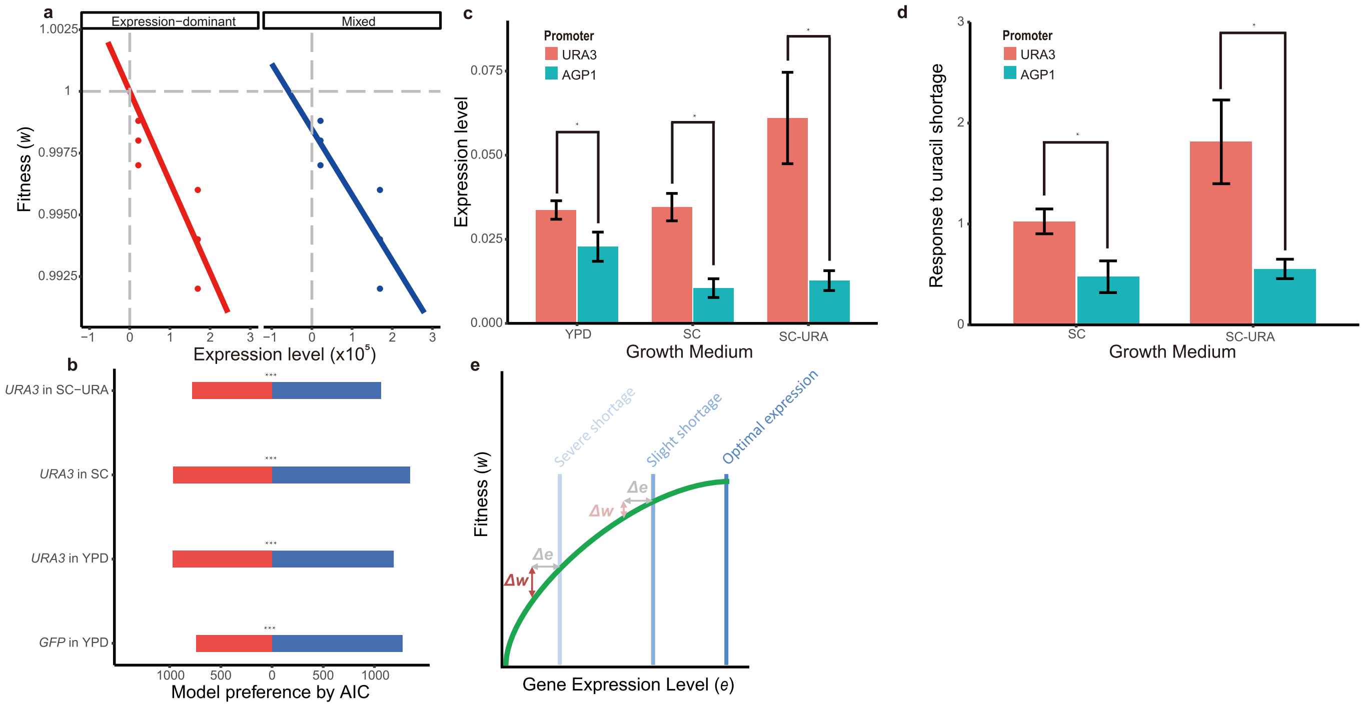
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



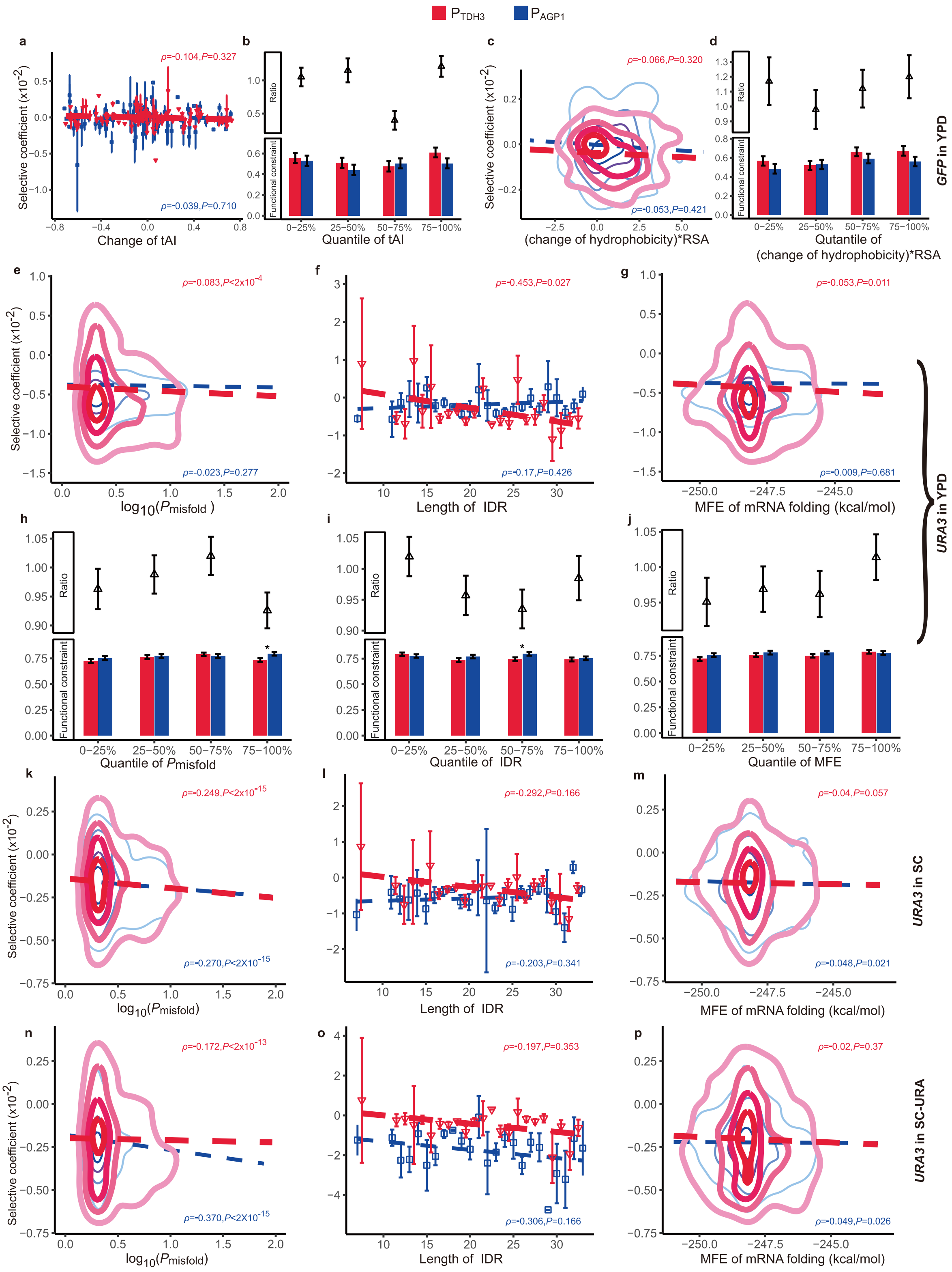
Extended Data Fig. 1 | Additional information on the experimental pipeline. (**a**) The CDS of the gene (*GFP* or *URA3*) was divided into non-overlapping regions of 50 bp. For each 50-bp region, one mutant primer was synthesized with the focal 50-bp region by using doped nucleotides (therefore introducing mutations) and a 20-bp invariable sequence flanking either side of the 50-bp region; see also Fig. 1a, Methods and Supplementary Table 6. (**b**) Different mutant primers were used in combination with the terminator primer to amplify PCR fragment 1 and in combination with the promoter primer to amplify PCR fragment 2. Then, PCR fragments 1 and 2 were fused using promoter primers and barcode (+ index) primers, giving rise to fusion PCR fragment 1, which was further fused with the *LEU2* marker. The final product, fusion PCR fragment 2, was ready for recombination; see also Fig. 1a, Methods and Supplementary Table 6. (**c** and **d**) Electrophoresis results for PCR fragment 1 (**c**) and PCR fragment 2 (**d**) of GFP. (**e** and **f**) Construction of the acceptor strain. The *HO* locus in *S. cerevisiae* strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) was individually replaced by two expression cassettes, P_{TDH3}-*GFP-KanMX* and P_{ADP1}-*GFP-KanMX*, via homologous recombination to construct two *GFP* acceptor strains (**e**). The two *URA3* acceptor strains were constructed by replacing the CDS of *GFP* in the *GFP* acceptor strains via homologous recombination (**f**). (**g**) Typical transformation results on plates selective for transformants, showing a relatively high transformation rate. (**h**) Histogram of between-sample correlations of genotype frequencies. The correlations were stratified as correlations between samples from the same timepoint (that is, biological replicates) (red) or different timepoints (green).



Extended Data Fig. 2 | The measured fitness landscape is overall accurate. (**a**) Protein expression levels driven by the two promoters of P_{TDH3} and P_{AGP1} assayed by flow cytometry using the corresponding GFP acceptor strains. The error bar represents standard deviation from 9 replicates (3 biological × 3 technological). (**b**) Distribution of reads ratio (*Day*$_7$/*Day*$_0$) of wild-type barcodes in four mutant libraries. Only wild-type barcodes whose reads ratios were no more than one standard deviation away from their average (black bars) were pooled and used for the estimation of read ratio of wild-type. (**c**) Coefficient of variation (CV) of fitness among biological replicates was calculated for each genotype, and collectively shown as a standard boxplot for all genotypes within a library. As comparison, similar estimates were shown for the previously reported fitness landscape of a tRNA gene. (**d** and **e**) Comparison of fitness of P_{TDH3}-GFP variants growing in YPD on day 7 among biological replicates. Pearson's correlation coefficients and the corresponding *P* values are shown. Correlations of other libraries were listed in Supplementary Table 3. (**f**) Correlation of fitness estimates from day 3 and day 7, and that among different growth media. The Pearson's correlation coefficients are shown by colours indicated by the colour scale bar. (**g**) Enrichment of deleterious mutations in activity centre of URA3. The fitness of each amino acid was calculated by the average fitness of all single mutants of the corresponding codon. The labelled amino acid are four known active site and the dashed ellipse outlines the activity centre of URA3. (**h**) Distributions of fitness for four types of mutants are shown as boxplots for each library. The four types of mutants are categorized as follows: synonymous (having one or more synonymous mutations), single missense (having one missense mutations, additional synonymous mutations allowed), multiple missense (having two or more missense mutations, additional synonymous mutations allowed), nonsense (having at least one nonsense mutations, additional missense or synonymous mutations allowed). Statistical significance of differences by Mann–Whitney U-test were indicated by asterisks: *: $P < 0.05$; **: $P < 0.001$; n.s.: not significant. (**i**) Ten least-fit and ten fittest genotypes picked from the P_{TDH3}-GFP library growing in YPD were individually measured by a spectrophotometer for their doubling time, which was further subtracted by the doubling time of wild-type and thereby plotted as relative doubling time (*y* axis). The relative doubling time of each genotype was tested for significant deviation from 0 by Mann–Whitney U-test using the three biological replicates, giving rise to filled circles for significant genotypes or empty circles for insignificant genotypes.

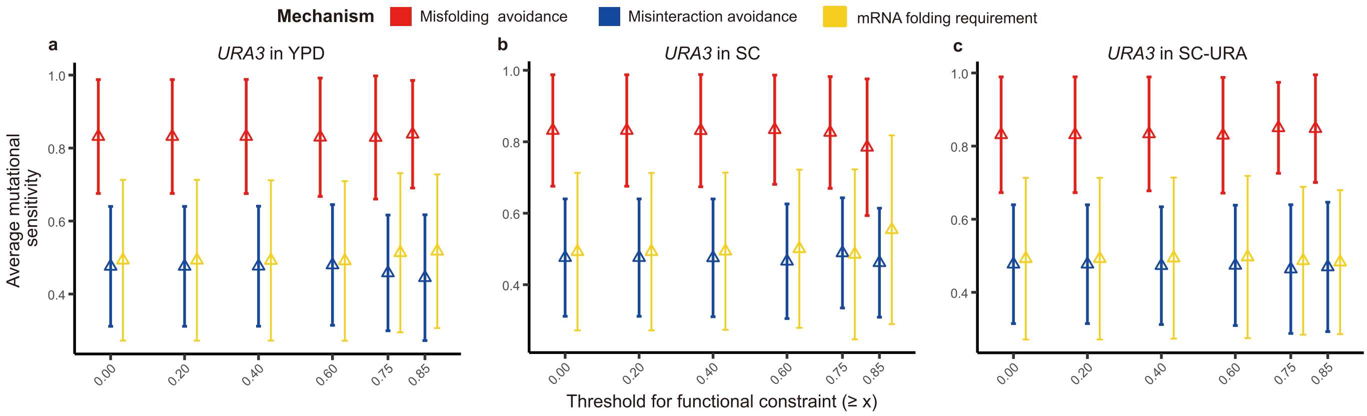


Extended Data Fig. 3 | Extrapolation of the expression-dependent and independent components of the fitness effect of CDS mutations. (**a**) Two types of linear models were regressed for the fitness effect of each single-nucleotide mutation (*y* axis) using the expression levels of P_{TDH3} and P_{AGP1}. In the 'expression-dominant' model, no expression-independent components were assumed, whereas the 'mixed' model contained both expression-dependent and expression-independent components. (**b**) The quality of the expression-dominant and mixed models in describing the data were assessed according to the Akaike information criterion (AIC). The number of mutations better described by the mixed model (blue bar) was always higher than the number better described by the expression-dominant model (red bar), regardless of the fitness landscapes (*y* axis) used. Binomial *P* values against the null expectation of equal preference for both models are indicated as **: $P < 10^{-3}$; ***: $P < 10^{-5}$. (**c** and **d**) The relative (to *ACT1*) activities of different promoters in yeast strain S288c growing in different media were measured by RT–qPCR of expression levels of corresponding native genes (c). The response of each promoter to uracil shortage was calculated as the ratio between its activity in SC/SC-URA and that in YPD (d). In both panels c and d, error bar represents the standard deviation of six replicates, and *P* values from Mann–Whitney U-tests are indicated as *: $P < 0.05$. (**e**) A simple model explaining why a functionally required gene is more sensitive to deleterious mutations when it is lowly expressed compared to when it is highly expressed. The green curve represents the relationship between expression level of the gene (*e*, on *x* axis) and organismal fitness (*w*, on *y* axis). Assuming diminishing return, that is, $dw/de = f(e)$ and *f* is a monotonic decreasing function. The dark, medium and light blue vertical lines represent optimal, slight shortage and severe shortage of gene expression, respectively. A deleterious mutation should trigger a small loss of function of the gene that is effectively equivalent to a small reduction of expression (Δe, the two grey segments), which shall lead to a corresponding reduction in fitness (Δw, the red or pink segments). Apparently, the same Δe should give rise to smaller Δw when *e* is higher at 'slight shortage' (the pink segment) compared to when *e* is lower at 'severe shortage' (the red segment), because $\Delta w/\Delta e = f(e)$ is smaller for larger *e* due to diminishing return. This relationship is also intuitively shown in the figure.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Additional tests on the error avoidance models. (**a** and **b**) These panels are similar to Fig. 4e, f except that the *t*RNA *a*daptation *i*ndex (tAI) was used as a proxy to test the mistranslation avoidance hypothesis. (**c** and **d**) These panels are similar to Fig. 4c and d except that the mutation-triggered increment of hydrophobicity was multiplied by the *r*elative *s*olvent *a*ccessibility (RSA) as a proxy for the probability of misinteraction so that the misinteraction avoidance hypothesis could be tested. Note that only the amino acids at the protein surface (RSA > 0.4) were considered. Altering the RSA criteria for the protein surface would not change the conclusion (data not shown). (**e**-**j**) These panels are similar to Fig. 4c, e, g, d, f, h except that the fitness landscape of *URA3* in YPD was used. These results were largely consistent with those presented in Fig. 4c–h, except for the functional constraints for misinteraction (i). For the mutations within the 75-100% quantile of P_{misfold}, we suspected that the fraction of misfolded URA3 was too high, making the number of correctly folded URA3 molecules insufficient when gene expression was driven by P_{AGP1} compared to when it was driven by P_{TDH3}. (**k**-**m**) These panels are similar to Fig. 4c, e, g except that the fitness landscape of *URA3* in SC was used. (**n**-**p**) These panels are similar to Fig. 4c, e, g except that the fitness landscape of *URA3* in SC-URA was used.



Extended Data Fig. 5 | Relative contribution of different error avoidance models in other fitness landscapes. (**a**-**c**) These panels are similar to Fig. 5c except that other measured fitness landscapes, as indicated on top of each panel, were used to estimate the relative contribution of misfolding avoidance (red), misinteraction avoidance (blue) and the mRNA folding requirement (green).

nature research

Corresponding author(s): Jian-Rong Yang

Last updated by author(s): Sep 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Raw sequencing data were directly obtained from sequencing service providers, without specific softwares invovled.
Data analysis	All codes used in data analyses was available on https://github.com/woson2020/Experror

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw data generated in this study will be available on NCBI BioProjects via accession number PRJNA681990

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Competitive co-culture experiments were replicated two or three times. The number of barcoded molecules for each genotype were directly determined from the sequencing data.
Data exclusions	Data exclusions applied during the data analyses were detailed in the Method section. Briefly, we excluded barcodes that cannot be reliably matched with genotypes, and we exclude some wild-type barcodes if they displayed significant deviation from other wild-type barcodes in terms of frequency change.
Replication	Competitive co-culture experiments were replicated two or three times.
Randomization	Each libraries were repetitively measured for the fitness landscape, with no need for randomization.
Blinding	Data analyses were conducted with identical pipeline for all fitness landscapes, therefore blinding is not necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging