

# Mistranslation suppresses mistranscription in eukaryotes

Received: 24 May 2025

Accepted: 12 February 2026

Published online: 25 February 2026

 Check for updates

Xiaoyi Zhang<sup>1</sup>, Gongwang Yu<sup>2</sup>, Ziyao Guo<sup>3</sup>, Jia Liao<sup>3</sup>, Weiyi Li<sup>1,3</sup>, Weijie Zhang<sup>3</sup>, Boyang Zheng<sup>1</sup>, Zhuoxing Wu<sup>3</sup>, Shuya Peng<sup>3</sup>, Dahui Tan<sup>3</sup>, Jian-Rong Yang<sup>3,4,5</sup>✉ & Xiaoshu Chen<sup>1,4</sup>✉

Phenotypic mutations include non-heritable sequence changes arising from synthetic errors such as mistranscription or mistranslation. Despite wide potential impact of both error types, their relationship remains under-explored. Using Circ-Seq and mass spectrometry data, we perform a genome-wide analysis of mistranscriptions and mistranslations in five model organisms, and find that genes with frequent mistranslations exhibit lower mistranscription rates. We hypothesize that this pattern is explainable by a negative epistasis between the two error types. We assess our hypothesis with systematic experimental measurements that confirm the prevalence of negative epistasis, as well as *in silico* evolutionary simulations that suggest the extra deleterious effects resulting from negative epistasis are sufficient to create selection against mistranscription in highly mistranslated genes. Moreover, empirical genomic analyses indicate that genes with frequent mistranslation purge nonsynonymous mistranscriptions more efficiently, and that highly translated transcripts show fewer mistranscriptions. Combined, our results reveal an unrecognized interaction between mistranscription and mistranslation.

Biological activities critically depend on the accurate expression of genetic information, which is largely determined by the molecular processes of transcription and translation. The fidelity with which they convey genetic information is, however, not ideal. It has been estimated that mistranscription, which is the incorporation of a nucleotide other than that encoded by the DNA template, occurs at an average rate of  $10^{-5}$  to  $10^{-6}$  per nucleotide during transcription<sup>1,2</sup>. The rate of mistranslation, i.e., the incorporation of an amino acid not encoded by the codon being translated on the mRNA, is on average  $10^{-3}$  to  $10^{-4}$  per codon<sup>3,4</sup>.

Besides producing erroneous proteins with reduced functional capacity, the biological consequences of mistranscription and mistranslation go far beyond the associated waste of time, energy, and

resources. For example, mistranscription has been shown to activate oncogene<sup>4</sup>, impair nucleotide and amino acid metabolism<sup>1</sup>, and shorten cellular lifespan<sup>5</sup>, among other effects<sup>6–8</sup>. Transcription error of skipping nucleotides produces toxic APP and UBB peptides, which are neuropathological markers of non-familial Alzheimer's disease<sup>9,10</sup>. Mistranslation, on the other hand, causes protein misfolding<sup>11,12</sup> and misinteraction<sup>13,14</sup>, induces proteotoxic stress by overwhelming the cell's protein quality control system<sup>11</sup>, leads to neurodegenerative diseases<sup>15,16</sup>, and correlates with aging<sup>17–19</sup>. Theoretically, all these phenotypic consequences of mistranslation can also be caused by nonsynonymous mistranscription, although at a lower rate.

Given these phenotypic effects, it is not surprising that mistranscription and mistranslation also play important roles in the

<sup>1</sup>Department of Microbiology, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. <sup>2</sup>The Affiliated Dongguan Songshan Lake Central Hospital, Guangdong Medical University, Dongguan, China. <sup>3</sup>Department of Medical Genetics and Bioinformatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. <sup>4</sup>Laboratory of Tropical Disease Control, Ministry of Education, Sun Yat-sen University, Guangzhou, China. <sup>5</sup>Guangdong Provincial Highly Pathogenic Microorganism Science Data Center, Guangzhou, China. ✉e-mail: [yangjianrong@mail.sysu.edu.cn](mailto:yangjianrong@mail.sysu.edu.cn); [chenxshu3@mail.sysu.edu.cn](mailto:chenxshu3@mail.sysu.edu.cn)

evolution of the genome. Natural selection against the cellular toxicity of mistranslated proteins has been suggested as the major force behind the more constrained sequence evolution of highly expressed genes, as otherwise they would produce a great number of erroneous proteins<sup>12–14,20–22</sup>. Additionally, mistranscription and mistranslation are both phenotypic mutations, which on the one hand decrease population-average fitness by creating deviations from the optimal phenotype<sup>23</sup>, but on the other hand facilitate evolution by creating diversity that ensures population survival in extreme environments<sup>24–26</sup>, purging slightly deleterious mutations through their negative epistasis with the mutations<sup>27</sup>, or promoting further heritable phenotypic changes<sup>28,29</sup>.

Despite the functional and evolutionary importance of both mistranscription and mistranslation and the inherent connection between them as sources of erroneous proteins, very little is known about their relationship, as previous studies have focused on only one or the other, rather than both<sup>30,31</sup>. It may be hypothesized, for example, that mistranscription makes downstream sequences more susceptible to mistranslation<sup>32</sup>, or that functionally more important genes may be subjected to more stringent natural selection<sup>11,12,20,33</sup> against transcription and translation errors, whereas less important genes may be subjected to a more relaxed selection against these errors. Considering this, comparing the rates of mistranscription and mistranslation across the genome should provide unique insight into the relationship between these two types of errors. As an example, both hypotheses mentioned above predict that mistranscription and mistranslation rates across different genes within the genome are positively correlated.

In this study, to investigate the relationship between mistranscription and mistranslation, we estimate the rate of mistranscription and mistranslation at the genome-scale in five model eukaryotes. Despite confirming the tenfold difference between mistranslation rates per codon and mistranscription rates per nucleotide, we also found that the difference between mistranslation and mistranscription probabilities per gene was less than threefold. Unexpectedly, we found strong evidence that genes with frequent mistranslation tend to have a lower mistranscription rate. We hypothesized that this pattern is created by the negative epistasis between the two types of errors, and tested this hypothesis via systematic experimental measurements of within-gene epistasis between genetic mutations. It is revealed that such epistasis is predominantly negative, which makes proteins affected simultaneously by both types of errors significantly more deleterious than expected. More importantly, an *in silico* simulation of molecular evolution suggests that the extra deleterious effects caused by the negative epistasis, when scaled by the error rates, can facilitate selection against mistranscription for genes with frequent mistranslation. Finally, our hypothesis is further corroborated by two independent lines of empirical evidence, including that genes with more frequent mistranslation purge nonsynonymous mistranscription more efficiently, and genes that are more frequently translated and therefore with higher penetration of mistranscriptions tend to exhibit a decreased mistranscription rate. Collectively, our results demonstrated a previously unknown interaction between mistranscription and mistranslation due to their negative epistasis.

## Results

### Genome-wide measurement of mistranscription and mistranslation rates

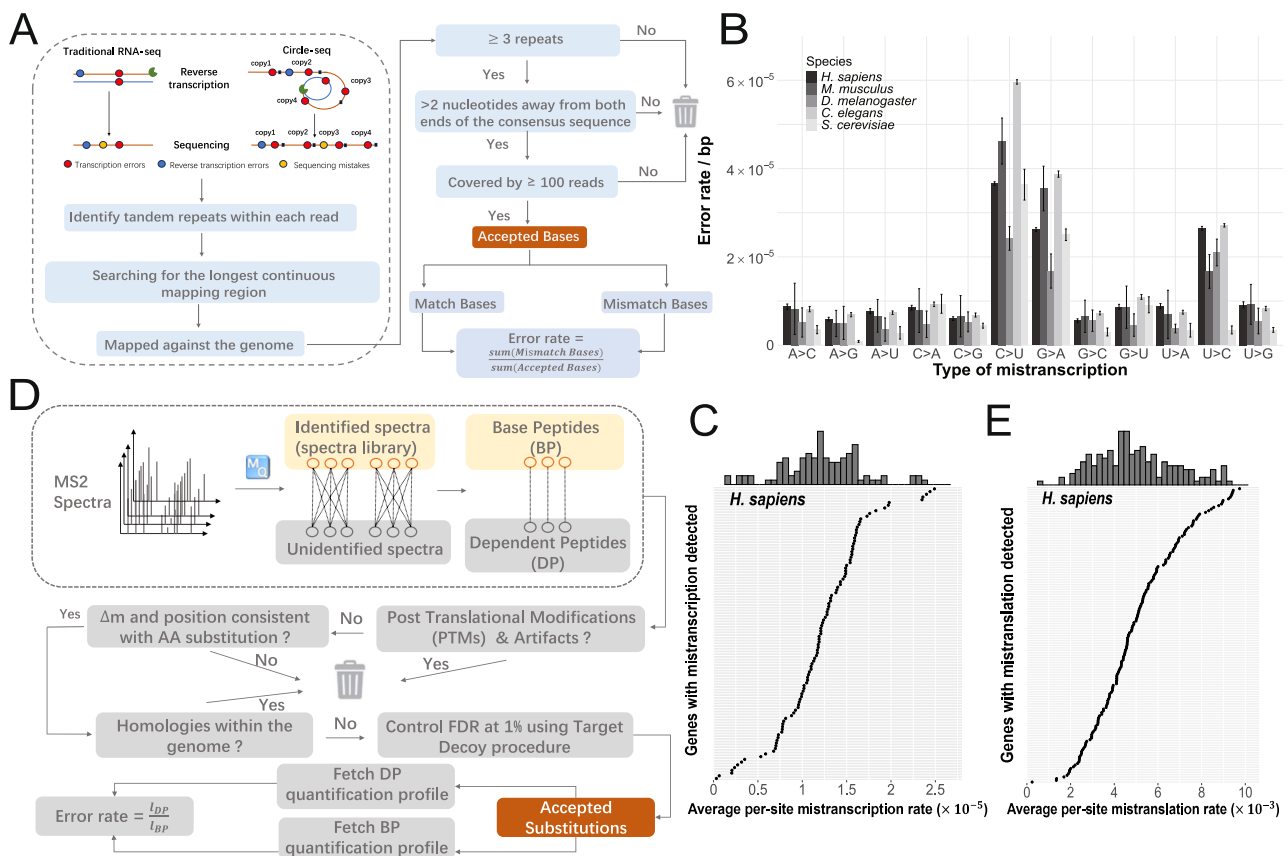
To detect mistranscription at genome scale, we used data from Circ-Seq, in which RNA template molecules are circularized by ligation to themselves and then reverse-transcribed to generate linear cDNA molecules with tandem copies of the original template sequence (Fig. 1A). A mistranscription can thereby be identified as recurring on all tandem copies, as opposed to reverse-transcription/PCR/sequencing errors occurring in only one copy<sup>1,34,35</sup>. We collected and analyzed

Circ-Seq data from *Homo sapiens*, *Mus Musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (Supplementary data 1), using a unified computational pipeline adapted from previous publications<sup>1,36,37</sup> (Fig. 1A and “Methods”), and identified high-confidence mistranscription in 125, 99, 436, 75 and 37 protein coding genes in these five model organisms, respectively (Supplementary data 2). Consistent with previous reports<sup>1,36,37</sup>, the mistranscription rates of C>U, G>A, and U>C are almost always higher than other types of mistranscription (Fig. 1B). In *H. sapiens*, the gene-wise average per-nucleotide mistranscription rate of the genes with detected mistranscriptions exhibits a near-normal distribution, with 95% confidence interval at  $1.2 \times 10^{-5}$  to  $1.3 \times 10^{-5}$  (Fig. 1C. Other species in Supplementary Fig. S1). The low mistranscription rates and the rare overlap with polymorphisms (e.g., only 1.5% human mistranscriptions overlap with polymorphisms in dbSNP) are indicative of genuine mistranscriptions rather than genetic polymorphisms.

To identify mistranslation, we used a mass-spectrometry(MS)-based method<sup>31</sup> that detects the mass shifts strictly explainable only by amino acid misincorporations but not by post-translational modifications (PTMs) or RNA editing (Fig. 1D. See “Methods”). The rate of a specific mistranslation event is calculated as the MS signal intensity of the misincorporation-containing peptide, divided by the sum of the MS signal intensities of the misincorporation-free peptides and the misincorporation-containing peptides. As a result of applying this method to high-throughput MS data from the five model organisms mentioned above (Supplementary data 3), we found mistranslations in 174 human genes, 190 mouse genes, 95 fly genes, 55 nematode genes, and 49 yeast genes (Supplementary data 4). The vast majority of the mistranslation events identified correspond to native codons being incorrectly decoded by near-cognate tRNAs rather than to artefacts such as polymorphisms (e.g., only 1.7% human mistranslations can be explained by polymorphisms in dbSNP), and the competition for decoding between cognate and near-cognate tRNAs significantly correlates with the mistranslation rate<sup>38,39</sup> (Supplementary Fig. S2). In *H. sapiens*, the gene-wise average per-codon mistranslation rates for all genes with detected mistranslations exhibit a near-normal distribution with 95% confidence interval at  $4.7 \times 10^{-3}$  to  $5.3 \times 10^{-3}$  (Fig. 1E and Supplementary Fig. S3). Collectively, these results suggest that we are able to estimate mistranscription and mistranslation rates accurately enough to capture major biological signals.

### Transcription is more accurate than translation, but not by much on a per-gene basis

The genome-scale estimation of mistranscription and mistranslation rates allows a direct comparison of the two, which indicates that the per-codon mistranslation rate ( $4.98 \times 10^{-3}$ ) is on average 400-fold higher than the per-nucleotide mistranscription rate ( $1.25 \times 10^{-5}$ , Fig. 2A) in *H. sapiens* (other species in Supplementary Fig. S4A). This value is largely consistent with previous reports<sup>30,38–40</sup>. We have found, however, that when extrapolating (see “Methods”) the probability of errors per gene based on the average per-site error rate and the number of sites across the full coding sequence, the magnitudinal difference in mistranslation and mistranscription is reduced to only 2.9-fold (Fig. 2B. See also Supplementary Fig. S4B). Note here that the probability of errors per gene is defined as the probability of a full-length product of the gene containing at least one error of a specific type, and our extrapolation likely over-estimates mistranslation relative to mistranscription (see “Methods”). This pattern is mainly attributable to two factors. First, an mRNA is at least three times longer than the protein it encodes, even when non-coding regions such as UTRs and introns are not considered. Second, the near-linear accumulation of per-site error rates to per-gene error probabilities gradually decelerates as it approaches saturation (i.e., per-gene error probabilities of -1). Therefore, the higher per-site mistranslation rate dictates that its accumulation with protein length rapidly decelerates,



**Fig. 1 | Genome-wide estimation of mistranscription and mistranslation rate.** **A** An overview of the method for detecting mistranscriptions. **B** The mistranscription rates of all 12 types of mistranscription among the five species ( $n = 125, 99, 436, 75, 37$ , from *H. sapiens* to *S. cerevisiae*). The bar and error bar represent, respectively, the mean and standard error of mistranscription rate for each type of mistranscription. **C** The average per-site mistranscription rate (x axis) of all human genes with detected mistranscriptions (y axis). Mistranscription rates

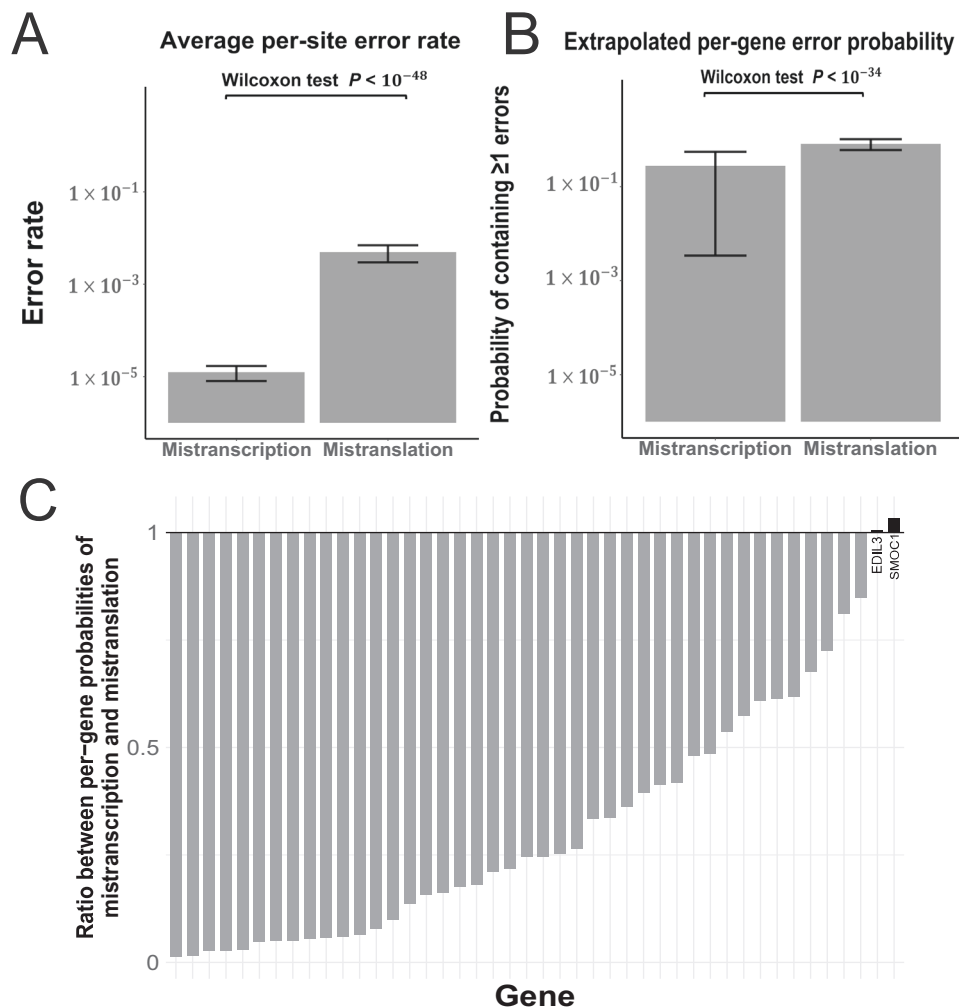
across genes are normally distributed (Kolmogorov–Smirnov test  $P = 0.67$ ). See Fig. S1 for other species. **D** An overview of the method for detecting mistranslations. **E** The average per-site mistranslation rate (x axis) of all genes with detected mistranslations (y axis). Mistranslation rates across genes are normally distributed (Kolmogorov–Smirnov test  $P = 0.40$ ). See Fig. S3 for other species. For panel C and E, a histogram summarizing the distribution is displayed on top of each panel. Source data are provided as a Source data file.

while the same accumulation remains near-linear for the lower per-site mistranscription rate (Supplementary Fig. S4C). We emphasize that while our estimation of average per-site error rate may be inaccurate due to variation in Circ-Seq/MS coverage, these two factors underlying the narrowed difference between the per-gene probabilities of mistranslation and mistranscription are likely to remain valid due to their mathematical nature.

The above observation points to the possibility that some genes might have a per-gene probability of mistranscription higher than that of mistranslation. Indeed, among human genes with both types of errors detected, we found two genes exhibiting higher probabilities of mistranscription than mistranslation (Fig. 2C), namely EDIL3 (EGF-like repeats and discoidin domains 3) and SMOC1 (encoding SPARC Related Modular Calcium Binding 1). Interestingly, these two genes are closely associated with cancer invasion and metastasis. Specifically, EDIL3 is involved in epithelial-mesenchymal transition (EMT)<sup>41,42</sup>. Variations in EDIL3 sequence are significantly associated with basal subtype breast cancer<sup>43</sup>. SMOC1 has been suggested as a tumor suppressor whose mutation is associated with poor prognosis in glioma<sup>44</sup>. It is also worth noting that, as all residues in the mRNA/protein are susceptible to mistranscription/mistranslation, the prevalence of erroneous products should be comparable or even exceed many commonly observed post-transcriptional/post-translational modifications. These highlight the biomedical significance of mistranscription and mistranslation, and suggest a possible cross-talk between them, as both are substantial sources of phenotypic mutations.

### Genes with high mistranslation rates tend to have a lower mistranscription rate

To directly compare mistranscription and mistranslation rates, we will hereinafter focus on the average per-site rates of mistranscription and mistranslation of a gene. And unless otherwise stated, results based on genes exhibiting at least one type of error are presented, whereas additional analyses that include genes exhibiting both types of errors, or all expressed genes regardless of the presence of errors, are shown in Supplementary Fig. S5. Gene-by-gene comparisons of mistranscription and mistranslation rates revealed that genes with high mistranslation rates tend to have a lower mistranscription rate (Fig. 3A–E). More importantly, such “trade-off” between mistranscription and mistranslation rates was consistently observed across all five model organisms, ranging from yeasts to humans, suggesting that the relationship is highly conserved. To further strengthen the reliability of this observation and eliminate the possibility of statistical artefacts, we assess this trade-off by four additional tests. First, we construct a  $2 \times 2$  contingency table by classifying each gene into one of four categories based on its mistranscription and mistranslation rate, and calculated an odds ratio from the table (Fig. 3F). The smaller the odds ratio, the stronger the support for the trade-off. We found that the odds ratios from all five model organisms are all significantly smaller than 1 (Fig. 3F,  $P < 0.001$ , Fisher’s exact test). Second, we analyzed all genes with non-zero coverage in both Circ-Seq and MS, regardless of whether they had mistranscription or mistranslation detected, and were able to replicate the negative correlation with Tobit model, which accounted



**Fig. 2 | Comparing the human mistranscription rate with the mistranslation rate.** Gene-wise average per-site rates (A) and extrapolated (see Methods) per-gene rates (B) of mistranscription and mistranslation ( $n = 125, 174$ ). The error bars represent the standard error of the mean across human genes. All genes with either mistranscription or mistranslation detected are considered.

Statistical significance was assessed using a two-sided Wilcoxon signed-rank tests. C For genes with both types of errors detected (x axis), the ratio between their extrapolated per-gene rates of mistranscription and mistranslation (y axis) is presented. A gene is colored black if its mistranscription rate exceeds its mistranslation rate, or gray otherwise. Source data are provided as a Source data file.

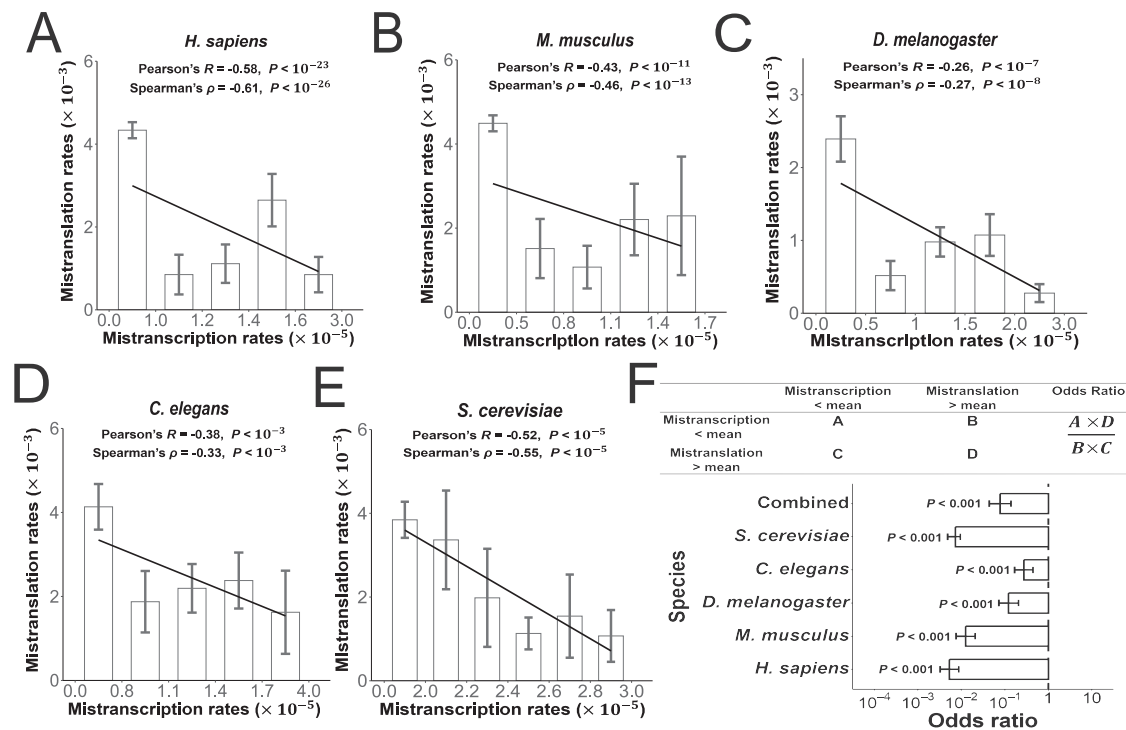
for gene expression as a confounding factor for error detection (Supplementary Fig. S5A, B. See “Methods”). Third, using all genes with both mistranscription and mistranslation detected, we again observed the negative correlations. Specifically, although the results are statistically insignificant due to the reduced sample size, their average value is significantly more negative than expected (Supplementary Fig. S5C), and the categorical analyses by contingency table and odds ratio remained significant by the collective Mantel-Haenszel test (Supplementary Fig. S5D). Fourth, all correlation-based results above are readily reproducible by weighted correlations using the inverse of the standard errors of the estimated error rates as weights, thereby precluding biased conclusions due to large estimation errors (Supplementary Fig. S6). These results suggest a conserved trade-off relationship between mistranscription and mistranslation rates.

### Negative epistasis between mistranscription and mistranslation may drive their anticorrelation

Our unexpected finding of a trade-off between mistranscription and mistranslation prompted us to investigate the underlying mechanism. While it is highly improbable that there would be a gene-specific molecular regulator that could sense one type of error and reverse-regulate the other, we instead sought to look for an evolutionary

mechanism. Our attention was drawn to a recent study revealing that mistranslation facilitates the selective purging of slightly deleterious mutations due to its negative epistasis with the mutations<sup>27</sup>. This report led us to similarly hypothesize that mistranslation can help eliminate deleterious mistranscriptions and thus reduce a population’s error load via its negative epistasis with mistranscription in the same gene. In other words, gene products affected by both types of errors are far more harmful than expected, which will result in a stronger natural selection against mistranscription among genes with a higher mistranslation rate, thus creating a trade-off between mistranscription and mistranslation.

To test our hypothesis, a direct quantification of the negative epistasis was required to determine if it is sufficiently pervasive and strong to explain the mistranscription-mistranslation tradeoff. Notably, this test should be conducted with three specific considerations in mind. First, negative epistasis can be caused by a greater-than-expected functional reduction of proteins containing both errors, but it can also be caused by the cellular toxicity<sup>11,12,14,33</sup> of extra misfolded proteins resulting from simultaneous mistranslation and mistranscription. Here, the impact of functional reduction heavily depends on the specific genes considered, whereas the extra toxicity is likely to be independent of the gene’s function and therefore more



**Fig. 3 | Genes with high mistranslation rates tend to have a lower mistranscription rate.** The mistranscription and mistranslation rates are negatively correlated in *H. sapiens* ( $n = 255$ ) (A), *M. musculus* ( $n = 245$ ) (B), *D. melanogaster* ( $n = 473$ ) (C), *C. elegans* ( $n = 95$ ) (D), and *S. cerevisiae* ( $n = 66$ ) (E). All genes with either mistranscription or mistranslation detected were grouped into bins defined by uniformly sized intervals of mistranscription rates, except for the first and last bins, which were extended to include outliers at the lower and upper ends of the distribution of mistranscription rates. The bars and error bars represent, respectively, the mean and standard error of mistranslation rates for genes in each bin. The black line represents the linear regression fitted to the binned average mistranslation rates. Pearson's and Spearman's correlation tests were performed

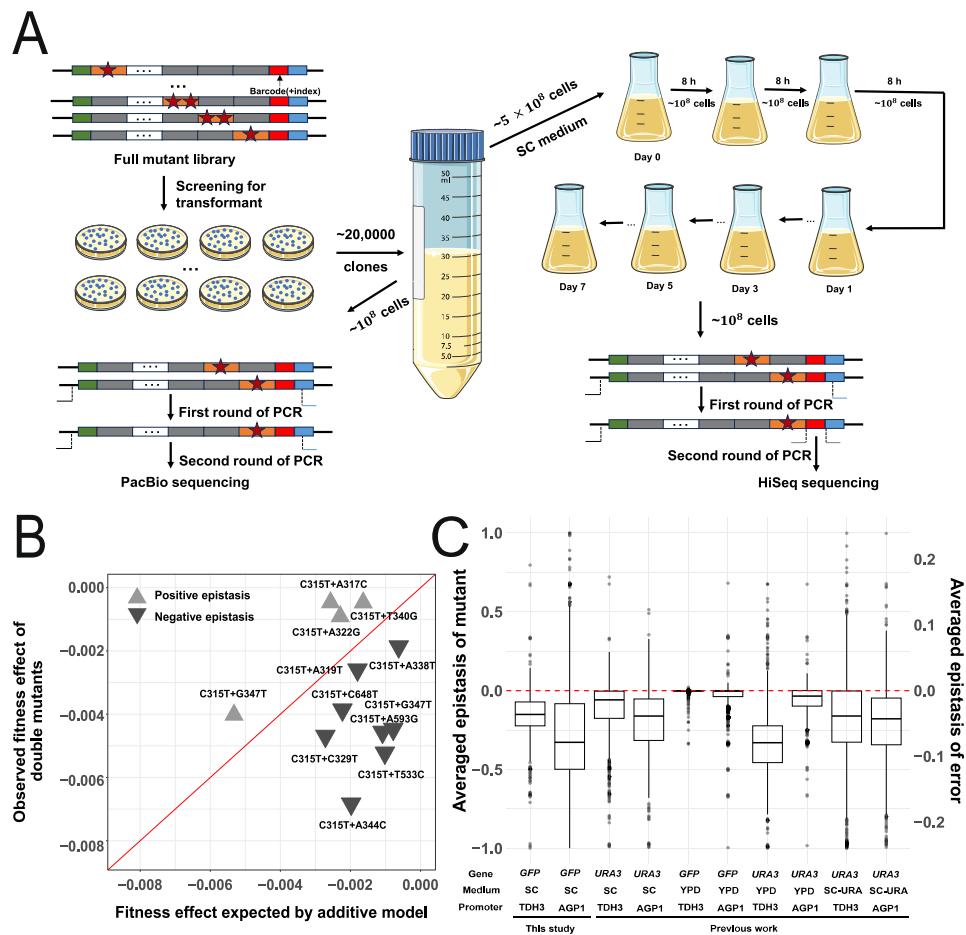
on the unbinned raw data, with statistical significance indicated at the top of each panel. **F** We constructed a  $2 \times 2$  contingency table by classifying each gene into one of four categories based on its mistranscription and mistranslation rates, and calculated an odds ratio from the table (top). The odds ratios calculated from the five model organisms are shown (bottom) as horizontal bars, with error bars representing the 95% confidence interval assessed by 1000 bootstraps of the genes. Fisher's exact test was used to determine the statistical significance as indicated on each bar. Thresholding by median instead of mean values gives qualitatively identical results. The contingency tables from the five species were collectively tested by Mantel-Haenszel tests, giving rise to the combined odds ratio and  $P$  value on top. Source data are provided as a Source data file.

generally applicable. Consequently, a gene with no native function in the cell should be tested so that the result is not dependent on the gene's function. Second, since mistranslations and mistranscriptions can occur anywhere in the CDS, negative epistasis should be evaluated across the entire CDS rather than within some specific functional domains. Last but not least, fitness or growth rate should be measured directly rather than some other types of functional capacities (such as fluorescence), as fitness is most important for natural selection and evolution.

As such, we utilized two 3'-barcoded mutant libraries of green fluorescent protein (*GFP*) that we had previously constructed<sup>21</sup> for examining the fitness effects of the single-nucleotide mutations (single mutants). The advantage of these libraries lies not only in their near-complete coverage of all single mutations across the CDS, but also in the fact that they maximize the percentage of molecules containing two single-nucleotide mutations (i.e., double mutants; see "Methods"), which is particularly suitable for measuring within-gene epistasis. In each library, the expression of *GFP* is driven by one of two promoters whose activity differs by -ten-fold ( $P_{AGP1}$  or  $P_{TDH3}$ )<sup>21</sup>, allowing for the assessment of how gene expression affects epistasis. The mutant libraries were bulk-transformed into *S. cerevisiae* BY4741 (MATa his3 $\Delta$ 1 leu2 $\Delta$ 0 met15 $\Delta$ 0 ura3 $\Delta$ 0) and integrated into the *HO* locus, and then subjected to PacBio-CCS to determine the correspondence between the barcodes and genotypes<sup>21</sup>. Together, we obtained unambiguous barcodes for 39238 single mutants and 31352 double mutants (Fig. 4A. See also Supplementary data 5).

Using each of these two strain pools of *GFP* mutants, we performed three replicates of competitive cocultures in Synthetic Complete (SC) media for seven days. The Wrightian fitness of any specific genotype relative to the wildtype is estimated by Hi-Seq targeting the genotype barcodes in samples before and after the competitive coculture ("This study" in Fig. 4A. See also Supplementary Fig. S7, Supplementary data 6 and "Methods"). In the two libraries driven by  $P_{AGP1}$  and  $P_{TDH3}$ , reliable fitness estimates were obtained for 2059 and 2180 single mutants, as well as 7631 and 14549 double mutants, among which 94.3% (7198 out of 7631) and 99.7% (14500 out of 14549) have both corresponding single mutants measured (Supplementary Fig. S7 and Supplementary data 7).

We then quantify the epistasis by focusing on one single mutation and all of its participating double mutants at a time. As an example, we extracted the fitness estimates for 12 double mutants containing the mutation C315T and one of other 12 single-nucleotide mutations whose fitness is also available. The observed fitness effect (selective coefficient  $s$ ) of each double mutant is then compared with its additive expectations derived from the individual effects of C315T and the other single-nucleotide mutation (Fig. 4B). The resulting differences (i.e., epistasis) were averaged for this focal single-nucleotide mutation (see "Methods"). Similarly, we individually evaluated all single-nucleotide mutations with participating double mutants and corresponding partner single mutants available, and found that the averaged epistasis is all negative (Fig. 4C "This study"). As a further confirmation that such pervasive epistasis is not limited to a specific



**Fig. 4 | Pervasive negative epistasis among single-nucleotide mutations within a gene.** **A** Schematic diagram of the experimental pipeline for measuring the within-gene epistasis between single-nucleotide mutations. More details are given in “Methods”. **B** The points represent double mutants that contain the C315T mutation of *GFP*. Observed selective coefficients relative to wildtype *GFP* are shown for these double mutants (*y* axis), which are compared with their corresponding additive expectations (*x* axis). The red diagonal line indicates absence of epistasis (i.e.,  $x = y$ ). Double mutants showing negative epistasis are indicated by down-pointing triangles, or by up-pointing triangles otherwise. **C** For each single-nucleotide mutation, the average epistasis (*y* axis) was calculated based on the

available double mutants that contain the focal mutation. Under various conditions in terms of gene, expression level, or medium (*x* axis), the average epistasis is predominantly negative. The red line indicates the absence of epistasis (i.e.,  $y = 0$ ). ( $n = 2094, 1823, 2245, 2255, 2094, 1823, 2245, 2255, 2245, 2255$ , from left to right). Box plots represent the distribution of average epistasis values: the center line indicates the median; the box bounds correspond to the 25th and 75th percentiles; whiskers extend to the minimum and maximum values; individual data points represent single-nucleotide mutations. Source data are provided as a Source data file.

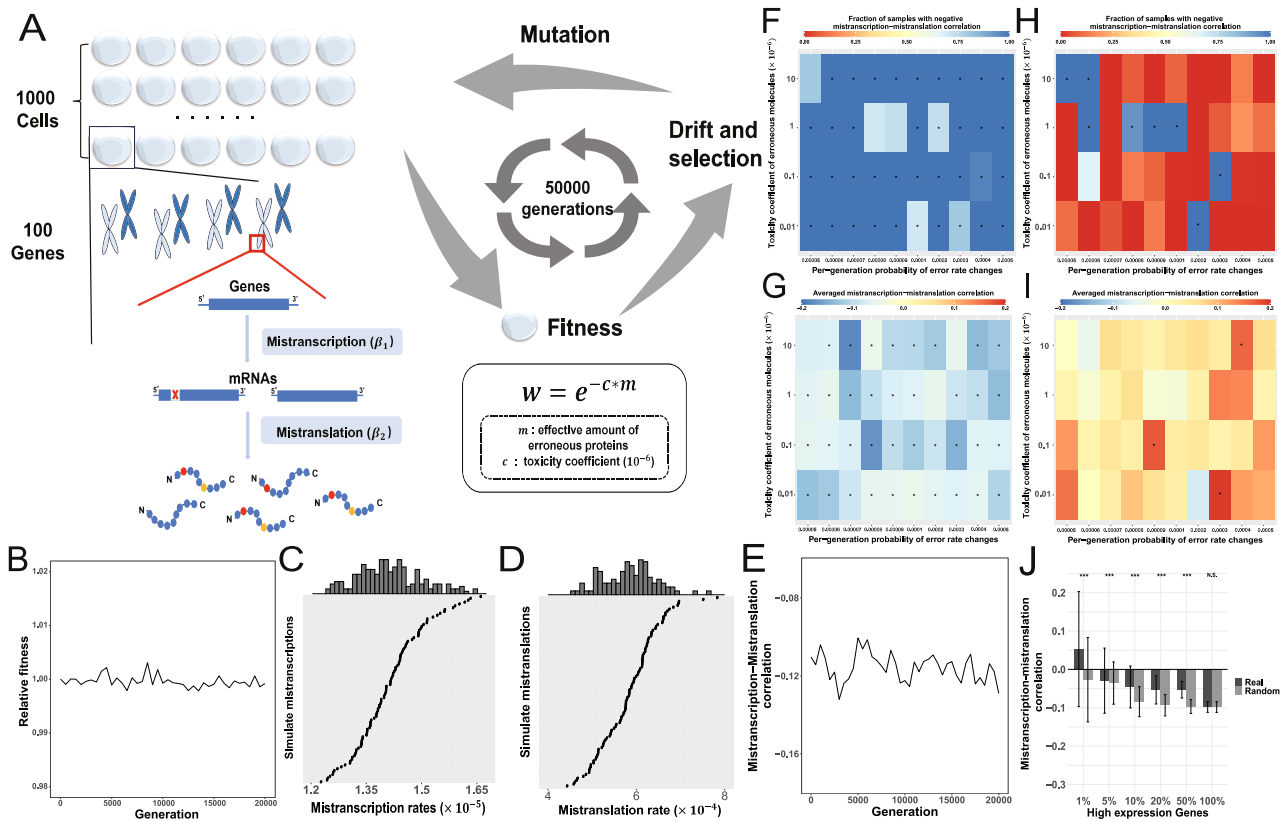
gene, expression level, or environment, we reanalyzed previously published data<sup>21</sup> on similar experiments under different conditions, and again found consistent patterns (Fig. 4C “Previous work”). These results were consistent with multiple previous reports<sup>45,46</sup>, which collectively confirmed the pervasive negative epistasis among point mutations.

### The negative epistasis between mistranscription and mistranslation suffices to generate their anti-correlation

To examine whether the measured epistasis suffices to generate the mistranscription-mistranslation trade-off, we conducted a simple calculation based on the assumption that the fitness effect of mistranscription/mistranslation is proportional to the fraction of erroneous molecules multiplied by the fitness effect of a 100% error (i.e., a genetic mutation). This assumption allows us to calculate the effect of negative epistasis between mistranscription/mistranslation by multiplying the effect of mutational epistasis by the product of mistranscription rates and mistranslation rates (see “Methods”). With the experimentally measured mutational epistasis (average  $-0.16$ ), as well as the per-gene probabilities of mistranscription ( $-0.30$ ) and

mistranslation ( $-0.81$ ), the negative epistasis between mistranscription and mistranslation would result in an average growth disadvantage of  $s = -0.02$  compared to the absence of epistasis (see “Methods”). This effect size is larger than the inverse of the effective population size of all model organisms examined here, providing initial support for the negative epistasis as a sufficient criterion for the mistranscription-mistranslation trade-off.

To further corroborate the role of negative epistasis in generating the mistranscription-mistranslation trade-off, we conducted a more detailed evolutionary simulation (Fig. 5A). Specifically, we simulated a population of 1000 cells, each expressing 100 genes with a fixed initial per-nucleotide mistranscription rate of  $10^{-5}$  and per-codon mistranslation rate of  $10^{-4}$ . Each gene is transcribed in 1000 nt mRNAs and translated into peptides of 300 amino acids long. The fraction of mistranscribed molecules and the fraction of mistranslated molecules for every gene were calculated based on these lengths, along with the total amount of erroneous proteins (see “Methods”). Note here that for simplicity and because synonymous mutations can also be deleterious due to their effects on gene expression<sup>47</sup>, elongation velocity or protein misfolding<sup>14,48,49</sup>, we ignored the differences between synonymous



**Fig. 5 | Evolutionary simulations support the role of negative epistasis in the mistranscription-mistranslation trade-off.** **A** Schematic diagram of the evolutionary simulation procedure. **B** The average fitness (y axis) of the population in different generations (x axis) along the simulation. Mistranscription (**C**) or mistranslation (**D**) rate (x axis) of simulated genes (y axis) in a typical cell in the final generation. **E** The average mistranscription-mistranslation correlation of the population (y axis) in different generations (x axis) along the simulation. **F, G** The key results are presented as heatmaps for simulations with various combinations of toxicity coefficient (y axis) and probability of error rate changes per generation (x axis). The fraction of individuals with a negative mistranscription-mistranslation correlation (**F**), and the averaged correlation coefficient of the simulated

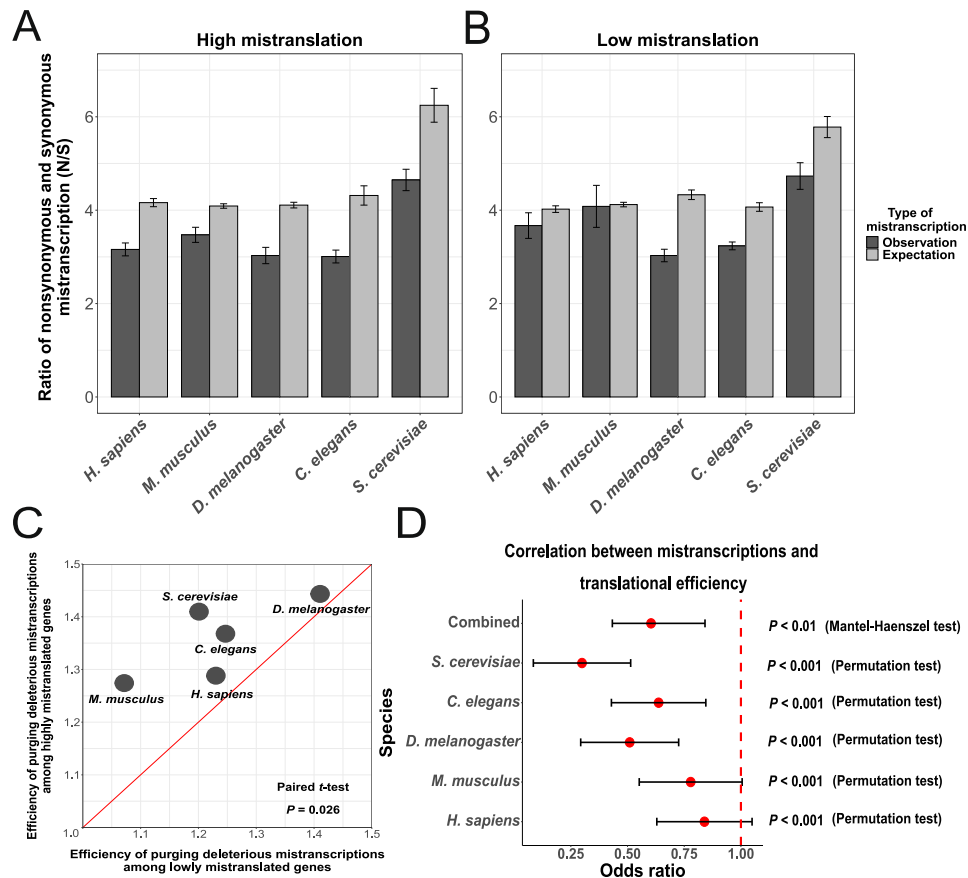
population (**G**) at the 20,000th generation are presented using the color scale on top of each panel, with an asterisk indicating statistical significance (Bonferroni-adjusted  $P < 0.05$ , Binomial test for **F** and Pearson's correlation test for **G**). **H–I** Similar to **F** and **G**, except the simulation is conducted without epistasis. **J** The average mistranscription-mistranslation correlation (y axis) among genes whose expression exceeds various thresholds (x axis, black bars), or among random genes of the same number (gray bars) are shown. The statistical significance by two-sided Wilcoxon signed-rank tests is indicated on top (\*:  $P < 0.05$ , \*\*:  $P < 0.01$ , \*\*\*:  $P < 0.001$ , N.S.: not significant). The bar and error bar represent, respectively, the mean and standard deviations assessed by bootstrapping the genes 1000 times. Source data are provided as a Source data file.

and nonsynonymous mistranscriptions. Following previous practice which assumed cellular toxicity of erroneous proteins<sup>11,12</sup>, the relative fitness  $w$  of a cell is then calculated as  $w = e^{-cm}$ , where  $c$  is a toxicity coefficient reflecting the average harm caused by each erroneous protein molecule, and  $m$  represents the effective amount of erroneous proteins, which is greater in the presence of negative epistasis than in the absence of epistasis (see “Methods”). This population was then subjected to sexual reproduction, mutations of mistranscription and mistranslation rates at a fraction of randomly selected sites by a predefined mutation rate, as well as drift and selection based on the relative fitness  $w$ . The simulation was initially run for a burn-in stage of 30,000 generations to reach mutation-selection-drift equilibrium, followed by a sampling stage of 20,000 generations. We sampled the population every 500 generations, and observed that it maintained a steady average fitness with  $< 1\%$  variation (Fig. 5B), a stable level of mistranscription rate at  $\sim 1.41 \times 10^{-5}$  (Fig. 5C), mistranslation rate at  $\sim 5.81 \times 10^{-4}$  (Fig. 5D), and the correlation between them at  $\sim -0.12$  (Fig. 5E). The overall pattern of mistranscription-mistranslation anticorrelation seems to be robust to changes in toxicity coefficient ( $c$ ) and the mutation rate of error rates within reasonable range (Fig. 5F, G). Most importantly, in parallel simulations without epistasis, the mistranscription-mistranslation anticorrelation disappeared (Fig. 5H, I), thus directly supporting the critical role of negative epistasis.

Collectively, we conclude that the negative epistasis, especially in genes that are lowly to intermediately expressed (Fig. 5J), is sufficient to generate the mistranscription-mistranslation trade-off.

### Functional genomic evidence for the mistranscription-suppression role of mistranslation

The hypothesized role of mistranslation in mistranscription suppression made three predictions that could be subject to empirical assessment. First, it predicts a negative correlation between rates of mistranscription and mistranslation, which has already been shown above. Second, it predicts that genes with higher mistranslation rates should show increased capacity of purging deleterious mistranscription when compared to those with lower mistranslation rates. We reasoned that because nonsynonymous mistranscriptions should be on average more deleterious than the synonymous ones, the relative strength of purging harmful mistranscriptions could then be approximated by the ratio between the nonsynonymous ( $N$ ) and synonymous ( $S$ ) mistranscriptions. Therefore, in each model organism, we calculated this ratio ( $N/S$ ) among genes with top 50% mistranslation rates (Fig. 6A, black bars). Compared with randomly simulated mistranscriptions (following rates in Fig. 1B) of identical numbers (Fig. 6A gray bars) from the same genome, the real mistranscriptions exhibit significantly lowered  $N/S$  ratios, suggesting that



**Fig. 6 | Functional genomic evidence for the mistranscription-suppression role of mistranslation.** **A, B** The observed (black bar,  $n = 15, 33, 13, 12, 11$ , from left to right) and expected (gray bar,  $n = 13, 22, 14, 19, 14$ , from left to right) ratios of nonsynonymous and synonymous mistranscriptions ( $N/S$ ) are presented. There was a significant reduction in  $N/S$  in real compared to simulated mistranscriptions in all five organisms (two-sided Wilcoxon signed-rank test, \*:  $P < 0.05$ , \*\*:  $P < 0.01$ , \*\*\*:  $P < 0.001$ ) among genes with high mistranslation (**A**) but less so among genes with low mistranslation (**B**). The bars and error bars represent, respectively, the mean and standard error of the ratios of  $N/S$  in each species. **C** For both highly mistranslated genes and lowly mistranslated genes, the efficiency of purging deleterious mistranscriptions was estimated by dividing the  $N/S$  of the simulated

mistranscription by the  $N/S$  of the real mistranscription. The median purging efficiency of highly mistranslated genes ( $y$  axis) tends to be higher than that of lowly mistranslated genes ( $x$  axis) ( $P = 0.026$ , paired  $t$ -test against equal purging efficiencies). The red diagonal line represents  $x = y$ . **D** Odds ratio for the correlation between translational efficiency and mistranscription rate was calculated (see “Methods”) for each of the five model organisms and assessed for statistical significance by the Permutation test ( $y$  axis, except the top row), and then combined and assessed for statistical significance by the Mantel-Haenszel test (top row). Error bars represent the standard deviation of the odds ratio assessed by 1000 bootstraps of the genes. Source data are provided as a Source data file.

a substantial fraction of deleterious mistranscriptions is eliminated. More importantly, the same analyses conducted for genes with low mistranslation rates (Fig. 6B) revealed a reduced difference in  $N/S$  ratios between the real and random mistranscriptions (Fig. 6C, paired  $t$ -test  $P = 0.026$ ), suggesting genes with lower mistranslation rates are indeed less capable of purging deleterious mistranscriptions.

A third prediction of our hypothesis concerns the relationship between a gene’s translational efficiency and its mistranscription rate. Consider a gene  $g_L$  that has a low translational efficiency, say, one protein per mRNA, and a gene  $g_H$  with a high translational efficiency, say, 100 proteins per mRNA. Each mistranscribed mRNA of  $g_L$  will therefore give rise to only one erroneous protein, whereas each mistranscribed mRNA of  $g_H$  will give rise to 100 erroneous proteins. For two reasons, highly translated  $g_H$  should have a lower mistranscription rate than the lowly translated  $g_L$ . First,  $g_H$  has a greater potential to exert functional effects (mistranscription suppression) of mistranslation due to its frequent translation, and also to the trade-off between translational efficiency and fidelity<sup>14,22</sup>. Second, since the phenotypic consequences likely depend on the amount, not fraction, of erroneous proteins<sup>20,21</sup>, reducing mistranscription for the highly translated  $g_H$  is obviously more beneficial than for  $g_L$ . Indeed, species-specific and

combined odds ratios suggest that genes with higher translational efficiency exhibit a lower mistranscription rate (Fig. 6D). In conclusion, functional genomic data support that mistranslation suppresses mistranscription.

## Discussion

To investigate the relationship between mistranslation and mistranscription, we analyzed mass spectrometry and Circ-Seq data to quantify gene-specific mistranslation and mistranscription rates in five eukaryotic model organisms. It is revealed that in all five species, genes with a high mistranslation rate tend to have a low mistranscription rate. We hypothesized that this observation could be explained by a mistranscription-suppression effect of mistranslation created by the negative epistasis between mistranslation and mistranscription. Via a deep-mutational scanning assay designed to probe within-gene epistasis between point mutations, we confirmed the prevalence of negative epistasis. More importantly, when taking into account the observed magnitude of negative epistasis, an *in silico* evolutionary simulation with realistic parameters of mistranslation and mistranscription rate stably recapitulates their inverse correlation, which is otherwise absent, suggesting that the observed negative epistasis

suffices to facilitate selection against mistranscription for genes with frequent mistranslation. Finally, we discovered from functional genomics data that genes with a high mistranslation rate exhibit increased efficiency in purging deleterious mistranscriptions, whereas genes with higher translational efficiency and therefore higher penetration of mistranscriptions tend to exhibit a decreased mistranscription rate. Our findings collectively demonstrated that mistranslation suppresses mistranscription, a previously unknown connection between transcriptional and translational fidelity.

Our study contains a few potential caveats that merit discussion. First, despite applying state-of-the-art methods (see “Methods”), the detectability and quantification accuracy of both mistranscription and mistranslation may not be very high, particularly for genes with low mRNA or protein abundances and peptides with low mass spectrometry detectability<sup>50</sup>. However, we found that the mistranscription and mistranslation profile, as quantified, is largely consistent with prior knowledge, suggesting that the biological signal has been efficiently extracted. For example, according to a recent study<sup>51</sup> examining the mistranscription rate in 21 species across the tree of life using Circ-Seq, within species, mistranscription rates of different genes are not correlated with expression levels, and across species, mistranscription rates are quite narrowly distributed. These observations are all recapitulated in our study. In addition, any noise in our measurements should have weakened rather than strengthened the mistranscription-mistranslation correlation, indicating that the correlation should be stronger than currently reported. Second, we cannot rule out the possibility that some identified mistranslations are the result of transcription errors. If that is common, a positive mistranscription-mistranslation correlation would be expected, further enhancing the significance of the observed negative correlation. Third, there may be alternative explanations for the mistranscription-mistranslation anticorrelation. For example, a transcriptional quality control mechanism might be activated when a large portion of a gene’s protein product is erroneous. Nevertheless, such a mechanism would require gene-specific regulation that links (or recognizes) both a protein and its corresponding mRNA, which are synthesized in separate parts of the cell, and is therefore highly unlikely. Fourth, the mistranscription-mistranslation anticorrelation may also be explained by an alternative, non-mutually exclusive hypothesis, namely that mistranscription suppresses mistranslation. However, we didn’t come up with predictions that exclusively support this alternative direction of causality, so we focused on the testable hypothesis. Fifth, our variant library used in the DMS assay contains only mutation pairs that are less than 50 bp apart (except for rare instances of template switching). This is a compromise on experimental feasibility due to the high number of mutation pairs in total. Nevertheless, there is no evidence that suggests that proximal mutation pairs are biased towards negative epistasis. The magnitude of negative epistasis we observed should therefore be generalizable to distant mutations. Finally, our evolutionary simulation does not explicitly model misfolded proteins or misfolding probability, which is believed to be more directly affected by natural selection than mistranscription or mistranslation<sup>52–54</sup> (but see ref. 12). This is a compromise for the computational complexity, since we are tracking a population of 1000 individuals, each expressing 100 genes, for up to 50,000 generations. We instead used a simplified approach, assuming a fixed deleterious effect per erroneous protein rather than per misfolded protein. The use of more computational resources or a more efficient simulation implementation is desirable in the future to explicitly model protein misfolding and investigate how the two types of errors contribute to misfolding.

Although the presented empirical evidence unanimously supports an evolutionary genetic explanation of the mistranscription-mistranslation anticorrelation, the proximal molecular mechanisms regulating the mistranscription and mistranslation rates of individual genes remain unexplored. In the case of mistranscription, two

mechanisms have been proposed, namely assembly of RNA polymerase with or without fidelity factors, or altering the speed of RNA synthesis<sup>30</sup>, albeit with limited empirical support. The codon usage bias<sup>22,55</sup> and modulation of translational elongation speed<sup>14</sup>, on the other hand, have found various support for their regulatory role in the mistranslation rate. Regardless of the specific molecular mechanisms, findings from this and other studies strongly suggest a local regulation of error rates for individual genes, as there is clear genomic variation in mistranscription and mistranslation rates, prioritizing fidelity in some genes over others<sup>19,24,28</sup>. The observed variation in gene-specific error rates also highlights potential use of phenotypic mutations to identify harmful genetic mutations<sup>56</sup>. It should be noted, however, that this does not preclude the global regulation of mistranscription and mistranslation rates.

Phenotypic mutations, such as mistranscription and mistranslation, have profound implications for evolution. On the one hand, they are deviations from the presumably optimal wildtype, thus causing a decrease in average fitness. On the other hand, they may also facilitate evolution by granting adaptive phenotypes before those phenotypes are fixed as genetic changes, such as observed in cancer evolution<sup>57</sup>. Based on our results, the transcription and translation level mutations interact in such a way that their combined error rates are constrained. This appears to be mainly the result of purifying selection that keeps fitness high. There is, however, another fascinating possibility that mistranslation is selectively maintained at a non-minimal level as a guard against excessive mistranscription, which could partly explain the observation that increasing ribosome proofreading fidelity reduces bacterial growth rates<sup>58</sup>. Another interesting topic would be how the mutation load-mitigation<sup>27,59</sup> and the mistranscription-suppression roles of mistranslation interplay.

The full ramifications of our discovery extend well beyond the evolutionary relationship between mistranslation and mistranscription and the underlying genetic mechanisms. As an example, phenotypic mutation is known to have broad medical implications, including the enhancement of antibiotic resistance in bacteria<sup>60</sup> and drug resistance in cancer<sup>24,61</sup>. There has long been a recognition that mistranslation can cause protein misfolding and, in turn, contribute to neurodegenerative disorders and aging<sup>11,17,18</sup>. Our results on the one hand suggest that mistranscription might also have a non-negligible contribution to this process, since it is NOT dwarfed by mistranslation by much on a per-gene basis (Fig. 2B). On the other hand, as the number of erroneous products by a mistranslation event is very limited as compared to that by a mistranscription event, some of the known phenotypic consequences that is potentially mistranscription-specific might therefore indirectly be regulated by mistranslation through its mistranscription-suppression effect. This might include, for example, the physiological process of cell fate switching<sup>28,29</sup>, or in pathological contexts such as proteotoxic stress and cellular lifespan<sup>5</sup>, amyloid precursor protein formation<sup>9</sup>, as well as mRNA heterogeneity during carcinogenesis<sup>62</sup>. It is anticipated that further studies into mistranscription and mistranslation, as well as their interactions, will enhance our understanding of how they influence overall health and provide new insights into diseases associated with protein aggregation.

## Methods

### Detection of transcription error and translation error

To detect transcription errors, we collected public Circ-Seq datasets from five model organisms (Supplementary data 1), and adapted a robust and sensitive computational pipeline previously established for detecting transcription errors from Circ-Seq data<sup>1,36,37</sup>. Please note that some Circ-Seq datasets from previous studies were based on mutated RNA polymerases, which were not used in this study. Briefly, using Tandem Repeats Finder (TRF)<sup>63</sup>, we scanned each adaptor-trimmed Circ-Seq raw read for tandem repeats whose repeat unit is at least 30

nts long and the repeated subsequences are at least 90% identical to each other. The 5'-end of the RNA template is identified by searching the TRF-reported consensus sequence of the repeat unit for the longest region that can be continuously mapped by Hisat2 (version 2.2.1) to the reference transcriptome. After reorganizing the consensus sequence to start at the 5'-end of the RNA template, it is mapped onto the genome using Hisat2. The mapped nucleotides were inspected and only those that met five criteria were retained: (1) they must be part of at least three repeats generated by the original RNA template; (2) all repeats must support the same nucleotide; (3) the sum of all quality scores for this nucleotide must exceed 100; (4) the distance from the focal nucleotide to either end of the consensus sequence must exceed 2 nts; (5) each nucleotide must be covered by 100 reads of which < 1% support base calls that differ from the reference genome. This last criterion eliminated polymorphic/heterozygous sites, as well as potential RNA editing events. We further discarded genes for which Circ-Seq reads only covered < 80% of their full-length transcript. There will be no change in our conclusion if we change the cutoff from 80% to 50% or 90% (Supplementary Fig. S8). Finally, the transcription error rate is calculated by dividing the number of mismatches by the total number of nucleotides that meet all quality criteria.

To detect translation errors in the form of amino acid misincorporations, we collected public MS datasets from the PRoteomics Identification Database (PRIDE) for the same five model organisms as above (Supplementary data 3). These datasets were all generated in data-dependent acquisition mode (DDA) and contained label-free quantifications. By using MaxQuant(v2.4.14.0)<sup>64,65</sup>, we analyzed the datasets following a previously published approach<sup>31</sup> based on the fact that misincorporation of amino acids can lead to specific mass shifts in peptides, while strictly filtering false calls due to polymorphisms or post-translational modifications (PTMs). Briefly, a standard database search (allowing 1% FDR) using MaxQuant with previously described<sup>31</sup> parameters yields “identified spectra” (peptides that have sequence and abundance information available) and “unidentified spectra”. To identify pairs of base peptides (error-free, from the identified spectra) and dependent peptides (error-containing, from the unidentified spectra), we used the identified spectra as a reference spectrum and performed a one-to-one comparison with the unidentified spectra. We applied the following steps to each pair of identified and unidentified spectra: (i) Calculate the mass shift ( $\Delta m$ ) between the identified and the unidentified spectra. (ii) Approximate theoretical spectra for all possible amino acid substitutions by an in silico simulation applied to the peptide sequence of the identified spectra, whose resulting mass shifts ( $\Delta m'$ ) are all recorded. (iii) Retain all substitutions whose  $\Delta m'$  is consistent (< 0.005 Da difference) with  $\Delta m$  as candidates. (iv) Compare candidates' theoretical and experimental spectra using a formula based on Andromeda's binomial score, and retain the one candidate with the best score. (v) The mass shift derived from the experimental spectra must have a MaxQuant-estimated localization probability  $\geq 0.95$  at the misincorporated site. A candidate that passes all five steps is considered a dependent peptide, while the corresponding misincorporation-free peptide from the identified spectra is deemed a base peptide.

The candidate pairs of base and dependent peptides were further filtered to eliminate various artifacts. First, we excluded any pair that could be explained by any PTM recorded in UNIMOD by the following criteria: (i) The  $\Delta m$  of a dependent peptide is consistent with the PTM's mass shift by < 0.01 Da difference. (ii) The known position of the PTM is consistent with the dependent peptide with a localization probability  $\geq 0.05$ . The PTM criteria are intentionally made less stringent than the criteria for translational misincorporations, so that we are reporting a highly confident set of translational misincorporation events. Second, we aimed to eliminate misincorporations that could be explained by RNA editing. Specifically, we retrieved a list of known RNA editing events from the REDportal database<sup>66</sup>, and extracted nonsynonymous

events that would result in amino acid substitutions. Any pair corresponding to an amino acid substitution that could be explained by these RNA editing events should be discarded. We found that none of the candidate pairs in human or mouse samples could be due to nonsynonymous RNA editing. Third, we removed misincorporations caused by known polymorphisms. Specifically, we downloaded genomic variant data from the Ensembl (*H. sapiens* GRCh38, *M. musculus* GRCm39, *D. melanogaster* genomeBDGP6.46, *C. elegans* WBcel235, *S. cerevisiae* R64-1-1, Ensembl release 113) and extracted missense variants causing amino acid changes. We removed candidate pairs of base and dependent peptides with the same type of error as these missense variants. Finally, the translation error rate is calculated by dividing the intensities of the mass spectral signals of the dependent peptides by the sum of the intensities of the base peptides and dependent peptides. Misincorporations with a rate  $> 10^{-2}$  are excluded, as they are unlikely real, given the misincorporation rate is estimated to be around  $10^{-3}$ .

### Extrapolating the probability of errors per gene by the per-site error rate and the gene length

According to the above per-site rates of error, we estimated the probability of errors per gene, which is defined as the probability of a full-length product (RNA or protein) of the gene containing at least one error of a specific type. For this purpose, we first calculated, for each gene, the average per-site mistranscription rate for all nucleotides with non-zero coverage in Circ-Seq. The per-site mistranslation rate is averaged only across codons with non-zero mistranslation rates, not flanking codons that are covered by MS but do not mistranslate. This results in an overestimation of the average mistranslation rate per codon since many codons with no mistranslation detected are not considered in the averaging process. Therefore, the list of genes exhibiting higher rates of mistranscription than mistranslation is likely to be conservative. Denoting the average per-site error rate as  $\beta$ , and the number of sites in the full coding sequence as  $L$  (i.e., length in units of nucleotides for RNA CDS and amino acids for protein), the per-gene error probability is then calculated as  $1 - (1 - \beta)^L$ .

### Accounting for expression-dependent detection threshold for errors

We aimed to estimate the correlation using all genes covered in both the Circ-Seq and MS datasets, regardless of whether any errors were detected. This effort is complicated by two factors. First, error rates below the error detection limits will appear as zero, similar to censoring of very low error rates. Second, the error detection limit is dependent on the gene expression level, as relatively low error rates are readily detectable for highly expressed genes, while only higher error rates are detectable for genes with a lower expression level. These two issues were taken into account using a bivariate Tobit model, which is composed of two analytical stages. In the first stage, we modeled the error rate (a latent variable whose real value is unobservable for some gene) by adjusting for the detection threshold via a Tobit regression, where the censoring point (detection threshold) is dependent on the gene expression (protein expression for mistranslation or RNA expression for mistranscription). In the second stage, the error rates inferred by the Tobit regression from the first stage are used as instruments to estimate the true correlation or odds ratio (Supplementary Fig. S5A, B).

### The relationship between mutational epistasis and epistasis between two errors

As transcription and translation errors are very rare, it is not possible to measure the epistasis between the two types of errors directly. We therefore take a step back to experimentally measure the epistasis between residue substitutions at the DNA level, which is, in theory, equal to the epistasis between a mistranscription occurring at a rate of

100% and a mistranslation occurring at a rate of 100%. By scaling the measured epistasis by the probability of mistranscription and mistranslation, we can then estimate the magnitude of the between-error epistasis.

More specifically, if we define the Wrightian fitness of a mutant relative to the wildtype as  $w$ , and the corresponding selective coefficient as  $s = w - 1$ , we can estimate the epistasis between two single-nucleotide mutations in a gene as

$$\varepsilon_{AB} = s_{AB} - (s_A + s_B) \tag{1}$$

by assuming that the effects of independent mutations are additive. Here,  $s_A$  and  $s_B$  are the selective coefficients of the single-nucleotide mutation A and B, respectively, so  $(s_A + s_B)$  is the expected effect when both A and B occur and have independent effects, and  $s_{AB}$  is the observed selective coefficient when both A and B occur. In the case of mistranscription and mistranslation, if mistranscription A occurs at a rate of  $\beta_A$  and mistranslation B occurs at a rate of  $\beta_B$ , we can express the total fitness effect in three parts. The first part is the effect of mistranscription as  $s_A\beta_A(1 - \beta_B)$ , where  $\beta_A(1 - \beta_B)$  is the fraction of molecules with only mistranscription A but no mistranslation B. Similarly, the second part, which represents the effect of mistranslation, is  $s_B\beta_B(1 - \beta_A)$ . Lastly, the third part representing the effect of molecules containing both types of errors is  $s_{AB}\beta_A\beta_B$ . In sum, the total fitness effect of mistranscription and mistranslation can then be expressed as

$$s = s_A\beta_A(1 - \beta_B) + s_B\beta_B(1 - \beta_A) + s_{AB}\beta_A\beta_B = s_A\beta_A + s_B\beta_B + \varepsilon_{AB}\beta_A\beta_B \tag{2}$$

If there is no epistasis, so that A and B just independently affect fitness (i.e.  $\varepsilon_{AB} = 0$ ), the independently combined fitness effect of mistranscription and mistranslation can be expressed as

$$s' = s_A\beta_A + s_B\beta_B \tag{3}$$

Then the net contribution of epistasis between mistranscription and mistranslation is simply

$$s - s' = \varepsilon_{AB}\beta_A\beta_B \tag{4}$$

By this logic, we can transform the question of epistasis between transcription and translation errors into a question of epistasis between two genetic mutations (i.e.  $\varepsilon_{AB}$ ).

Note that alternatively assuming a log-additive model, or multiplicative model, will not qualitatively alter our conclusion, because that is equivalent to replacing all the “s” ( $s_A, s_B$  and  $s_{AB}$  included) in the equation above by their corresponding  $\log(w)$ . More specifically, the Eqs. (1–4) above will be transformed as follows. Equation (1) for the mutational epistasis becomes

$$\varepsilon_{AB} = w_{AB} / (w_A w_B) \tag{5}$$

Equation (2) for the total fitness effect of mistranscription and mistranslation becomes

$$\begin{aligned} \log(w) &= \log(w_A)\beta_A(1 - \beta_B) + \log(w_B)\beta_B(1 - \beta_A) + \log(w_{AB})\beta_A\beta_B \\ &= \log(w_A)\beta_A + \log(w_B)\beta_B + \log(\varepsilon_{AB})\beta_A\beta_B \end{aligned} \tag{6}$$

Equation (3) for the independently combined fitness effect of mistranscription and mistranslation becomes

$$\log(w') = \log(w_A)\beta_A + \log(w_B)\beta_B \tag{7}$$

Equation (4) for the net contribution of epistasis between mistranscription and mistranslation becomes

$$\log(w) - \log(w') = \log(\varepsilon_{AB})\beta_A\beta_B \tag{8}$$

### Experimental assessment of epistasis

To measure the mutational epistasis  $\varepsilon_{AB}$  by experiment, we utilized a mutant library of green fluorescent protein (*GFP*) that we previously<sup>21</sup> constructed by doped oligonucleotide synthesis and fusion PCR. Specifically, the full 733-bp CDS of *GFP* is divided into 15 non-overlapping 50-bp subsequences (the last one is 33 bp long). For each subsequence, a 90-bp doped oligonucleotides centered on the focal 50-bp were synthesized (<https://www.idtdna.com/>), with the center 50-bp variable sites each synthesized by mixing 97% wildtype nucleotide with 1% each of the other three nucleotides, while the flanking 20-bp on both ends are invariable sites, same as the corresponding positions in wildtype *GFP*. Despite the fact that mutations in double mutants were limited to  $\leq 50$  bp apart, this design has the unique advantage of maximizing the fraction of double mutants in the library<sup>67</sup>, which in theory equals to  $0.97^{48} \times 0.03^2 \times C_{50}^2 = 25.55\%$ . Using the synthesized oligonucleotides individually as forward mutant primers, and a fixed reverse primer, we amplified 15 “PCR fragment 1” each starting at 20 bp upstream of the focal 50-bp subsequences and ending at the *GFP* terminator. Then another “PCR fragment 2”, amplified from a selected promoter ( $P_{AGP1}$  or  $P_{TDH3}$ , whose activities differ by  $\sim$  ten-fold<sup>21</sup>) to the immediate upstream region of the focal 50-bp subsequences, was fused with the corresponding PCR fragment 1 using a promoter-specific primer and a terminator primer containing a 20-bp region of fully degenerated (randomized) nucleotides serving as a barcode. The final mutant libraries were bulk-transformed<sup>21</sup> into *S. cerevisiae* BY4741 (MATa his3 $\Delta$ 1 leu2 $\Delta$ 0 met15 $\Delta$ 0 ura3 $\Delta$ 0) and integrated into the HO locus, and then subjected to PacBio-CCS to determine the correspondence between the barcodes and genotypes following previous pipelines<sup>21,68</sup>. The PacBio results show that the wild-type strains have been  $\sim$ two-fold enriched in the final strain library ( $\sim 45\%$  compared to the expected  $0.97^{50} = 21.81\%$ ), so that both single ( $\sim 20.85\%$ ) and double ( $\sim 16.65\%$ ) mutants have reduced relative to their expected fractions (Supplementary data 5).

In this study, we performed three replicates of competitive cocultures in Synthetic Complete (SC) media for seven days using the above strain pool of *GFP* mutants. In each competition, samples were inoculated into 700 ml of SC medium and then cultivated at 250 r.p.m. and 30 °C. Every eight hours during the competition, we transferred  $\sim 1.5 \times 10^8$  cells ( $> 15 \times$  the effective population size of yeast) to fresh medium, and estimated the concentration of cultured cells by OD660 readings to ensure that the population remained in the exponential phase of growth. Based on a comparison of the pre- and post-competition variant frequency as measured by high-throughput sequencing of the barcodes, the relative fitness of variants with known barcodes is determined. Specifically, we PE150-sequenced the barcode region, retained all barcodes that appeared as 100% consistent from both ends, and mapped all retained barcodes to genotypes (Supplementary data 6). To ensure accurate measurement of wildtype, we calculated the frequency changes of individual wildtype barcodes, or d7/d0 ratios, using wildtype barcodes with at least 100 reads on day 0 and at least five reads on day 7. Wildtype barcodes whose d7/d0 ratios deviated more than one standard error from the average d7/d0 ratio were discarded. To determine the relative fitness of mutants, we combined the reads for all barcodes belonging to the same genotype, and also required that the total number of reads for a variant be at least 100 at day 0 in order to guarantee accuracy and

resolution. The Wrightian fitness  $w$  of a specific genotype relative to the wildtype is then calculated as:

$$w = \left( \frac{f_7/f_0}{F_7/F_0} \right)^{\frac{1}{g}} \quad (9)$$

Here,  $f_0$  and  $f_7$  is the frequency (number of reads) of the genotype on day 0 and 7 respectively;  $F_0$  and  $F_7$  are the  $f_0$  and  $f_7$  of the wildtype;  $g$  is the number of generations of cell growth during the competitive coculture as estimated from the OD660 obtained at each passage according to  $g = \log_2(\text{OD660} \times 10^7 \times \nu/n)$ , where  $\nu$  is the culture volume in liters (0.7), and  $n$  is the number of cells transferred at the last passage. The genotypes with fitness values less than 0.5 were discarded. Genotype frequencies from all replicates were pooled together, which essentially gives a weighted average of the replicate-separated fitness estimates, with the weight being the confidence level of the fitness estimate for the wildtype<sup>21</sup>.

Using the abovementioned Eq. (1), we compared the relative fitness of double mutants and corresponding single mutants to estimate the fitness effects of mutational epistasis ( $\varepsilon_{AB}$ , Figure 4c, y axis to the left). To extrapolate the fitness effect of epistasis between mistranscription and mistranslation based on the abovementioned Eq. (2), we picked one double mutant (which gives  $s_{AB}$  and therefore  $\varepsilon_{AB}$ ) and its two corresponding single mutants (which give  $s_A$  and  $s_B$ ), as well as a per-gene mistranscription probability of 0.3 (i.e.,  $\beta_A$ ) and a per-gene mistranslation probability of 0.8 ( $\beta_B$ ) (from Fig. 2B). Note that although Eq. (2) was introduced for per-site error rates, we are now using it for per-gene error rates. The justification for this is not only that they are mathematically equivalent, but also because epistasis may occur not only between two specific mutations/errors, but also between any two mutations/errors within the gene. The net fitness effect of epistasis between errors is then estimated by contrasting the  $s$  calculated by Eq. (2) with the real  $\varepsilon_{AB}$  of the picked double mutant, with that calculated with  $\varepsilon_{AB} = 0$  (absence of epistasis). We similarly extrapolated the fitness effect of epistasis between mistranscription and mistranslation using all available double mutants, and thereby assessed the resulting distribution (Fig. 4c, y axis to the right). To examine whether the pattern is different for other genes or in other environments, we also extracted the relative fitness of double mutants from our previous similar experiments<sup>21</sup> conducted for a conditionally essential gene *URA3*, and in other environments, such as in YPD (a rich medium) and SC-URA (in which *URA3* becomes essential) media (Supplementary data 6).

### Evolutionary simulation

A population of 1000 cells, each containing 100 genes whose expression follow a power-law distribution<sup>69</sup> was simulated. The 100 genes were evenly distributed on four independently assorting chromosome pairs of 100Kbp per chromosome, whose recombination rate is set at  $10^{-8}$  per generation per gamete (corresponding to 1 cM/Mbp). Our simulation considers selection (fitness), mutation and drift. (i) For the fitness, let us denote a gene's transcript length as  $L_N$  nucleotides, and the cell's per-nucleotide mistranscription rate as  $\beta_N$ , then the probability that a transcript contains at least one transcription error, is  $E_N = 1 - \prod_{i=1}^{L_N} (1 - \beta_{N,i})$ . Similarly, given that the length of the gene's protein product is  $L_P$  amino acids, and per-codon mistranslation rate is  $\beta_P$ , then the probability that a protein contains at least one translation error, is  $E_P = 1 - \prod_{i=1}^{L_P} (1 - \beta_{P,i})$ . We used realistic parameters of fixed lengths  $L_N = 1000$  and  $L_P = 300$  (shorter than 1000/3 to reflect UTRs), and variable (subject to evolution) error rates initially set as  $\beta_N = 10^{-5}$  and  $\beta_P = 10^{-4}$  for every site in every gene. Following previous practice that assumed cellular toxicity of erroneous proteins<sup>11,12</sup>, the relative fitness  $w$  of a cell is then calculated based on the total amount of erroneous (containing at least one mistranscription and/or

mistranslation) proteins  $m$  as  $w = e^{-cm}$ . Here,  $c$ , set as  $10^{-6}$ , is a constant reflecting the toxicity per erroneous protein molecule. Further,  $m = \sum_{g=1}^{500} (M_g E_{N,g} (1 - E_{P,g}) + M_g E_{P,g} (1 - E_{N,g}) + M_g \varepsilon E_{N,g} E_{P,g})$ , where  $\varepsilon$  reflects epistasis (see below), and  $g$  represents a gene that is expressed at a level of  $M_g$  protein molecules and that has a per-gene probability of mistranscription at  $E_{N,g}$  and a per-gene probability of mistranslation at  $E_{P,g}$ . A minimum requirement for the fraction of correct (error-free) molecules is applied so that  $(1 - E_{P,g}) \times (1 - E_{N,g}) > 0.5$  or the cell is considered dead. (ii) For the mutation, with a predetermined mutation rate ( $x$  axis in Fig. 5F–I), random changes are applied to the error rates ( $\beta_N$  and  $\beta_P$ ) of a fraction of sites in each cell. The factors of random changes are drawn from gamma distribution created by fitting the empirically observed error rates of the corresponding type and scaled to a mean of 1 (i.e., relative change). The fitness of each cell is then recalculated with  $E_N$  and  $E_P$  updated by the mutated error rates. (iii) For the drift, we employed a roulette wheel algorithm to randomly sample (with replacement) 1000 cells from the current population to create the next generation, using the relative fitness  $w$  as the relative probability of each cell being sampled. This whole procedure of fitness calculation, mutation, and reproduction was repeated for an initial burn-in stage of 30,000 generations to allow the population to reach mutation-selection-drift equilibrium, followed by a sampling stage of 20,000 generations (Fig. 5A).

To illustrate the effect of negative epistasis between mistranscription and mistranslation, we constructed two versions of the model by changing the value of  $\varepsilon$  in the aforementioned definition of  $m$ . More specifically, the “independent model”, which assumes no epistasis between the two types of errors, sets  $\varepsilon$  as 1. The other “epistatic model” sets  $\varepsilon$  as 100, which, compared to the independent model, reduces the fitness of the initial genotype by 2.7% (or the genotype at sampling stage by 0.8%), a difference consistent with the empirical observation (-2%) of the effect of negative epistasis among mutations in our experiment.

### Simulated mistranscriptions

To assess the expected ratios of nonsynonymous (N) to synonymous (S) mistranscriptions (Fig. 6A, B), we first extracted the protein coding genes with detected mistranscriptions. For each mistranscribed nucleotide of these genes, we randomly generated a simulated mistranscription event whose type of base substitution is sampled from the genome-wide frequencies of the twelve distinct base substitution types (in empirically observed mistranscriptions). We then calculated the expected N/S ratio for each gene by determining whether the simulated mistranscription events were synonymous or not.

### Translational efficiency and mRNA abundance

We obtained Ribo-seq and RNA-seq raw data from Gene Expression Omnibus under accession numbers GSE280878 (*H. sapiens*), GSE263991 (*M. musculus*), GSE246393 (*D. Melanogaster*), GSE280071 (*C. elegans*), and GSE171392 (*S. cerevisiae*). These datasets were chosen based on (i) their preprocessed translational efficiency data give the maximum number of genes with mistranscription or mistranslation, and (ii) the most recent dataset as of Nov. 2024. The raw reads were processed by TrimGalore v0.6.10 (<https://github.com/FelixKrueger/TrimGalore>) for trimming the sequencing adaptors, as well as to remove low quality bases (-q 25 -stringency 3) and reads too short ( $\leq 36$ nt) for meaningful alignment. Those reads alignable to non-coding transcripts by Hisat2 (version 2.2.1)<sup>70</sup> were excluded. The remaining reads were aligned to the reference transcriptomes by Hisat2, and the low-quality alignments (MAPQ < 30) were removed. For RNA-seq and Ribo-seq data, we use cufflinks<sup>71</sup> (version 2.2.1) to calculate FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) in per gene. Translational efficiency (TE) is defined as the ratio of FPKM at the translational level to FPKM at the transcriptional level.

### Odds ratios and Mantel–Haenszel test

We evaluated the relationship between mistranslation rate and tRNA abundance by combined odds ratios (OR) calculated by Mantel–Haenszel test. Take the human dataset as an example, for each type of original codon, we constructed a  $2 \times 2$  contingency table by classifying each type of mistranslated amino acids (i.e., a tile corresponding to mistranslation in Supplementary Fig. S2) into one of four categories based on (i) whether its mistranslation rate is greater than the median mistranslation rate of this original codon and (ii) whether its tRNA genomic copy number (summing up all tRNAs encoding this amino acid) is greater than the median tRNA copy number among all amino acids. Let the number of mistranslated amino acids that fall in the four categories be: *a* (yes to both questions), *b* (yes to only question i), *c* (yes to only question ii), and *d* (no to both questions), respectively.  $OR = (ad)/(bc)$ . Thus,  $OR > 1$  if mistranslation is more frequent for amino acids with higher genomic copy number. The Mantel–Haenszel test was then conducted by combining the 61 contingency tables from the 61 original codons by the “mantelhaen.test” function in R package “stats”. This method is applied to all five species in assessing whether genes with higher mistranslation rate tend to have lower mistranscription rate (Fig. 3F and Supplementary Fig. S5), whether synonymous codons with higher tRNA availability tend to have lower mistranslation rate (Supplementary Fig. S2), and whether genes with higher translational efficiency tend to have lower mistranscription rate (Fig. 6D).

### Reporting summary

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

### Data availability

The data supporting the findings of this study are available from the corresponding authors upon request. All raw data from high-throughput sequencing generated in this study have been deposited in the NCBI BioProjects under accession number PRJNA1347074. Source data for the figures and Supplementary Figs. are provided as a Source Data file. Source data are provided with this paper.

### Code availability

Custom R scripts were used in data analysis and are available on GitHub ([https://github.com/zhangxy77/Mistranscription\\_mistranslation](https://github.com/zhangxy77/Mistranscription_mistranslation)).

### References

- Gout, J. F. et al. The landscape of transcription errors in eukaryotic cells. *Sci. Adv.* **3**, e1701484 (2017).
- Springgate, C. F. & Loeb, L. A. On the fidelity of transcription by *Escherichia coli* ribonucleic acid polymerase. *J. Mol. Biol.* **97**, 577–591 (1975).
- Zaher, H. S. & Green, R. Fidelity at the molecular level: lessons from protein synthesis. *Cell* **136**, 746–762 (2009).
- Saxowsky, T. T., Meadows, K. L., Klungland, A. & Doetsch, P. W. 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc. Natl. Acad. Sci. USA* **105**, 18877–18882 (2008).
- Vermulst, M. et al. Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nat. Commun.* **6**, 8065 (2015).
- Anagnostou, M. et al. Transcription errors in aging and disease. *Transl. Med. Aging* **5**, 31–38 (2021).
- Chung, C. S. et al. Transcript errors generate amyloid-like proteins in human cells. *Nat. Commun.* **15**, 8676 (2024).
- James, K., Gamba, P., Cockell, S. J. & Zenkin, N. Misincorporation by RNA polymerase is a major source of transcription pausing in vivo. *Nucleic Acids Res.* **45**, 1105–1113 (2017).
- van Leeuwen, F. W. et al. Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer’s and Down patients. *Science* **279**, 242–247 (1998).
- van Leeuwen, F. W., Burbach, J. P. & Hol, E. M. Mutations in RNA: a first example of molecular misreading in Alzheimer’s disease. *Trends Neurosci.* **21**, 331–335 (1998).
- Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
- 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).
- 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).
- Yang, J. R., Chen, X. & Zhang, J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.* **12**, e1001910 (2014).
- Lee, J. W. et al. Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature* **443**, 50–55 (2006).
- Kapur, M. & Ackerman, S. L. mRNA translation gone awry: translation fidelity and neurological disease. *Trends Genet.* **34**, 218–231 (2018).
- Lindner, A. B. & Demarez, A. Protein aggregation as a paradigm of aging. *Biochim Biophys. Acta* **1790**, 980–996 (2009).
- Orgel, L. E. The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc. Natl. Acad. Sci. USA* **49**, 517–521 (1963).
- Zheng, B. et al. Translational fidelity and longevity are genetically linked. *Nat. Commun.* **16**, 7521 (2025).
- Zhang, J. & Yang, J. R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).
- Wu, Z. et al. Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nat. Ecol. Evol.* **6**, 103–115 (2022).
- Wu, X., Xu, M., Yang, J. R. & Lu, J. Genome-wide impact of codon usage bias on translation optimization in *Drosophila melanogaster*. *Nat. Commun.* **15**, 8329 (2024).
- Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci. USA* **108**, E67–E76 (2011).
- Sharma, S. V. et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).
- Whitehead, D. J., Wilke, C. O., Vernazobres, D. & Bornberg-Bauer, E. The look-ahead effect of phenotypic mutations. *Biol. Direct* **3**, 18 (2008).
- van Dijk, D. et al. Slow-growing cells within isogenic populations have increased RNA polymerase error rates and DNA damage. *Nat. Commun.* **6**, 7972 (2015).
- Zheng, J., Guo, N. & Wagner, A. Mistranslation reduces mutation load in evolving proteins through negative epistasis with DNA mutations. *Mol. Biol. Evol.* **38**, 4792–4804 (2021).
- Gordon, A. J. et al. Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. *PLoS Biol.* **7**, e44 (2009).
- Satory, D. et al. DksA involvement in transcription fidelity buffers stochastic epigenetic change. *Nucleic Acids Res.* **43**, 10190–10199 (2015).
- Chung, C. et al. The fidelity of transcription in human cells. *Proc. Natl. Acad. Sci. USA* **120**, e2210038120 (2023).
- Mordret, E. et al. Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. *Mol. Cell* **75**, 427–441.e5 (2019).
- Wohlgemuth, I. et al. Translation error clusters induced by aminoglycoside antibiotics. *Nat. Commun.* **12**, 1830 (2021).

33. 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).
34. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
35. Acevedo, A. & Andino, R. Library preparation for highly accurate population sequencing of RNA viruses. *Nat. Protoc.* **9**, 1760–1769 (2014).
36. Chung, C. et al. Evolutionary conservation of the fidelity of transcription. *Nat. Commun.* **14**, 1547 (2023).
37. Fritsch, C., Gout, J. P. & Vermulst, M. Genome-wide surveillance of transcription errors in eukaryotic organisms. *J. Vis. Exp.* **13**, 57731 (2018).
38. Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13**, 87–96 (2007).
39. Kramer, E. B., Vallabhaneni, H., Mayer, L. M. & Farabaugh, P. J. A comprehensive analysis of translational missense errors in the yeast *Saccharomyces cerevisiae*. *RNA* **16**, 1797–1808 (2010).
40. Fritsch, C. et al. Genome-wide surveillance of transcription errors in response to genotoxic stress. *Proc. Natl. Acad. Sci. USA* **118**, e2004077118 (2021).
41. Gasca, J. et al. EDIL3 Promotes epithelial–mesenchymal transition and paclitaxel resistance through its interaction with integrin AVβ3 in cancer cells. *Cell Death Discov.* **6**, 86 (2020).
42. Ke, B. et al. EDIL3 is a potential prognostic biomarker that correlates with immune infiltrates in gastric cancer. *PeerJ* **11**, e15559 (2023).
43. Mamoor, S. A single nucleotide variant on chromosome 5 residing within EDIL3 distinguishes patients with basal-like human breast cancer. (2022).
44. Wang, J. et al. Prognostic potential of secreted modular calcium-binding protein 1 in low-grade glioma. *Front. Mol. Biosci.* **8**, 666623 (2021).
45. Gonzalez, C. E. & Ostermeier, M. Pervasive pairwise intragenic epistasis among sequential mutations in TEM-1 β-lactamase. *J. Mol. Biol.* **431**, 1981–1992 (2019).
46. Schenk, M., Szendro, I. G., Salverda, M. L. M., Krug, J. & Visser, J. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Mol. Biol. Evol.* **30**, 1779–1787 (2013).
47. Shen, X., Song, S., Li, C. & Zhang, J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature* **606**, 725–731 (2022).
48. O'Brien, E. P., Vendruscolo, M. & Dobson, C. M. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nat. Commun.* **5**, 2988 (2014).
49. Yang, J. R. Does mRNA structure contain genetic information for regulating co-translational protein folding? *Zool. Res* **38**, 36–43 (2017).
50. Wacholder, A. & Carvunis, A. R. Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biol.* **21**, e3002409 (2023).
51. Li, W. et al. A narrow range of transcript-error rates across the Tree of Life. *Sci. Adv.* **11**, eadv9898 (2025).
52. Karve, S., Dasmeh, P., Zheng, J. & Wagner, A. Low protein expression enhances phenotypic evolvability by intensifying selection on folding stability. *Nat. Ecol. Evol.* **6**, 1155–1164 (2022).
53. Dasmeh, P., Serohijos, A. W., Kepp, K. P. & Shakhnovich, E. I. The influence of selection for protein stability on dN/dS estimations. *Genome Biol. Evol.* **6**, 2956–2967 (2014).
54. Serohijos, A. W., Rimas, Z. & Shakhnovich, E. I. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* **2**, 249–256 (2012).
55. Akashi, H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935 (1994).
56. Tretyachenko, V. et al. Encoded and non-genetic alternative protein variants expand human functional proteome. *bioRxiv*, <https://doi.org/10.1101/2025.02.17.638604> (2025).
57. Frank, S. A. & Yanai, I. The origin of novel traits in cancer. *Trends Cancer* **10**, 880–892 (2024).
58. Kurland, C. G. Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* **26**, 29–50 (1992).
59. Zheng, J., Guo, N. & Wagner, A. Selection enhances protein evolvability by increasing mutational robustness and foldability. *Science* **370**, eabb5962 (2020).
60. Rego, E. H., Audette, R. E. & Rubin, E. J. Deletion of a mycobacterial divisome factor collapses single-cell phenotypic heterogeneity. *Nature* **546**, 153–157 (2017).
61. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
62. Brulliard, M. et al. Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. *Proc. Natl. Acad. Sci. USA* **104**, 7522–7527 (2007).
63. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
64. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
65. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
66. Mansi, L. et al. REDportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. *Nucleic Acids Res.* **49**, D1012–D1019 (2021).
67. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
68. Guo, Z. et al. Assessment of the reversibility of resistance in the absence of antibiotics and its relationship with the resistance gene's fitness cost: a genetic study with *mcr-1*. *Lancet Microbe* **5**, 100846 (2024).
69. Awazu, A., Tanabe, T., Kamitani, M., Tezuka, A. & Nagano, A. J. Broad distribution spectrum from Gaussian to power law appears in stochastic variations in RNA-seq data. *Sci. Rep.* **8**, 8339 (2018).
70. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
71. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (32222018, 32170635 to X.C., 32361133555, 32122022 to J.-R.Y.), the National Key R&D Program of China (grant number 2021YFA1302500 to J.-R.Y.). We thank Jianzhi Zhang and Haipeng Li for their comments on the manuscript.

## Author contributions

X.C. and J.-R.Y. conceived the idea and designed the study; X.C. and J.-R.Y. supervised the study and obtained fundings; X.Z., G.Y., Z.G., J. L, W.L., W.Z., and Z.W. acquired data by conducting experiments or from public datasets; X.Z., B.Z., S. P., D. T., J.-R.Y., and X. C. conducted formal data analyses. X.Z., J.-R.Y., and X. C. wrote the paper with inputs from all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-69969-x>.

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

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026