


# Determinants of genome-wide distribution and evolution of uORFs in eukaryotes

Hong Zhang <sup>1</sup>, Yirong Wang<sup>1,2</sup>, Xinkai Wu <sup>1</sup>, Xiaolu Tang <sup>1</sup>, Changcheng Wu <sup>1</sup> & Jian Lu <sup>1</sup>✉

Upstream open reading frames (uORFs) play widespread regulatory functions in modulating mRNA translation in eukaryotes, but the principles underlying the genomic distribution and evolution of uORFs remain poorly understood. Here, we analyze ~17 million putative canonical uORFs in 478 eukaryotic species that span most of the extant taxa of eukaryotes. We demonstrate how positive and purifying selection, coupled with differences in effective population size ( $N_e$ ), has shaped the contents of uORFs in eukaryotes. Besides, gene expression level is important in influencing uORF occurrences across genes in a species. Our analyses suggest that most uORFs might play regulatory roles rather than encode functional peptides. We also show that the Kozak sequence context of uORFs has evolved across eukaryotic clades, and that noncanonical uORFs tend to have weaker suppressive effects than canonical uORFs in translation regulation. This study provides insights into the driving forces underlying uORF evolution in eukaryotes.

<sup>1</sup>State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing, China. <sup>2</sup>College of Biology, Hunan University, Changsha, China. ✉email: [LUJ@pku.edu.cn](mailto:LUJ@pku.edu.cn)

Upstream open reading frames (uORFs) are short open reading frames (ORFs) that have start codons located in the 5' untranslated regions (UTRs) of eukaryotic mRNAs. uORFs can attenuate the translational initiation of downstream coding sequences (CDSs) by sequestering or competing for ribosomes<sup>1–10</sup>. For an AUG triplet in the 5' UTR (defined as “uAUG” hereafter), it can function as the start codon of a uORF that has a stop codon either preceding the start codon of the downstream CDS (nonoverlapping uORF, nORF) or residing in the body of the downstream CDS (out-of-frame overlapping uORF, oORF)<sup>4,11–18</sup>. Less frequently, an uAUG can function as the start codon of an ORF whose stop codon overlaps with the stop codon of the downstream CDS (N-terminal extension, NTE)<sup>4,19–21</sup>. The advent of ribosome profiling<sup>22–26</sup>, a method that determines the ribosome occupancy on mRNAs at the codon level, has enabled the genome-wide characterization of uORFs and NTEs that showed evidence of translation in various species with high sensitivity and accuracy<sup>12,16,17,27–36</sup>. Besides the canonical uORFs (beginning with an AUG start codon and ending with a UAA/UAG/UGA stop codon), the modified ribosome profiling methods<sup>4,37</sup>, which detect initiating ribosomes in cells treated with harringtonine<sup>32,34</sup> or lactimidomycin<sup>38–40</sup>, have provided further evidence showing that many noncanonical uORFs (beginning with a non-AUG codon and ending with a UAA/UAG/UGA stop codon) might be prevalent and functionally important. Collectively, recent studies have demonstrated that uORFs are prevalently translated in eukaryotic cells and that uORF-mediated regulation plays important roles in tuning the translational program during development<sup>32,41–45</sup> or stress responses<sup>10,27,46–55</sup>.

It is well accepted that canonical uORFs are generally deleterious and are depleted in the 5' UTRs of eukaryotic genomes<sup>56–60</sup>, and mutations that generate polymorphic uORFs are also usually deleterious and selected against in humans<sup>19,61–66</sup> and flies<sup>32</sup>. Nevertheless, our recent study indicated that many uAUGs recently fixed in *Drosophila melanogaster* were driven by positive Darwinian selection<sup>32</sup>, which suggests that some uORFs and NTEs might be adaptive. Despite these exciting progress, the principles underlying the genomic distribution and evolution of uORFs and NTEs are poorly understood. For example, the following questions remained unanswered: (1) What is the role of natural selection in shaping the genome-wide contents of uORFs and NTEs in eukaryotes at the micro- and macroevolutionary scales? (2) Can we detect signatures of positive selection on uORFs and NTEs in clades other than *Drosophila*? (3) Are the sequence characteristics that influence the efficacy of uORF-mediated translational repression conserved between different eukaryotic species? Answers to these questions will not only help elucidate the role of translational regulation in adaptation, but also advance our understanding of the mechanisms underlying protein homeostasis in health and disease.

Here, we systematically characterize 16,907,129 uAUGs in 478 eukaryotic species and explore various factors and forces that determine the genome-wide distributions of uORFs and NTEs across genes and species. Our results suggest that differences in uORF occurrences across genes are mainly influenced by gene expression levels, while the interspecific variability of uORFs is shaped by the effective population size ( $N_e$ ). We also compare the conservation patterns of start codons versus coding regions of the canonical uORFs in different clades, disentangled the relationship between the Kozak sequence context and the translational efficiency of uORFs, and explore the evolution of Kozak contextual characteristics across eukaryotes. Our analyses present a broad overview of the interspecies variability of uAUGs in eukaryotes and provide insights into the general principles underlying the distribution and sequence evolution of uORFs and NTEs in eukaryotes.

## Results

**Characterization of putative canonical uORFs and NTEs in 478 eukaryotes.** We developed a bioinformatic pipeline and characterized uAUGs in the genomes of 478 eukaryotic species, including 242 fungi, 20 protists, and 216 multicellular eukaryotes that comprise plants and animals. As most species surveyed in this study currently have no ribosome profiling data, and it is very challenging to predict the noncanonical uORFs in silico reliably, we only focused on the putative canonical uORFs that start with the AUG start codon. In what follows, the uORFs analyzed in this study are restricted to the putative canonical uORFs unless explicitly stated otherwise (all the annotated uORFs and NTEs are presented in figshare<sup>67</sup>).

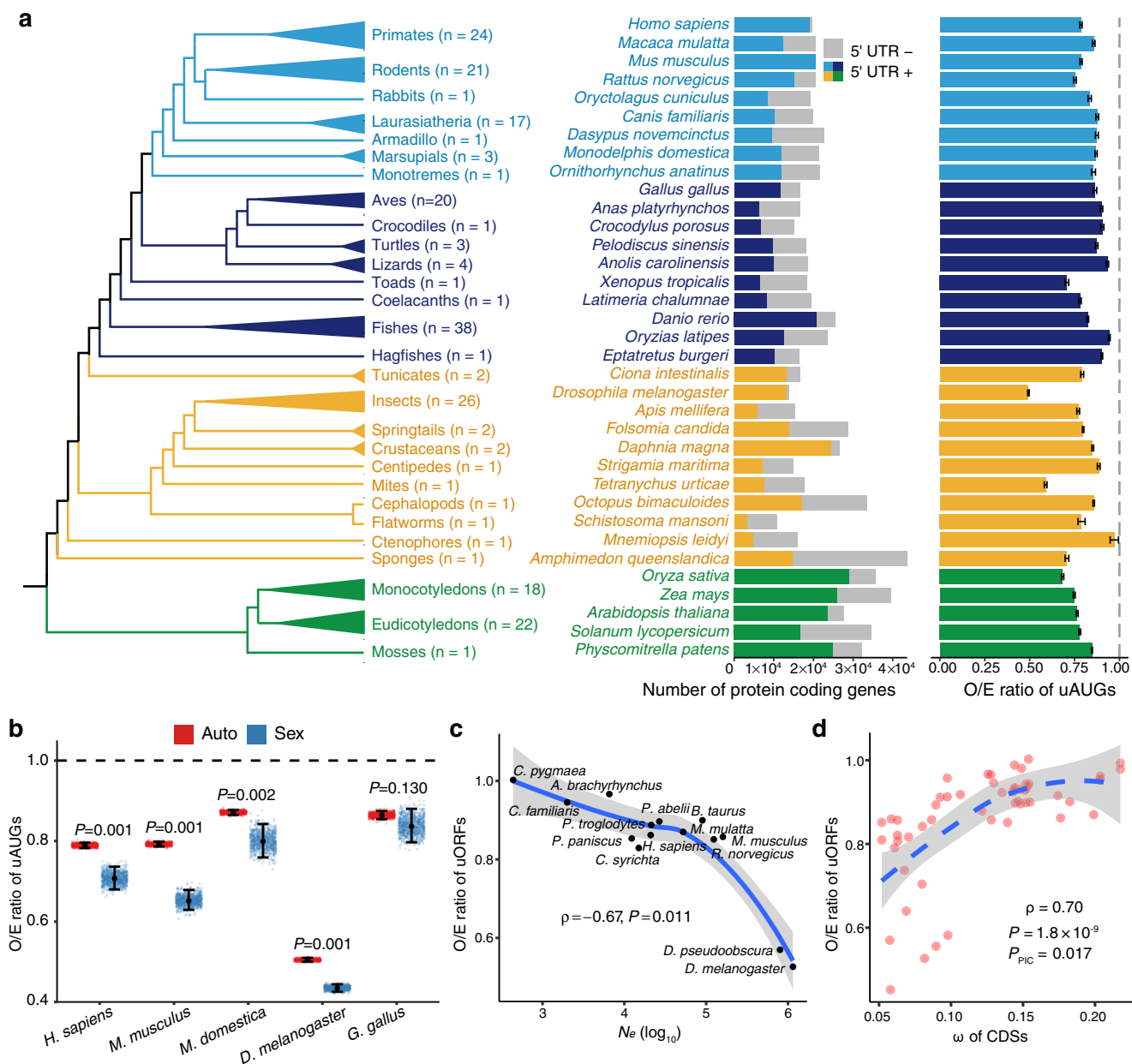
The number of annotated protein-coding genes in the 242 fungi ranged from 3623 (*Pneumocystis murina*) to 32,847 (*Fibularhizogonia sp.*). A total of 3,469,095 uAUGs were identified in these fungal genomes, with the number ranging from 1233 (*Malassezia sympodialis*) to 94,695 (*Verticillium longisporum*) (Supplementary Data 1). Since many protists use alternative nuclear genetic codes involving stop-codon reassignments<sup>68–73</sup> or obligatory frameshifting at internal stop codons<sup>74</sup>, here we only focused on 20 protists that use the standard genetic code (Supplementary Data 1). Among the 20 protists, the number of annotated protein-coding genes ranged from 5389 (*Plasmodium vivax*) to 38,544 (*Emiliania huxleyi*), and the number of uAUGs ranging from 1903 (*Plasmodium falciparum*) to 99,859 (*Cystoisospora suis*), which resulted in a total of 391,565 uAUGs in these protist genomes (Supplementary Data 1).

The 216 multicellular plants and animals, whose last common ancestor was dated to 1.5 billion years ago<sup>75</sup>, span the following taxa: (1) 41 plants, including mosses, eudicotyledons, and monocotyledons; (2) 38 invertebrates, including sponges, ctenophores, flatworms, cephalopods, mites, centipedes, crustaceans, springtails, insects, and tunicates; and (3) 137 vertebrates, including hagfishes, fishes, coelacanths, toads, lizards, turtles, crocodiles, birds, and mammals (Fig. 1). Among these species, mammals, including monotremes ( $n = 1$ ), marsupials ( $n = 3$ ), an armadillo ( $n = 1$ ), laurasiatherians ( $n = 17$ ), a rabbit ( $n = 1$ ), rodents ( $n = 21$ ), and primates ( $n = 24$ ), constituted the largest clade ( $n = 68$  species). Nematodes were excluded from the analyses because *trans*-splicing alters the 5' UTR sequences of many mRNAs in their transcripts<sup>45,76</sup>.

The number of annotated protein-coding genes in the 216 multicellular plants and animals ranged from 10,581 (*Bombus terrestris*) to 107,545 (*Triticum aestivum*), and 3388 (*Schistosoma mansoni*) to 68,741 (*T. aestivum*) of these genes exhibited annotated 5' UTRs for at least one transcript (Fig. 1a and Supplementary Data 1). In these species, the annotated 5' UTRs were usually shorter than 500 nt (the median length of the annotated 5' UTRs ranged from 22 nt in *Arabidopsis lyrata* to 477 nt in *Physcomitrella patens*; Supplementary Fig. S1 and Supplementary Data 1). The number of uAUGs ranged from 3249 (*Drosophila willistoni*) to 798,433 (*Theobroma cacao*). Altogether, we identified a total of 13,046,469 uAUGs in the 216 multicellular plants and animals, although the number varied greatly across species.

The vast majority (>97%) of the uAUGs identified in the 478 eukaryotic species were start codons of putative canonical uORFs. Specifically, in a species, the percentage (mean  $\pm$  s.e.) of nORFs, oORFs, and NTEs was  $83.45 \pm 0.41\%$ ,  $14.24 \pm 0.34\%$ , and  $2.31 \pm 0.15\%$ , respectively. The detailed information for the uORFs (nORFs and oORFs) and NTEs is presented in Supplementary Data 1.

**Purifying selection is the major force shaping the prevalence of uAUGs in eukaryotic genomes.** The number of uAUGs varied



**Fig. 1** Variability of upstream AUG (uAUG) prevalence among eukaryotes and evolutionary driving forces. **a** Overview of the 216 eukaryotes analyzed in this study. The left panel is the cladogram of the 216 eukaryotes. The number of species in each clade is shown in brackets. The middle panel shows the total number of protein-coding genes in 35 representative species. Genes with an annotated 5' untranslated regions (5' UTR+) are colored by clade, and those without 5' UTR annotation (5' UTR-) are shown in gray. The unavailability of annotated 5' UTRs for many genes in less-studied organisms is presumably caused by the lack of accurate annotations. The right panel shows the ratio of the observed number of uAUGs to the expected number of uAUGs (O/E ratio) in the 35 species. The error bars indicate the 95% confidence interval of the O/E ratio. **b** O/E ratios of uAUGs in sex chromosome (X or Z) genes (Sex, blue) and autosomal genes (Auto, red) in humans, mice, opossum, flies, and chickens.  $n$  = 1000 permutation replicates for each category of genes in each species. Center point, median; error bars, 95% confidence intervals.  $P$  values were obtained by two-sided Wilcoxon signed-rank tests, and no correction for multiple testing was made. **c** Relationship between the effective population size ( $N_e$ ) and the O/E ratio of uORFs among 14 animals. The blue line indicates the local polynomial regression fit of the O/E ratio against  $N_e$ , and the gray band indicates the standard error of the fit. Spearman's correlation ( $\rho$ ) between  $N_e$  and the O/E ratio and the two-sided  $P$  value are shown in the plot. **d** Relationship between the genome-wide median number of nonsynonymous changes per nonsynonymous site over the number of synonymous changes per synonymous site ( $\omega$ ) of coding sequences (CDSs) and the O/E ratio of uORFs among 56 animals. The blue line indicates the local polynomial regression fit and the gray band indicates the standard error of the fit. Both Spearman's correlation and the significance of the two-sided phylogenetic independent contrast (PIC) between  $\omega$  and the O/E ratio ( $P_{PIC}$ ) are shown. Source data are provided as a Source Data file.

wildly across species, either due to the differences in the sequencing coverage of genomes, the accuracy and completeness of 5' UTR annotation, the number of protein-coding genes, the length of 5' UTRs, or mutational bias in 5' UTRs<sup>77</sup>. To control for the compounding factors, in each species, we compared the observed number of uAUGs (O) versus the expected number (E) that was obtained with the assumption of randomness by randomly shuffling the 5' UTR sequences. We maintained the same dinucleotide frequencies in each sequence during shuffling for two reasons. First, the stacking energy of a new base pair is influenced by the neighboring base pairs in an RNA molecule<sup>78,79</sup>. Second, the biased mutations in certain dinucleotide contexts, such as from CpG to TpG mutations in mammals, might also affect the occurrences of uAUGs. The O/E ratio enabled the efficient measurement of selective pressure on uAUG depletion in a given species. As expected<sup>32,56–59</sup>, the O/E ratio of uAUGs was significantly lower than 1 in nearly all the examined species (473 out of 478 species, Fig. 1a and Supplementary Data 1). As a negative control, we also calculated the O/E ratio of all the other 63 possible triplets in 5' UTRs and 3' UTRs separately in each species. Of note, AUG had the lowest relative O/E ratio (5' UTRs over 3' UTRs) among all the 64 possible triplets (Supplementary Fig. S2), supporting the notion that purifying selection is the major force shaping the prevalence of uAUGs in the eukaryotic genomes. Interestingly, some AUG-like triplets (e.g., AUU, UUG, AUC, and GUG) tended to have higher O/E ratios in 5' UTRs than in 3' UTRs in all the clades. Such AUG-like triplets were either selectively maintained in 5' UTRs as they can be used as noncanonical start codons, or alternatively, were the consequence of the depletion of uAUGs because point mutations can easily convert AUG to AUG-like triplets (e.g., from AUG → UUG) in the 5' UTRs. However, further studies are required to separate these two possibilities.

Within a species, the O/E ratio of uAUGs was significantly lower in the 5' UTR regions within a distance  $L$  from the start codons of CDSs (cAUGs) than in the remaining 5' UTR regions ( $P = 3.5 \times 10^{-37}$ , two-sided Wilcoxon signed-rank test when  $L$  was set to 100 nt; other values of  $L$  did not affect the conclusion, see Supplementary Fig. S3). This pattern is consistent with previous observations that uAUGs closer to CDSs showed a higher tendency to be depleted from 5' UTRs<sup>57</sup>. Notably, the O/E ratio of oORFs was significantly lower than that of nORFs (Supplementary Fig. S4), suggesting oORFs tend to be more repressive and thus under stronger purifying selection than nORFs. Interestingly, NTEs showed lower O/E ratios than both oORFs and nORFs in 457 out of 478 species (Supplementary Fig. S4), suggesting that novel NTEs were selected against as they might alter protein functions<sup>21</sup>.

X-linked mutations experience stronger selection than autosomal mutations if the fitness effects of the mutations are (partially) recessive<sup>80–82</sup>. If purifying selection is the dominant force acting on the occurrences of uAUGs in a genome, we expect to observe lower O/E ratios of uAUGs on X chromosomes than on autosomes. Indeed, significantly lower O/E ratios of uAUGs were found in X chromosomes than in autosomes, and this finding was obtained with both vertebrates and insects (Fig. 1b). In birds, which present female heterogamety (males ZZ, females ZW), selection is more efficient on the Z chromosome than autosomes<sup>83</sup>. Accordingly, a slightly lower O/E ratio of uAUGs was observed on the Z chromosome than on autosomes (Fig. 1b). Thus, the comparison between sex chromosomes and autosomes reinforces the thesis that purifying selection is the major force governing the prevalence of uORFs and NTEs in eukaryotes.

Overall, these results suggest that uAUGs were selected against in 5' UTRs, and the NTEs, which only accounted for a small fraction (~2.31% on average) of the uAUGs, were also shaped by

strong purifying selection during evolution. Since uORFs (nORFs and oORFs) and NTEs might have different mechanisms in regulating gene expression and function, in what follows, we only focused on the putative canonical uORFs.

**Gene expression level as an important factor influencing the genome-wide distributions of uORFs across genes.** In humans, genes with uORFs exhibited lower expression levels than genes without uORFs<sup>84</sup>. Similarly, our analysis of previously published mRNA and protein abundance data of fly, human, mouse, mustard plant, and yeast revealed uORFs were infrequently detected in housekeeping genes, and there were significant anticorrelations between the gene expression level and the number of uORFs (Supplementary Fig. S5a and Supplementary Data 2). Meanwhile, gene ontology analysis revealed that genes containing putative uORFs tend to be enriched in the categories of signal transduction, transcription factors, and membrane proteins (Supplementary Fig. S5b; Supplementary Data 3). These patterns still held when we focused on the uORFs supported by previously published ribosome profiling data in fly<sup>32</sup> and other species collected in the GWIPs-viz database<sup>85</sup> (Supplementary Table S1; Supplementary Fig. S5b). Noteworthy, the anticorrelation between uORF occurrences and gene expression level well reconciles with the gene ontology analyses as housekeeping genes tend to be highly (or broadly) expressed<sup>86</sup>.

Since gene expression level affects the efficacy of natural selection<sup>87,88</sup>, we further asked whether the efficacy of purifying selection is reduced in removing deleterious uORFs in lowly expressed genes. We grouped genes of a species into 20 equal-sized bins based on increasing expression levels and calculated the O/E ratio of uORFs in each bin. In all the five species we examined, the O/E ratio was lower than 1 in each bin (Supplementary Fig. S5c), suggesting that purifying selection was the dominant evolutionary force acting on the uORF occurrence regardless of gene expression levels. Interestingly, we observed significant anticorrelations between the gene expression level and O/E ratio of uORFs in each species, suggesting that purifying selection acting on uORFs is relatively weak for lowly expressed mRNAs.

Thus, our results suggest that gene expression level is an important factor influencing uORF distribution across genes in a eukaryotic species. Excessive uORFs in highly expressed genes might cause insufficient protein output, which is harmful to the organisms. We postulate that purifying selection has removed deleterious uORFs in the highly expressed genes more efficiently than in the lowly expressed genes. On the other hand, genes in specific functional categories, such as transcriptional factors, which are likely to be lowly expressed, might be preferentially suppressed by uORFs at the translational level for optimizing protein production. Further studies are needed to investigate the relative importance of the two mechanisms in shaping the anticorrelation between gene expression level and uORF occurrence.

**Differences in  $N_e$  influence interspecies differences in uORF occurrences.** The O/E ratio of uORFs varied widely across the eukaryotic species (Fig. 1a and Supplementary Data 1), suggesting that the efficacy of natural selection differs across these species. Because the efficiency of natural selection is determined by  $N_e$ <sup>89</sup>, we questioned whether the differences in the O/E ratios of uORFs between different eukaryotes are due to the differences in  $N_e$ . We reasoned that the O/E ratio should be lower in species with a larger  $N_e$  because purifying selection is the dominant force acting on uORFs, and deleterious uORFs will be depleted more efficiently by purifying selection. Indeed, we uncovered a significant



negative correlation between the O/E ratio and  $N_e$  (Spearman's  $\rho = -0.67$ ,  $P = 0.011$ ) for 14 animals for which the  $N_e$  value was estimated in previous studies (Fig. 1c and Supplementary Table S2).

Because the  $N_e$  value was unknown for most eukaryotes investigated in this study, we calculated the genome-wide average  $dN/dS$  ratio ( $\omega$ , number of nonsynonymous changes per nonsynonymous site over the number of synonymous changes per synonymous site) of CDSs between closely related species as an indirect measure of the average  $N_e$  for a clade based on the following rationale: if a clade includes species with a large  $N_e$ , deleterious nonsynonymous mutations in CDSs will be more efficiently removed by natural selection, resulting in a smaller  $\omega$  value for that clade. Therefore, if the purifying selection is the major force acting on uORF prevalence, a positive correlation between the O/E of uORFs and the  $\omega$  of CDSs would be expected across different species. We aligned orthologous CDS sequences at the genomic scale for 37 pairs of closely related species, and calculated the genome-wide  $\omega$  value for each pair of species (Supplementary Table S3). In this analysis, we assumed that two closely related species would have the same  $\omega$  values and obtained both the O/E ratios of uORFs and the  $\omega$  values of CDSs for 56 species. We uncovered a significant positive correlation between the O/E ratio and the  $\omega$  value ( $\rho = 0.70$ ,  $P = 1.8 \times 10^{-9}$ ; Fig. 1d), which further confirms that the differences in  $N_e$  determine the differences in uORF depletion among eukaryotic genomes. Interestingly, a significant positive correlation between the median 5' UTR length and the  $\omega$  was also observed ( $\rho = 0.54$ ,  $P = 1.4 \times 10^{-5}$ ; Supplementary Fig. S6), suggesting that the 5' UTR length is also under selective constraints. This finding is not surprising because the number of uORFs is generally positively correlated with the 5' UTR length<sup>90</sup>. To exclude the possibility that the observed positive correlations were confounded by the phylogenetic relationships of the eukaryotic species, we also performed phylogenetic independent contrasts<sup>91</sup> and still detected significant positive correlations between the O/E ratio and the  $\omega$  ( $P = 0.017$ ) and between the 5' UTR length and the  $\omega$  ( $P = 0.021$ ). Together, our analyses suggest that purifying selection is the dominant force governing the contents of uORFs in eukaryotes and that the degree of uORF depletion in a species is mainly determined by the  $N_e$  of that species.

### Role of positive selection in influencing the prevalence of uORFs in eukaryotes.

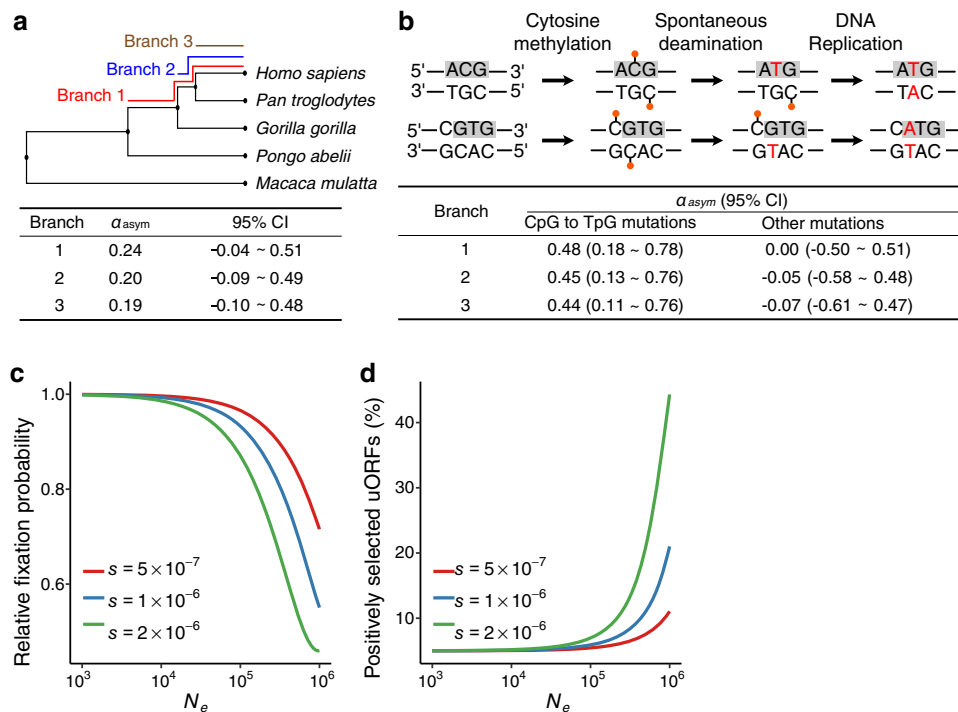
Although uAUGs are generally depleted in the 5' UTRs of *Drosophila*, our previous results indicated that a considerable fraction of the uORFs recently fixed in *D. melanogaster* were driven by positive Darwinian selection<sup>32</sup>. Our results are consistent with the notion that the very large  $N_e$  of *D. melanogaster* increases the efficacy of both positive selection and purifying selection<sup>89</sup>. Nevertheless, whether positive selection drives the fixation of uORFs in a eukaryote with a small  $N_e$ , such as humans, remains unclear. To address this research gap, we analyzed the new uORFs that were newly fixed in the lineages leading to extant humans after divergence from *Pongo abelii*, *Gorilla gorilla*, or *Pan troglodytes* using the asymptotic McDonald-Kreitman test (asymptoticMK)<sup>92,93</sup>. We detected weak signals of positive selection on the newly fixed uORFs in all three branches, and the value of  $\alpha_{asym}$ , which represents the fraction of newly formed uORFs driven to fixation by positive selection, was 0.24 (95% confidence interval [CI],  $-0.04$ – $0.51$ ), 0.20 (95% CI,  $-0.09$  to  $0.49$ ), and 0.19 (95% CI,  $-0.10$  to  $0.48$ ) in the three branches, respectively (Fig. 2a). Noteworthy, C>T mutations at CpG dinucleotides are highly frequent in mammals<sup>94</sup>, and new AUGs can be generated from CpG to TpG mutations through two approaches<sup>95</sup>: (1) from ACG to ATG, and

(2) from CGTG to CATG (Fig. 2b). Thus, we further examined new uORFs derived from the CpG contexts and the remaining new uORFs separately. Roughly speaking, ~33% of the new uORFs fixed in each of the three branches were generated by CpG to TpG mutations. Interestingly, the CpG-derived uORFs were under strong positive selection (the  $\alpha_{asym}$  was 0.48 (95% CI, 0.18–0.78), 0.45 (95% CI, 0.13–0.76), and 0.44 (95% CI, 0.11–0.76) in the three branches, respectively), while the  $\alpha_{asym}$  for the remaining uORFs was close to 0 (Fig. 2b). Noteworthy, the  $\alpha_{asym}$  values were even higher when we focused on the new uORFs that were derived from the CpG contexts in the highly expressed genes (Supplementary Table S4). Of note, for the new uORFs fixed in *D. melanogaster* we previously analyzed<sup>32</sup>, a higher  $\alpha_{asym}$  value was also observed for the highly expressed genes (Supplementary Table S4). Therefore, although the prevalence of uORFs in a species was generally under purifying selection, we still found a fraction of uORFs might be favored by positive selection even in primates that typically have a small  $N_e$ .

To further explore how positive and purifying selection coupled with differences in  $N_e$  shaped the repertoire of uORFs in a given species, we mathematically modeled the O/E ratio of uORFs by treating this ratio as the average fixation probability of mutations with different fitness effects. Considering that uORFs are generally deleterious, we assumed that 20%, 75%, and 5% of newly originated uORFs are neutral, deleterious, and beneficial, respectively. We also assumed that both beneficial and deleterious mutations present the same absolute selection coefficient in a diploid organism and that the fitness reduction in heterozygotes is half of that in homozygous mutants. We then calculated the overall fixation probability of newly originated uORFs relative to the neutral expectation, which is, by definition, similar to the O/E ratio. In our modeling, the relative fixation probability of newly originated uORFs gradually decreased with increases in the  $N_e$  (Fig. 2c), which resembled our observation that species with larger  $N_e$  values tend to exhibit lower O/E ratios. Moreover, a higher fraction of fixed uORFs that are driven by positive selection was obtained with a higher  $N_e$  value (Fig. 2d). Together, our results suggest that both purifying selection and positive selection act on uORF occurrences during eukaryotic evolution and that differences in  $N_e$ , which affects the efficiency of both types of natural selection, plays a major role in shaping the differences in uORF prevalence among eukaryotic species.

### Selective constraints on start codons of uORFs in eukaryotes.

Next, we questioned how the uORFs were maintained during eukaryotic evolution. The start codon is the most important definitive characteristic of a uORF<sup>4</sup>, and a uORF with a more conserved start codon tends to be more repressive toward the translation of the downstream CDS<sup>19,32</sup>. Here, we first quantitatively measured the selective pressures on the AUG start codons of uORFs (uoAUGs) in vertebrates, insects, and yeasts. Among the 78,003 uoAUGs identified in the human reference genome, 98.7% have conserved uoAUGs in other vertebrates (Fig. 3a). Interestingly, 1030 (1.3%) uoAUGs were only observed in humans and not in any other species, suggesting that these have a recent origin. Whether the human-specific uoAUGs are associated with unique human features remains to be investigated. For each uoAUG identified in the human reference genome, we calculated the branch length score (BLS) based on the conservation patterns of the orthologous sites among 100 vertebrate species using a previously described method<sup>96</sup> (Fig. 3b). To estimate the number of uoAUGs that are more conserved than the neutral expectation, we also calculated the BLS for all 63 other triplets present in 5' UTRs based on the assumption that these triplets evolve neutrally. The start codons of 173,290 noncanonical



**Fig. 2 Selection and effective population size ( $N_e$ ) shape the upstream open reading frame (uORF) prevalence in eukaryotes. a** Asymptotic McDonald–Kreitman (AsymptoticMK) test of newly fixed uORFs in the lineages leading to extant humans (branches 1, 2, and 3). The left panel shows the phylogeny of the five primates related to the analysis. Rhesus macaque (*Macaca mulatta*) was used as the outgroup. The fraction of newly fixed uORFs driven by positive selection ( $\alpha_{asym}$ ) is shown in the bottom panel. **b** The result of AsymptoticMK tests for newly fixed uORFs derived from CpG to TpG mutations and the other mutations on each branch. The two approaches by which CpG to TpG mutations create new ATGs in 5' UTRs are illustrated above. Relative fixation probability of newly originated uORFs (**c**) and the fraction of uORFs driven by positive selection (**d**) as a function of the  $N_e$  of a simulated population. In the simulation, we assumed that beneficial and deleterious mutations presented the same absolute selective coefficient ( $s$ ) and that there was no dominance ( $h = 0.5$ ). The fractions of newly originated uORFs that are deleterious, neutral, or beneficial are 75%, 20%, and 5%, respectively. Source data are provided as a Source Data file.

uORFs identified in humans by McGillivray et al.<sup>17</sup> were excluded from the neutral controls.

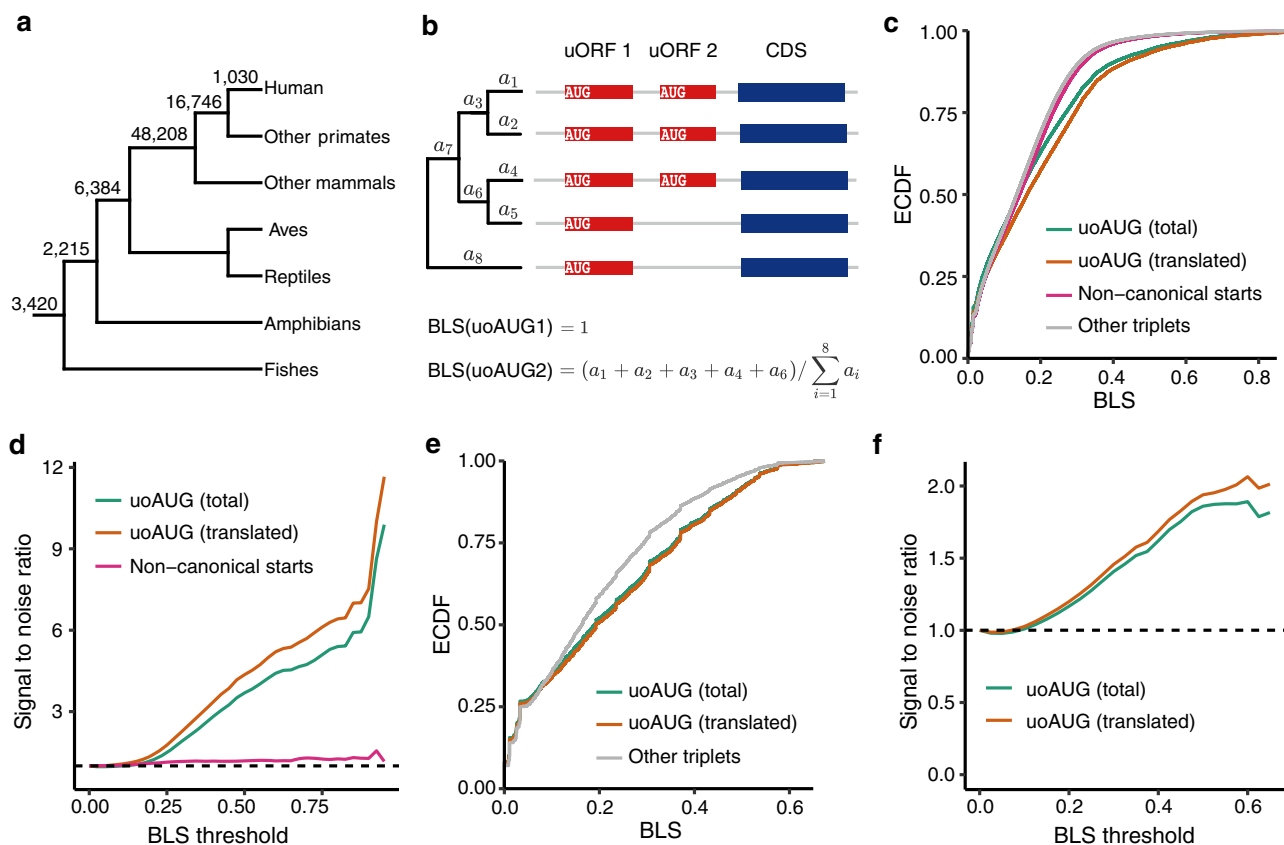
Compared with the other triplets, uoAUGs showed significantly higher BLS values ( $P = 7.6 \times 10^{-58}$ , two-sided Wilcoxon rank-sum test [WRST]; Fig. 3c), suggesting that the uoAUGs are under selective constraints during evolution. At a BLS cutoff of 0.5, the signal-to-noise ratio (fraction of uoAUGs that meet a minimum BLS cutoff divided by the fraction of other triplets with the same minimum BLS) was 3.71, and this value increased with increases in the BLS cutoff (Fig. 3d). Moreover, the BLS values of translated uoAUGs supported by ribosome profiling data from human samples were significantly larger than those of untranslated uoAUGs ( $P = 8.6 \times 10^{-262}$ , two-sided WRST; Fig. 3c). Accordingly, at a BLS cutoff of 0.5, a markedly higher signal-to-noise ratio (4.40) was obtained for the translated uoAUGs (Fig. 3d), suggesting that uoAUGs from which translation is initiated are under even stronger functional constraints. We also calculated the BLS values for the start codons of the 173,290 noncanonical uORFs previously identified in humans by McGillivray et al.<sup>17</sup>. Since conservation was used as a feature to identify the noncanonical uORFs in that study, it is not surprising that these noncanonical start codons were slightly (~1.2 times) more conserved than the other random triplets ( $P = 2.1 \times 10^{-77}$ , two-sided WRST; Fig. 3d). However, they were significantly less conserved than the canonical uoAUGs ( $P = 1.3 \times 10^{-12}$ , two-sided WRST).

The uoAUGs identified in *D. melanogaster* were also more conserved than the random triplets in 5' UTRs across the 27 examined insect species (Fig. 3e, f and Supplementary Fig. S7),

and the uoAUGs with translational evidence from ribosome profiling data were more conserved (Figs. 3e, f). Analogously, the uoAUGs in *S. cerevisiae* were significantly more conserved than the other triplets in the 5' UTRs across the seven yeast species we examined, no matter we used all the uoAUGs or the translated ones only ( $P < 9.5 \times 10^{-11}$ , two-sided WRST; Supplementary Fig. S8). Altogether, the start codons of the canonical uORFs, particularly the translated ones, are more likely to be maintained by functional constraints during eukaryotic evolution.

### Coding regions of uORFs are overall under neutral evolution.

How many uORFs can encode functional peptides remains unclear<sup>4,10,32,97</sup>. If a uORF encodes a functional peptide, one expects that the coding region of that uORF should be under selective constraints. In contrast, if the function of a uORF is to tune the translation of the downstream CDS by sequestering or competing for ribosomes, the coding regions of uORFs might be under neutral evolution or weaker selective constraints. Thus, we investigated the conservation patterns of the uORF peptides in the vertebrates, insects, and yeasts. While NTEs were not included as uORFs in our analysis, we further excluded CDS-overlapping portions of uORFs due to the confounding effects of selective constraints on CDS evolution. Briefly, for each of the 48,286 human uORFs that encode peptides of at least ten amino acids, we searched the putative homologous peptide sequences in other vertebrate species and calculated the BLS for that peptide (homologous sequences that have stop codons or frameshifts within 80% of the start regions were excluded). 48.8% of these

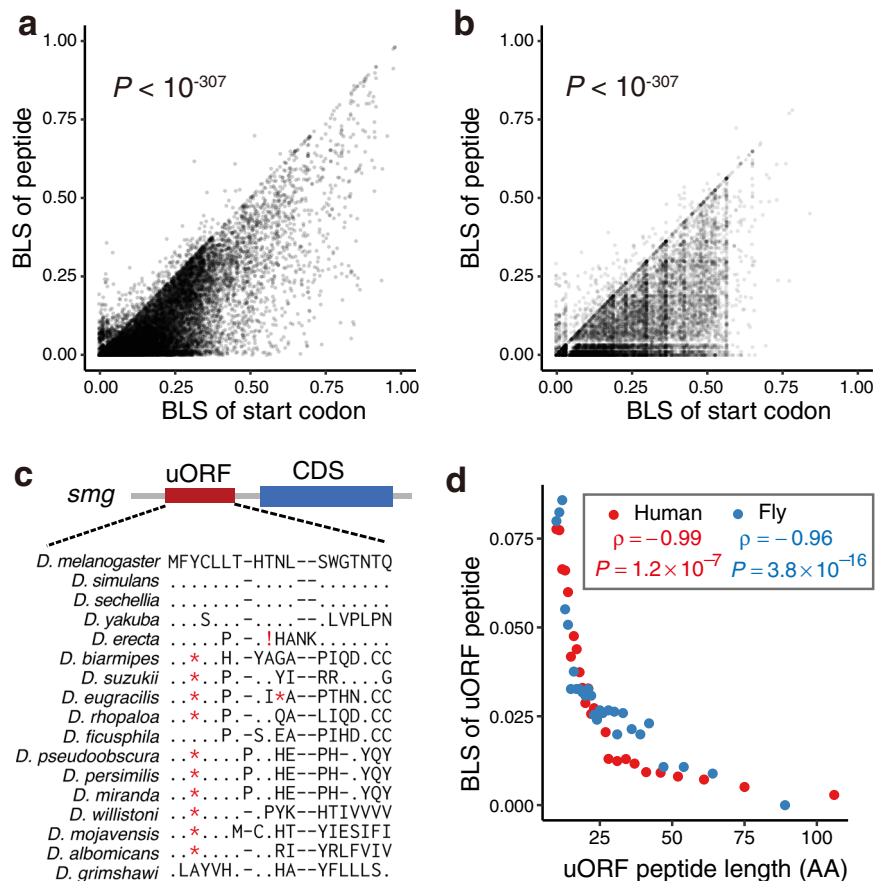


**Fig. 3 Selective constraints on the upstream open reading frame (uORF) start codons.** **a** Age distribution of start codons (uoAUGs) of human uORFs. The number of origination events assigned to each branch was inferred with the maximum parsimony method. **b** The scheme showing how the branch length scores (BLSs) for the start codons of two uORFs are calculated based on their presence or absence across species. In this hypothetical example, the length of each branch is denoted with  $a_i$  ( $i = 1-8$ ). **c** Empirical cumulative distribution function (ECDF) of the BLSs for uoAUGs, noncanonical start codons, and the other triplets in human 5' untranslated regions (UTRs). The BLS of uoAUGs (total or translated) or noncanonical start codons was significantly larger than that of the other triplets ( $P < 8 \times 10^{-58}$ , two-sided Wilcoxon rank-sum tests). **d** Signal-to-noise ratios of the BLSs of uoAUGs and noncanonical start codons relative to other triplets in humans based on different thresholds of minimum BLS. The dashed line delineates a signal-to-noise ratio of 1 expected under neutral evolution. **e** ECDF of BLS for uoAUGs and the other triplets in fly 5' UTRs. The BLS of uoAUGs (total or translated) was significantly larger than that of the other triplets ( $P < 2 \times 10^{-88}$ , two-sided Wilcoxon rank-sum tests). **f** The signal-to-noise ratio of BLSs for uoAUGs relative to other triplets in flies based on different minimum BLS thresholds. Source data are provided as a Source Data file.

human uORFs putatively presented conserved peptide sequences only in primates; 36.6% of them putatively exhibited conserved peptide sequences in mammals other than primates, and 1.82% of them putatively exhibited conserved peptide sequences in fishes. Of note, the BLS values of the uORF coding sequences were significantly lower than those of the uoAUGs (Fig. 4a; Supplementary Fig. S9 for other cutoffs of the minimum number of AAs required for uORF peptides). Analogously, for the uORFs identified in *D. melanogaster*, the coding regions of uORFs were also less conserved than uoAUGs in the 27 examined insect species (Fig. 4b and Supplementary Fig. S9). A similar pattern was also observed in the seven-way alignments of yeasts (Supplementary Fig. S10). Of note, a strong anticorrelation was observed between the BLSs and the lengths of uORF peptides in both humans and flies (see Fig. 4c and d), suggesting the peptides encoded by long uORFs are less likely to be maintained during evolution because they were more likely disrupted by stop codons or frameshifts. Also, if the major function of uORFs is to regulate CDS translation, a longer uORF might be less advantageous than a shorter one because the translation of a longer uORF consumes more energy and metabolites, which might be harmful to the host organisms. (The analysis was not conducted in yeasts because for 69% of the uORF peptides in *S. cerevisiae* we could not reliably identify the orthologous sequences in other yeast species).

Together, these results suggest that the coding regions of uORFs tend to be less conserved than start codons of uORFs.

To further test the selective pressure on the coding sequences of uORFs, we calculated the  $\omega$  for coding regions of uORFs between humans and macaques. To reduce the noise in estimating  $\omega$  values, we ranked the uORFs based on the Kozak scores of their start codons and equally grouped the uORFs into 1000 bins. For each bin, we concatenated the alignments of the uORF coding sequences and calculated the  $\omega$  value. In contrast to CDSs, which present  $\omega$  values markedly lower than 1, the  $\omega$  value of the uORF coding region was roughly equal to 1 between humans and macaques (median  $\omega = 1.05$ ; Fig. 5a and Supplementary Fig. S11a). Similarly, the  $\omega$  of uORFs was also close to 1 between *D. melanogaster* and *D. simulans* (median  $\omega = 0.99$  for all uORFs or 0.98 for translated uORFs only; Fig. 5b and Supplementary Fig. S11b). Moreover, we also grouped the single nucleotide polymorphisms (SNPs) in uORFs of humans (1000 Genomes Project<sup>98</sup>) and flies (*Drosophila* Genetic Reference Panel<sup>99</sup>) based on the derived allele frequencies (DAF) and calculated the ratio of nonsynonymous SNPs to synonymous SNPs ( $pN/pS$ ) in each bin. In parallel, we performed the same analyses on SNPs in CDSs. In CDSs of both humans and flies, the  $pN/pS$  ratios were substantially lower than the values expected under randomness, and the  $pN/pS$  ratio was significantly



**Fig. 4 Conservation of upstream open reading frame (uORF)-encoded peptides.** The branch length score (BLS) of the coding region of a uORF is significantly lower than that of the start codon of that uORF in humans (a) and flies (b). c Example of a typical uORF with a conserved start codon in the fly *smg* gene. The orthologous peptide sequences in distant lineages exhibit many nonsynonymous substitutions and are frequently disrupted by stop codons (\*) or frameshifts (!). d Relationship between the length and the BLS of uORF peptides in humans and flies. The uORFs were grouped into custom bins of increasing peptide length. The median peptide length and BLS value of each bin are displayed and were used to calculate Spearman's correlations ( $\rho$ ) with two-sided  $P$  values. "AA" refers to amino acid. Source data are provided as a Source Data file.

negatively correlated with the DAF bins in both species (Fig. 5c, d; Supplementary Fig. S11c, d). In contrast, in uORFs of both humans and flies, the  $pN/pS$  ratio fluctuated around expected values, and there was no correlation between  $pN/pS$  and DAF bins. Thus, these contrasting patterns indicated that at the population level, the nonsynonymous SNPs in CDSs were under strong purifying selection, while the nonsynonymous SNPs in uORFs were nearly neutral. Together, these analyses further revealed that the coding regions of uORFs are overall under neutral evolution in both primates and flies.

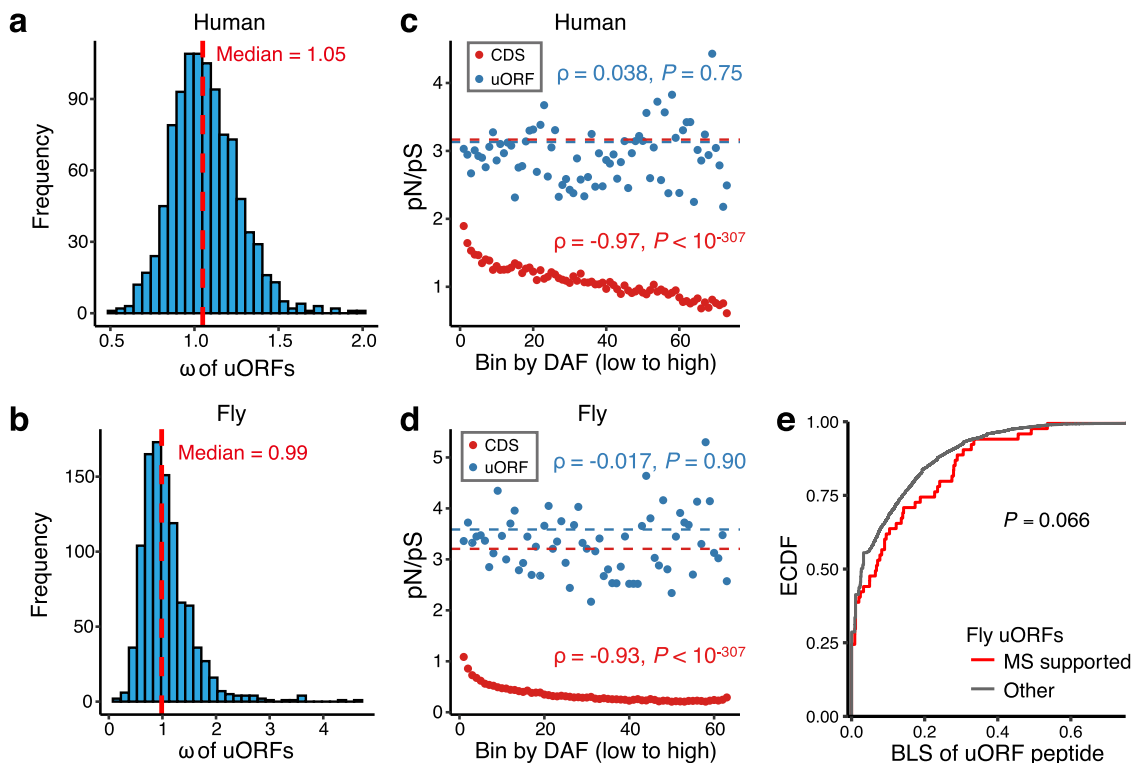
To estimate the proportion of uORFs that might encode conserved peptides, for each uORF, we also calculated PhyloCSF score, which predicts whether a genomic region potentially represents a conserved protein-coding region or not based on multiple sequence alignments<sup>100</sup> (a positive PhyloCSF score means that region is more likely to encode a peptide). As a negative control, we also calculated the PhyloCSF scores for 20,000 randomly selected ORFs in 3' UTRs (downstream ORFs, dORFs), as these dORFs have little chance of translation. Among the 36,655 uORFs that are  $\geq 10$  codons and evidenced of translation in humans, only 361 (0.985%) had positive PhyloCSF scores (Supplementary Fig. S12a). In contrast, the PhyloCSF score was positive for 0.545% (109 out of 20,000) dORFs. Thus, after controlling for the background noises, only 0.44% (161) of the translated uORFs showed evidence of encoding conserved peptides. In *Drosophila*, 1.19% (152 of 12,745) translated uORFs

and 0.39% (78 out of 20000 dORFs) had positive PhyloCSF scores (Supplementary Fig. S12b), yielding an estimate of 0.80% (102 of 12,754) of the translated uORFs might encode conserved peptides. Overall, these analyses suggest that <1% canonical uORFs might encode conserved peptides.

To test whether our evolutionary analyses of uORFs were supported by experimental evidence, we analyzed the mass spectrometry (MS) data from 38 samples of different developmental stages or tissues of *D. melanogaster* (Supplementary Data 4)<sup>101–105</sup>. Among the 23,321 uORFs that met our parameter settings (Methods), 57 (0.24%) had peptides detected in at least one sample (Supplementary Data 5). Interestingly, the BLS analysis revealed that the MS-supported uORFs present slightly more conserved coding regions than the other uORFs (Fig. 5e), suggesting these MS-supported uORF peptides might be functionally important. Collectively, our results support the notion that most uORFs play regulatory roles and their start codons are maintained due to functional constraints, and only a tiny fraction (<1%) of the uORFs might encode peptides that are maintained by natural selection during evolution.

**Evolution of Kozak sequence contextual characteristics that influence uORF translation.** The Kozak sequence context (−6 to +4 nucleotides) around the uoAUG plays a prominent role in influencing the translational initiation of that uORF<sup>16,32,106,107</sup>.





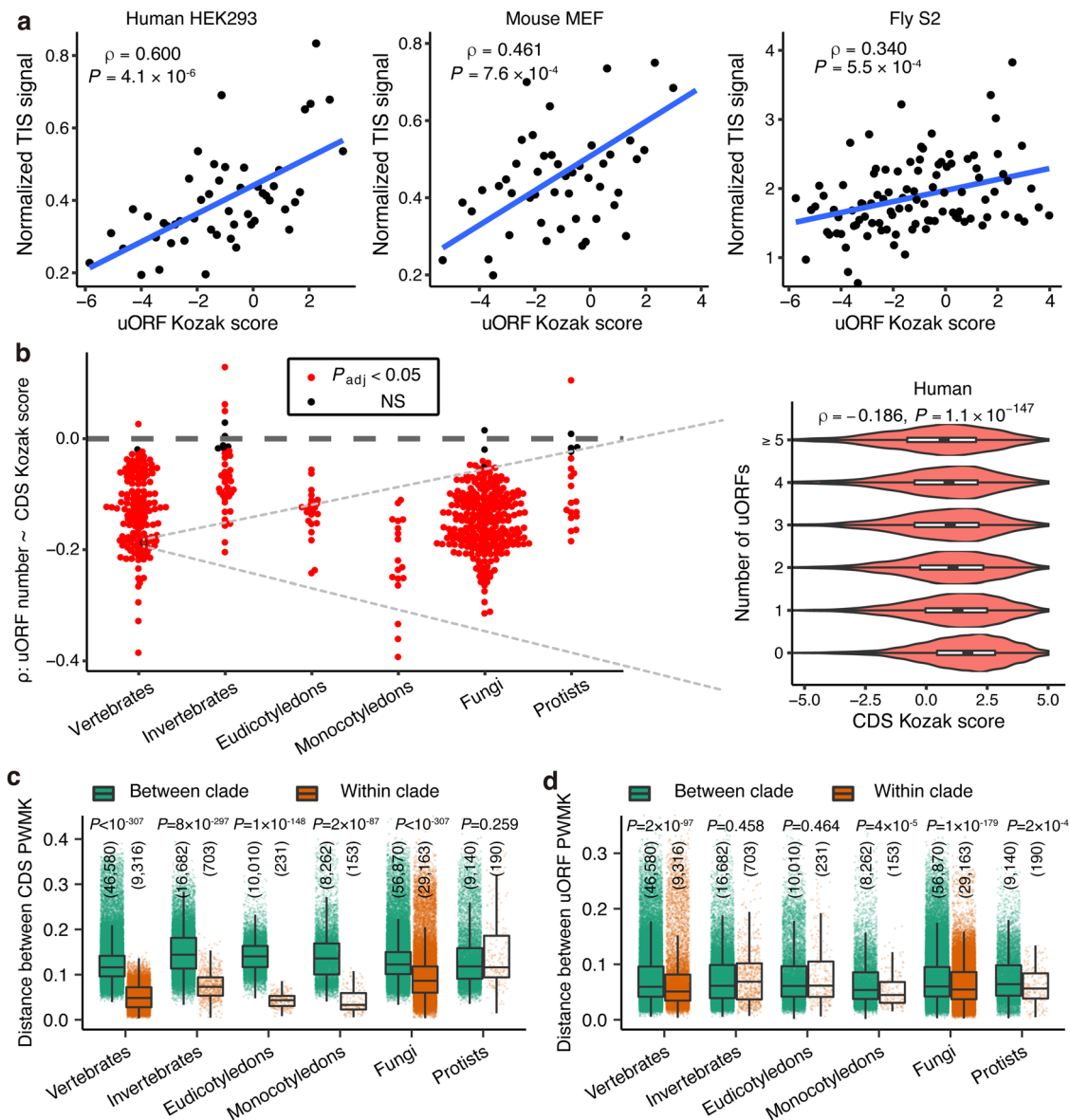
**Fig. 5** Selective constraints on coding regions of upstream open reading frames (uORFs). **a** Distribution of the number of nonsynonymous changes per nonsynonymous site over the number of synonymous changes per synonymous site ( $\omega$ ) of uORFs between humans and rhesus macaques. Human uORFs were equally divided into 1000 bins based on the start codon of uORFs with an increasing Kozak score. For each bin, the alignments of uORF sequences between human and rhesus macaque were concatenated to calculate the  $\omega$  value. **b** Distribution of the  $\omega$  values of uORFs between *D. melanogaster* and *D. simulans*. The procedure for  $\omega$  calculation was similar to that described in **a**. **c** The ratio of the nonsynonymous to synonymous SNP numbers ( $pN/pS$ ) in coding sequences (CDSs, red) and uORFs (blue) in bins with an increasing derived allele frequency (DAF). Spearman's correlation ( $\rho$ ) between the  $pN/pS$  ratio and the median DAF of each bin of uORFs and CDSs is displayed in the plot with two-sided  $P$  values. **d** Same as **c** but for fly uORFs. **e** The empirical cumulative distribution function (ECDF) of peptide branch length score (BLS) for mass spectrometry (MS)-supported uORFs and the remaining uORFs in flies. uORFs with <10 amino acids were excluded. The one-sided  $t$ -test was performed to test differences in BLS. Source data are provided as a Source Data file.

As the nucleotide compositions differ between eukaryotes<sup>108–110</sup>, the preferential Kozak sequence context around cAUGs also differs across species<sup>111,112</sup>. Nevertheless, whether Kozak contextual characteristics around uoAUGs evolve remains unclear. To address this research gap, in each of the 478 species, we reconstructed a position weight matrix of the Kozak sequence context (PWMK) for all the CDSs (Supplementary Data 6 and Supplementary Fig. S13). Subsequently, in each species, we calculated the Kozak score for each cAUG or uoAUG with the PWMK of that species as previously described<sup>32</sup>.

To test the performance of the Kozak score in predicting the translational initiation of a uORF (or CDS), we analyzed translation initiation site (TIS) profiling data from three species (human, mouse, and fly)<sup>32,38,40</sup>. We detected the translation of 26,344, 16,245, and 15,195 canonical uORFs that were supported by TIS data in at least one sample for human, mouse, and fly, respectively (Supplementary Table S1). For each uoAUG in a sample, we calculated the normalized TIS signal by dividing its initiating ribosome-protected fragment (RPF) count by its mean coverage in the matched RNA-Seq data. Strong positive correlations were found between the Kozak score and the normalized TIS signal for both cAUGs (Supplementary Fig. S14) and uoAUGs (Fig. 6a and Supplementary Fig. S15), suggesting that start codons with an optimized Kozak sequence context exhibit a higher translation initiation efficiency for both CDSs and uORFs.

Interestingly, the number of uORFs was negatively correlated with the Kozak score of the cAUGs in most species (Fig. 6b), suggesting that uORFs tend to suppress genes translated at low levels, as previously suggested<sup>59</sup>. Also, the Kozak scores of the uORFs were significantly lower than those of the CDSs in each species (Supplementary Fig. S16a), supporting the notion that uORFs are generally located in less optimal contexts than CDSs<sup>16,32,113</sup>. To test whether the sequence contexts of uoAUGs are optimized, in each species, we also calculated the Kozak scores of the AUG triplets in 3' UTRs (downstream AUGs, dAUGs) as neutral controls. The Kozak scores of uoAUGs were significantly higher than those of dAUGs in most (82.4%, 112 out of 136) vertebrates, (61.0%, 25 out of 41) plants, and (71.9%, 174 out of 242) fungi; however, an opposite trend was observed in invertebrates, and no obvious trend was observed in protists (Supplementary Fig. S16b). These results suggest that the optimization of the Kozak sequence context of uORFs is different across eukaryotic clades.

To examine whether the Kozak contextual characteristics of uORFs evolved, in each of the 478 species, we calculated the pairwise Euclidian distance of the PWMK for uORFs (or CDSs) between two species ("Methods"). Interestingly, for both uORFs and CDSs, the distance between two species from a clade tend to be significantly shorter than that between one species in that clade and another species outside of that clade (Fig. 6c). A similar pattern was observed for the uORF PWMK as well (Fig. 6d).



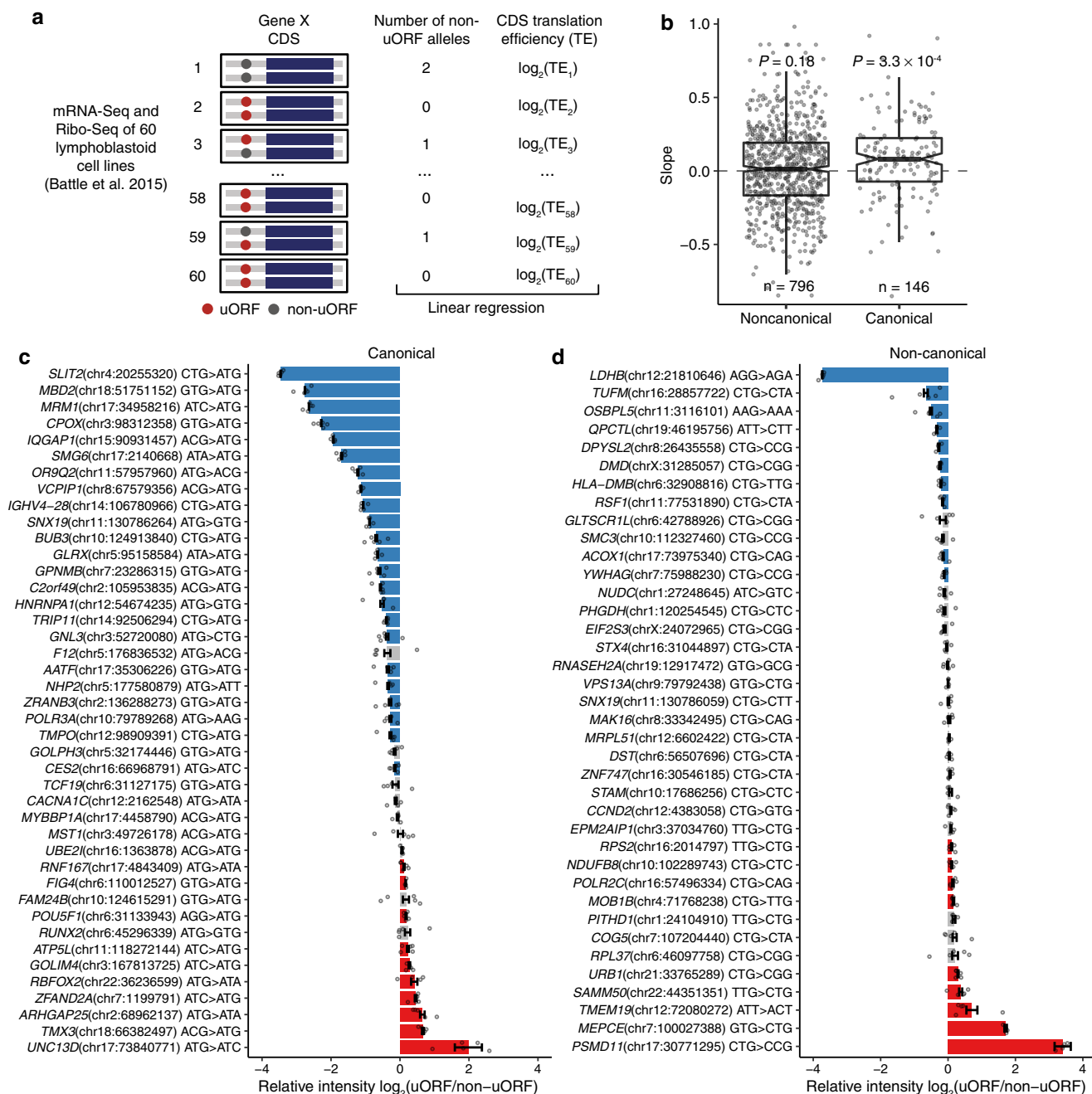
**Fig. 6 The Kozak sequence contextual characteristics that influence upstream open reading frame (uORF) translation.** **a** Relationship between the Kozak scores and normalized translational initiation signals of uORF start codons (uoAUGs) in human HEK293 cells, mouse MEF cells, and fly S2 cells. In each sample, we ranked uORFs based on increasing Kozak scores and divided them into 50 bins (100 bins for S2 cells) with equal numbers of uoAUGs. The median Kozak score and normalized TIS signal for each bin were used to calculate Spearman's correlations ( $\rho$ ) and two-sided  $P$  values. The linear fit was indicated with a blue line. **b** The distribution of Spearman's correlation coefficients between the coding sequence (CDS) Kozak scores and the number of uORFs for that gene in eukaryotes in different taxa. In the left panel, each dot represents one species. The right panel shows that in humans, genes that have multiple uORFs tend to have weaker Kozak sequence context around the start codon of CDSs.  $P_{adj}$ , two-sided  $P$  value after correction for multiple testing; NS, not significant. Box plots showing the distribution of the Euclidian distance of the position weight matrix of Kozak sequences (PwMK) for cAUGs (**c**) and uAUGs (**d**) between species within the same taxa (brown) or species in different taxa (green). Differences in distances were compared with two-sided Wilcoxon rank-sum tests. Exact  $P$  values (no correction for multiple testing were made) and the number of pairwise distances in each group were shown in the plot. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5 times the interquartile range. Source data are provided as a Source Data file.

Together, our results suggest that, although the Kozak sequence context plays a pivotal role in regulating the translational initiation of uORFs and CDSs in eukaryotes, its contextual characteristics evolved during eukaryotic evolution.

**Comparing the canonical versus noncanonical uORFs in repressing CDS translation in human populations.** Recent studies have demonstrated that noncanonical uORFs are very abundant<sup>17,34</sup>, and many of them might have diverse

functions<sup>4,20</sup>. Moreover, hundreds of noncanonical uORFs are conserved between different yeast species, suggesting they might be functionally important<sup>114</sup>. In the above analyses, we mainly focused on the canonical uORFs because (1) the majority of the species analyzed in this study had no ribosome profiling data available, and (2) it is still challenging to identify noncanonical uORFs without experimental data.

To test whether the noncanonical uORFs influence the translation of CDSs, we extracted high-quality genotyping,



**Fig. 7 Experimental verification of canonical and noncanonical upstream open reading frames (uORFs).** **a** The scheme showing how to determine the effect of uORF variations in the human population on the translation efficiency (TE) of downstream coding sequences (CDSs). With the mRNA-Seq and Ribo-Seq data of 60 human lymphoblastoid cell lines<sup>115</sup>, we calculated the translation efficiency of CDS for each gene and obtained the genotypes of each subject from the 1000Genomes Project. For each uORF variant, we performed a linear regression between the number of non-uORF alleles and the  $\log_2(TE)$  of downstream CDS across the 60 cell lines. **b** The distribution of slopes in the linear regression between genotypes (the number of non-uORF alleles) and CDS translational efficiency among 60 cell lines. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5 times the interquartile range. The number of variants in each category was shown in the plot. Exact  $P$  values of two-sided Wilcoxon signed-rank tests were shown in the plot. The ratio of relative luciferase intensity ( $\log_2$ ) between the reporters with the uORF allele or the non-uORF allele for each variant of canonical uORFs (**c**) or noncanonical uORFs (**d**). The bars are displayed in blue or red when the relative intensity of uORF-allele is significantly lower or higher than that of the non-uORF allele (one-sided Wilcoxon rank-sum tests,  $P_{\text{adj}} < 0.1$ ), respectively. Measures of center, mean; error bars, standard errors.  $n = 4$  or 5 independent biological replicates for each variant (details are presented in source data). Source data are provided as a Source Data file.

mRNA-Seq, and Ribo-Seq data of 60 human lymphoblastoid cell lines from previous studies<sup>98,115</sup>, and examined whether variations in uORF start codons influence the translation efficiency of the main CDSs among different samples (Fig. 7a). Among the potentially functional uORFs in humans predicted by McGillivray et al.<sup>17</sup>, 146 canonical and 796 noncanonical uORFs had genetic

variants in their start codons among these samples (only variants with minor allele frequency  $\geq 5\%$  were considered in the analysis). We performed linear regressions to assess the regulatory impact of uORF alteration on the translation of down-stream CDSs, with a positive slope value in the regression meaning that the presence of a uORF in certain individuals is associated with a decrease in

the translation efficiency of the downstream CDS in those individuals, and vice versa (“Methods”). A general trend was the slope values were overall positive for the canonical uORFs, while the slope values for the noncanonical uORFs fluctuated around 0 (Fig. 7b). This comparison suggests that in human populations, the noncanonical uORFs overall have relatively limited repressive effects on CDS translation compared to the canonical uORFs, although we cannot exclude the possibility that a small fraction of the noncanonical uORFs might have strong repressive effects on the translation of downstream CDSs.

To experimentally verify the influence of both types of uORFs on CDS translation, we sampled 80 human uORFs and performed luciferase reporter assays in HEK293FT cells (Supplementary Fig. S17). These tested uORFs, which included 42 canonical and 38 noncanonical ones, were predicted potentially functional by McGillivray et al.<sup>17</sup> and had polymorphic start codons in human populations. For each uORF, we compared the repressive effect of the annotated uORF allele versus that of the non-uORF allele in suppressing translation of the reporter gene. Although occasionally the non-uORF allele had a stronger repressive effect than the uORF allele, the general trend was that the uORF allele had a stronger effect than the non-uORF allele in suppressing translation (Fig. 7c, d). Moreover, a significantly higher proportion of the canonical (55%, 23/42) than the noncanonical (26%, 10/38) uORFs exhibited the pattern that the annotated uORF allele showed a significantly stronger repressive effect on the CDS translation than the non-uORF allele ( $P = 0.013$ , Fisher’s exact test, Fig. 7c, d). Also, the difference in CDS translation suppression between the uORF and the non-uORF allele is significantly larger for the canonical than the noncanonical uORFs ( $P = 0.006$ , one-sided WRST). Altogether, these results reinforced the thesis that the noncanonical uORFs overall have weaker repressive effects on CDS translation than the canonical uORFs.

## Discussion

In this study, we analyzed ~17 million uAUGs,  $97.69 \pm 0.15\%$  of which are start codons of putative canonical uORFs in 478 eukaryotic species that span the majority of extant taxa of eukaryotes. Although the prevalence of canonical uORFs in a species was generally under purifying selection, we still found a fraction of new canonical uORFs might be favored by positive selection even in primates that typically have a small  $N_e$ . These observations are consistent with the evolution model of uORFs we previously proposed<sup>4,32</sup>. Under that model, the majority of newly formed uORFs are deleterious and quickly removed from the population, and a relatively smaller fraction of the new uORFs are beneficial and rapidly fixed in populations under positive selection. After fixation, the functional uORFs, particularly the start codons, are maintained by natural selection during evolution. Hence, although in a species the occurrence of uORFs is influenced by positive or purifying selection, the opposing effects of positive selection and purifying selection acting on new uORFs result in a pattern that uORFs are overall depleted in 5' UTRs. As shown in our population genetic modeling, the efficacies of both positive and purifying selection on uORF fixation in a species are influenced by the effective population size. Moreover, we also found that the gene expression level affects the efficacy of natural selection acting on uORF occurrences. Thus, our results have systematically demonstrated how positive and purifying selection, coupled with differences in gene expression level and  $N_e$ , influence the genome-wide distribution and contents of uORFs in eukaryotes. Together, our analyses reveal the general principles underlying the distribution and sequence evolution of uORFs in eukaryotes. As uORFs often control posttranscriptional gene

expression in combination with other regulators such as microRNAs<sup>90</sup>, further studies are required to elucidate how uORFs coevolve with other regulatory elements.

We found that start codons of canonical uORFs, particularly the translated ones, tend to be maintained by functional constraints during evolution. These results might also be pertinent to the translational buffering mechanism, which indicates that protein expression levels are more conserved between species than mRNAs<sup>116–120</sup>. Nevertheless, our analyses suggest the coding regions of uORFs are overall under neutral evolution. It is not uncommon that some uORF-encoded peptides are conserved across species; however, the conservation of such a peptide does not necessarily mean that peptide might be functional since the coding region of a uORF can be constrained to optimizing translation elongation of that uORF<sup>54,121,122</sup>. Overall, our results suggest that the major function of uORFs is to fine-tune CDS translation rather than to encode conserved peptides. Nevertheless, we do not deny that some uORFs can encode functional peptides, as clearly demonstrated by the previous studies<sup>15,123,124</sup>. Of note, both our PhyloCSF analyses and MS data analyses suggest that a small fraction (<1%) of uORFs might produce peptides.

We found the start codons of the noncanonical uORFs McGillivray et al.<sup>17</sup> identified in humans are overall slightly (~1.2 times) more conserved than the other random triplets across vertebrates. Moreover, our re-analyses of the previously published gene expression data revealed that the noncanonical uORFs tend to have weaker repressive effects on CDS translation than the canonical uORFs, and this pattern was further confirmed by our luciferase reporter assays. Of note, these results do not necessarily suggest that noncanonical uORFs are functionally unimportant, as it has been well established that many noncanonical uORFs might have diverse functions in various biological processes<sup>4,20</sup>, such as stress responses<sup>125,126</sup> or tumor initiation<sup>127</sup>. Overall, our current understanding of the prevalence and function of the noncanonical uORFs are still very limited. Further studies are required to reliably identify the noncanonical uORFs and elucidate their regulatory functions and evolutionary principles.

Protists have a very high phylogenetic diversity<sup>128</sup>, and many protists use alternative nuclear genetic codes involving stop-codon reassignments<sup>68,69</sup> and obligatory frameshifting at internal stop codons<sup>74</sup>. In protists with no dedicated stop codons<sup>71</sup>, such as *Condylostoma magnum*<sup>70,71</sup>, *Parduzia* sp.<sup>71</sup>, *Blastocrithidia*<sup>72</sup>, and *Amoebophrya* sp. ex *Karlodinium veneficum*<sup>73</sup>, translation from any possible uAUG is supposed to terminate near the end of a transcript and overlaps with the main CDS, which results in a different protein. Thus, the occurrence of uORFs in protists with alternative genetic decoding schemes might differ considerably from that of most other eukaryotes. In this study, we only focused on 20 protists that use the standard genetic code. Although the O/E ratio of uAUGs was significantly <1 in all the fungi, multicellular plants and animals we examined, such a pattern was observed in only 15 of the 20 protists. The O/E ratio of uAUGs was close to or higher than 1 in the remaining five protists, including *Cystoisospora suis* (1.161, 95% CI 1.154–1.169), *Toxoplasma gondii* (0.998, 95% CI 0.989–1.007), *Nannochloropsis gaditana* (0.997, 95% CI 0.986–1.007), and two malaria vectors *Plasmodium yoelii* (1.016, 95% CI 1.008–1.025), and *Plasmodium vivax* (0.989, 95% CI 0.975–1.004). However, these five protists tended to have significantly longer 5' UTRs than the other 15 protists (Supplementary Fig. S18), suggesting this observation might be an artifact caused by inaccurate 5' UTR annotations in these five species. Indeed, the O/E ratio of uAUGs in the 5' UTR regions that are proximal to CDS (within 100 or 150 nt) were significantly lower than 1 in all the five protists (Supplementary Table S5), suggesting that uAUG occurrence in 5' UTR regions



proximal to CDSs is still under purifying selection in these protists.

The Kozak sequence context around the uoAUG plays a crucial role in controlling the translation of a uORF<sup>16,32,106,107</sup>, which subsequently regulates translation of the downstream CDS. There has been a growing interest in engineering uORFs for precise translation control of the main protein products<sup>129–131</sup>. Our results revealed the Kozak sequence context evolved across eukaryotic clades, which suggests that the species-specific Kozak sequence contextual features should be considered in designing uORFs for a specific desired trait.

## Methods

**Identification of putative canonical uORFs.** We downloaded the gene models and cDNA sequences of all eukaryotes that are annotated in the Ensembl Genome Browser (release 96)<sup>132</sup>, Ensembl Metazoa (release 43), Ensembl Plants (release 43), Ensembl Protists (release 46), and Ensembl Fungi (release 46). Transcript ends of yeast mRNAs were obtained from a previous study<sup>133</sup>. Putative uORFs and NTEs that start with AUG codons and end with stop codons (UAA/UAG/UGA) were identified from the annotated 5' UTRs of protein-coding genes. uORFs and NTEs with start codons located in CDSs of other transcripts were excluded from the analysis. Only the species for which 5' UTR annotation information was available for more than 25% of the protein-coding genes were considered in the analyses. Among all the 479 species meet this criteria, *Ichthyophthirius multifiliis* was excluded since UAA and UAG are reassigned to encode glutamine in this species<sup>134</sup>, which would interfere with the uORF and NTE prediction.

**Calculation of the O/E ratio.** A permutation analysis was performed to determine the ratio of the observed to the expected number of uAUGs (O/E ratio) for each species. For genes that exhibited more than one transcript, only the longest transcript was used in the analysis. Unusually long 5' UTRs in each species (longer than the mean + 3s.d. of the lengths of the 5' UTRs in that species) were excluded because these are likely annotation artifacts. We denoted the number of AUG triplets in the 5' UTRs of a species as  $n_{obs}$ . We subsequently shuffled the 5' UTRs with 1000 replicates while maintaining the same dinucleotide frequency using uShuffle<sup>135</sup>. We calculated the median and 2.5% and 97.5% quantiles of the number of AUGs in the shuffled 5' UTRs and denoted these numbers as  $n_{exp}$ ,  $n_{0.025}$ , and  $n_{0.975}$ , respectively. We then calculated the O/E ratios for the species as  $n_{obs}/n_{exp}$  with a 95% confidence interval of  $[n_{obs}/n_{0.975}, n_{obs}/n_{0.025}]$ . The O/E ratio of other triplets in 5' UTRs or 3' UTRs was calculated using the same procedure.

**Estimation of the genome-wide  $\omega$  of protein-coding genes.** To estimate the genome-wide average  $\omega$  of protein-coding genes between two closely-related species in a clade, we performed a reciprocal best BLAST<sup>136</sup> of protein sequences between the two species ( $E < 10^{-10}$ ). We identified orthologs of protein-coding genes at the genomic scale for 37 pairs of closely related species, which spanned 56 species. For each pair of orthologs between two species, we aligned their protein sequences with MUSCLE (3.8.31)<sup>137</sup> using the default parameters and generated codon alignments with tranalign from the EMBOSS package<sup>138</sup>. We then calculated  $\omega$  using yn00 from PAML<sup>139</sup> with the codon alignments as input. The median  $\omega$  of all pairs of orthologs between two species was used as the genome-wide  $\omega$  of protein-coding genes. For species that were compared with multiple other species, the median  $\omega$  values obtained from different comparisons were averaged.

**Phylogenetic independent contrasts.** For the 56 metazoan species for which  $\omega$  values were estimated, we obtained the phylogenetic tree from the Open Tree of Life<sup>140</sup>. We used BUSCO<sup>141</sup> to identify single-copy protein orthologs that were conserved in all 56 species, concatenated the protein sequences of the single-copy orthologs in each species, and performed multiple alignments using MUSCLE with the default parameters. Poorly aligned regions in the resulting alignment were removed using trimAl<sup>142</sup> with the “-automated1” method. The branch length of the tree was calculated using codeml from PAML with the JTT substitution model (“seqtype = 2, runmode = 0, model = 2, aaRateFile = jones.dat”). Phylogenetic independent contrasts were performed using the “pic” function in the ape package<sup>143</sup>. The O/E ratio,  $\omega$ , and median 5' UTR length of each species were log-transformed before the contrasts.

**McDonald-Kreitman test of newly fixed uORFs in humans and primates.** To identify fixed differences in AUG triplets in 5' UTRs and introns, we downloaded whole-genome pairwise alignments between humans (hg19 freeze) and other primates (*Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*, and *Macaca mulatta*) from the UCSC Genome Browser<sup>144</sup>. AUG triplets that were newly fixed in the human or hominid lineages were inferred using the parsimonious method with *M. mulatta* as the outgroup. We obtained all human SNPs and their ancestral allele information from the phase 3 data of the 1000 Genomes Project<sup>98</sup>. Both fixed and polymorphic AUG differences located in repetitive regions were excluded from the downstream

analysis. Newly fixed or polymorphic AUGs in 5' UTRs that form NTEs were removed. AsymptoticMK<sup>92,93</sup> tests were performed to detect the signal of positive selection. The data for asymptoticMK tests in flies was obtained from our previous study<sup>32</sup>. To determine the effect of gene expression on positive selection, fixed and segregating mutations were divided into two halves based on the median expression level of genes with fixed new AUGs in 5' UTR. The average protein abundances across different tissues<sup>145</sup> were used from humans, and the average Reads per Kilobase per Million mapped reads (RPKM) values in Ribo-Seq of 12 different developmental stages or tissues were utilized for flies<sup>32</sup>.

**The fixation probability of new uORFs.** For a new autosomal mutation with a selective coefficient  $s$  in a diploid population of size  $N_e$ , the fixation probability of the mutation relative to a neutral mutation was calculated as

$$f(s) = 2N_e \int_0^{\frac{s}{2N_e}} G(x) dx / \int_0^1 G(x) dx, \text{ where } G(x) = \exp[-4N_e s h x - 2N_e s (1 - 2h)x^2]$$

and  $h$  is the dominance coefficient<sup>146</sup>. For mutations that introduce new uORFs into the population, the fractions of neutral, deleterious, and beneficial mutations are denoted as  $p_1$ ,  $p_2$ , and  $p_3$ , respectively. Based on the assumption that the selective coefficients for deleterious and beneficial mutations have the same absolute value, we can obtain the overall relative fixation probability of mutations as  $p_1 + p_2 f(-s) + p_3 f(s)$ . In the simulation, we used a fixed  $h = 0.5$ , and  $p_1$ ,  $p_2$ , and  $p_3$  were set to 0.2, 0.75, and 0.05, respectively.

**Processing of ribosome profiling data.** We obtained pre-calculated ribosome profiling coverage data from humans, mice, rats, zebrafish, and *A. thaliana* from the GWIPs-viz database<sup>85</sup>. Fly ribosome profiling data generated by our group<sup>32</sup> and other researchers<sup>147</sup> that covered all the major developmental stages were also used in this study. Each RPF was assigned to a P-site with plastid<sup>148</sup>. For a uORF in a species, the number of RPFs whose P-sites were within the uORF was calculated with BigWigAverageOverBed<sup>149</sup>. For uORFs that overlapped with CDSs, the overlapping regions were excluded. A uORF was considered as translated if it was covered by the P-site of at least one RPF read across different ribosome profiling datasets in a species.

The genome-wide coverage of initiating ribosome profiling and matched mRNA-Seq data from human and mouse cell lines were downloaded from the GWIPs-viz database. The RNA-Seq data from S2 cells and corresponding Ribo-Seq data after harringtonine treatment were obtained from our previous study<sup>32</sup>. As previously conducted<sup>40</sup>, we first counted the number of initiating RPFs whose P-sites were within the 1-nt flanking region (i.e., -1 to +4) of each uORF or CDS start codon and then normalized the initiating RPF count with the mean coverage of the RNA-Seq data in the same region. We only used start codons with at least 2 initiating RPFs and at least 4 mRNA reads for the downstream analysis of human and mouse cell lines, and those with at least 5 initiating RPF reads and at least 10 mRNA reads were used for the analysis of S2 cells.

**Gene ontology analysis.** Gene ontology (GO) annotations for human, mouse, rat, zebrafish, fly, *A. thaliana*, and yeast were downloaded from the Gene Ontology Resource (2019-06-09 release). Because not all genes under a GO term were provided in the GO annotation files, we parsed the gene annotation files to obtain the complete list of genes under each term using topGO<sup>150</sup>. For each species, all the GO terms belonging to Molecular Function, Biological Process, and Cellular Component were combined in the enrichment analysis. The GO terms enriched in uORF-containing genes or uORF-free genes were determined using Fisher's exact tests. Multiple testing correction was performed with the Benjamini-Hochberg method<sup>151</sup>, and significant terms were determined at a false discovery rate of 0.1 for each species. Nonredundant representative terms that were significantly enriched in at least five species were chosen for visualization.

**Branch length score calculation.** We downloaded the 100-way vertebrate genome alignments based on human (hg19), 27-way insect alignments based on *D. melanogaster* (dm6), the 7-way alignments of yeast species based on *S. cerevisiae* (sacCer3), and the corresponding phylogenetic trees from UCSC Genome Browser and used the Galaxy platform<sup>152</sup> to parse the multiple sequence alignments of 5' UTRs in vertebrates or insects. For the start codon of each human uORF (uoAUG), we calculated the sum of the branch lengths of the subtree composed of the species in which the uAUG was present in the orthologous sites ( $B_0$ ) and then calculated the BLS value by dividing  $B_0$  by the total branch lengths for the phylogenetic tree of the 100 species. Similarly, the BLS was calculated for the start codon of each uORF in *D. melanogaster* across 27 insect species.

For each predicted uORF peptide in humans, we searched its peptide against the orthologous sequences of other species in the 5' UTR alignments using Exonerate (V2.2)<sup>153</sup>. uoAUGs located in repeat regions (downloaded from UCSC Genome Browser) were excluded. For oORFs, only the portion that was not overlapping with CDSs were considered in the analysis. To avoid spurious matching, we only considered human uORF peptides containing at least  $m$  amino acids ( $m$  was set at 10, 15, and 20 in the analysis). We identified uORFs with conserved peptides in other species using the following criteria: (1) the first codon of the matched sequence was AUG; (2) no stop codons or frameshifts were present in the first 80%

of the matched sequence; and (3) between humans and the studied species, the identity of the uORF peptide should be greater than the 2.5% quantile of the genome-wide identity of the main protein sequences. For each uORF, we also calculated the BLs for peptide sequences based on the presence of conserved peptides in other vertebrates as described above. A similar analysis was performed for the fly and yeast uORFs.

Based on these alignments of uORF peptides, we generated alignments of uORF coding regions between humans and macaques and between *D. melanogaster* and *D. simulans*. Due to the short length of uORFs, we ranked the uORFs based on their Kozak score and divided them into 1000 bins with equal numbers of uORFs. For the uORFs in each bin, we concatenated their alignments and calculated  $\omega$  values using yn00 as described above.

**pn/pS analysis.** To study the population variation within uORFs, we merged the genomic intervals of human uORFs and excluded the regions overlapping with CDSs and repeats. We then extracted the SNPs overlapping with uORF regions from the phase 3 data of the 1000 Genomes Project. SNPs in the CDS-overlapping portion of oORFs were excluded. We annotated the effect of SNPs on human uORFs (nonsynonymous or synonymous) using custom scripts and excluded ambiguous SNPs that were annotated as both nonsynonymous and synonymous in different uORFs. For comparison, we also extracted the SNPs in CDS regions and determined their effect on CDSs using SnpEff<sup>154</sup>. The same analysis was performed for uORFs of *D. melanogaster* using the freeze 2 data of the *Drosophila* Genetic Reference Panel<sup>99</sup>.

**PhyloCSF score calculation.** The alignments of human uORFs with at least 10 codons were extracted from the 100-way vertebrate genome alignment based on humans as described above. PhyloCSF for each uORF was calculated with PhyloCSF software<sup>100</sup> using the parameter set “100vertebrates”. As a negative control, we annotated all the possible ORFs in 3' UTRs (dORFs) with at least ten codons using getorf from EMBOSS suit<sup>138</sup>. dORFs overlapping with any CDS or uORF were excluded. We randomly selected 20,000 unique dORFs from the remaining dORFs and calculated PhyloCSF scores with the same procedure as for uORFs. The same analysis was performed for uORFs and dORFs in flies, except that the parameter set “23flies” was used when calculating PhyloCSF scores.

**MS data analysis.** MS datasets for multiple tissues, developmental stages, and cell lines of *D. melanogaster* were obtained from ProteomeCentral<sup>155</sup>. Information on these datasets is listed in Supplementary Data 4. In peptide search, we used a custom database composed of the annotated proteome of *D. melanogaster* and all the peptides encoded by regions between two consecutive in-frame stop codons in cDNA sequences with at least 7 amino acids. To recover as many uORF-encoded peptides as possible, each sample was searched with three different search engines (MaxQuant v1.6.5<sup>156</sup>, OpenMS v2.3.0<sup>157</sup>, and pFind3<sup>158</sup>) at a 1% false discovery rate. Enzyme specificity was set to trypsin, and at most two missing cleavages were allowed. Cysteine carbamidomethylation was included as the fixed modification and methine oxidation as the variable modification. Both the precursor and fragment tolerance were set to 20 ppm for higher-energy collisional dissociation (CID) datasets. Peptides with <7 amino acids were excluded during searching. Peptides that match the built-in contaminants in MaxQuant, yeast proteins and the annotated fly proteome were removed with PeptideMatchCMD v1.0<sup>159</sup> allowing mismatches of leucine and isoleucine. The remaining peptides were mapped to peptides encoded by all the putative canonical uORFs. uORFs with uniquely mapped peptides were kept as MS-supported uORFs.

**Calculation of the Kozak score.** For each species, we retrieved the six nucleotides upstream of the CDS start codons and one nucleotide downstream of these codons and built a position probability matrix (PWM) as the Kozak sequence context. We then determined the Kozak score for the start codon of a uORF or CDS, as well as for each AUG in 3' UTRs by calculating the log-odds ratio of their flanking sequences using the above-derived PWM<sup>32</sup>. The Euclidian distance between two PWMs of uORF or CDS Kozak sequences was calculated using TFBSTools<sup>160</sup>.

**Effect of uORF variation on CDS translation in human populations.** The RNA-Seq and ribosome profiling data from lymphoblastoid cell lines (LCLs) were obtained in a previous study<sup>115</sup>. High-quality genotyping data from 60 LCLs were obtained from the 1000 Genomes project<sup>98</sup>. After pre-processing, the RNA-Seq reads and RPFs were mapped to the human reference genome with STAR<sup>161</sup>. Reads mapped to the CDS region of each protein-coding gene were tabulated with htseq-count<sup>162</sup>. CDS read counts were normalized across different cell lines with DESeq2<sup>163</sup> separately for RNA-Seq and RPFs. The translation efficiency of a gene in a sample was calculated as the ratio of the normalized RPF read count over the normalized RNA-Seq read count. To control for false positives, only SNPs that disrupt the canonical and noncanonical uORFs annotated by McGillivray et al.<sup>17</sup> were analyzed. SNPs with a minor allele frequency of <5% among the 60 LCLs were excluded. A SNP is classified as a canonical uORF variant if the wild-type start codon or mutant start codon is AUG and is classified as noncanonical otherwise. For each uORF variant, a linear regression was performed between the CDS

translational efficiency and the number of non-uORF alleles (0, 1, or 2) across different LCLs.

**Experimental verification of uORF variants.** The effects of uORF variants were assayed with dual-luciferase reporter assays (psiCHECK-2 vector, Promega). HEK293FT cells were purchased from the Cell Bank of the Chinese Academy of Sciences. RNA was extracted using the TRIzol reagent (15596018, Thermo Fisher Scientific), and cDNAs were synthesized using the PrimeScript First-strand cDNA Synthesis Kit (6110B, TaKaRa). For each uORF variant, the wild-type (WT) 5' UTR or mutated 5' UTR were cloned from cDNAs by PCR. All the primers used for cloning 5' UTR fragments were listed in Supplementary Data 7. The reporter plasmid was linearized using Nhe1 (R3131S, NEB). The product and the 5' UTR sequence were assembled using the NEBuilder HiFi DNA Assembly Cloning Kit (E5520S, NEB). Plasmid libraries were extracted using a QIAGEN Miniprep kit (27106, QIAGEN) according to the manufacturer's instructions. The constructed vectors were transfected into HEK293FT cells using Lipofectamine 3000 Transfection Reagent (L3000015, Thermo Fisher Scientific). The cells were cultivated in Dulbecco's modified Eagle's medium (DMEM) with 10% FBS for 32 h. Then the psiCHECK-2 dual-luciferase reporter assay system (Promega) was used to detect levels of the Renilla luciferase with WT or mutant 5' UTR and normalized to the firefly luciferase as an internal control. At least four biological repetitions were performed for each WT or mutated 5' UTR plasmid.

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

## Data availability

The putative uORFs and NTEs annotated in this study are available from figshare<sup>67</sup> (<https://doi.org/10.6084/m9.figshare.9980441.v4>). The following public data were analyzed in this study: (1) gene annotations, cDNA sequences and genome sequences from Ensembl Genome Browser (<https://www.ensembl.org> and <http://ensemblgenomes.org>); (2) the transcript ends of yeast mRNAs from Gene Expression Omnibus (GEO) under the accession number GSE49026; (3) functional annotation of gene categories from The Gene Ontology Resource (<http://geneontology.org>); (4) gene expression data in model organisms from previous studies as listed in Supplementary Data 2; (5) Ribo-Seq data from GWIPs-viz database (<https://gwips.ucc.ie>) and our previous study<sup>32</sup>; (6) the effective population size reported in previous studies as listed in Supplementary Table 2; (7) single nucleotide polymorphisms from the 1000 Genomes Project (<https://www.internationalgenome.org/data>) and DGRP2 (<http://dgrp2.gnets.ncsu.edu>); (8) multiple genome alignments from UCSC Genome Browser (<https://genome.ucsc.edu>); (9) the annotation of potential functional uORFs in humans from McGillivray et al.<sup>17</sup>; (10) mass spectrometry datasets from ProteomeCentral (<http://proteomecentral.proteomexchange.org>) as listed in Supplementary Data 4; (11) RNA-Seq and Ribo-Seq data of human lymphoblastoid cell lines from GEO under the accession number GSE61742 and the Gilad/Pritchard group ([http://eqtl.uchicago.edu/RNA\\_Seq\\_data](http://eqtl.uchicago.edu/RNA_Seq_data)). Source Data have been deposited in figshare (<https://doi.org/10.6084/m9.figshare.12612068.v2>) and are provided with this paper.

## Code availability

The data investigated in this study were analyzed using R statistical software (v3.6). The custom scripts used in this study are available from figshare (<https://doi.org/10.6084/m9.figshare.12612068.v2>).

Received: 30 October 2019; Accepted: 20 January 2021;

Published online: 17 February 2021

## References

- Jackson, R. J., Hellen, C. U. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).
- Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745 (2009).
- Ruiz-Orera, J. & Alba, M. M. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.* **35**, 186–198 (2018).
- Zhang, H., Wang, Y. & Lu, J. Function and evolution of Upstream ORFs in eukaryotes. *Trends Biochem. Sci.* **44**, 782–794 (2019).
- Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
- Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642 (2000).
- Wethmar, K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev.: RNA* **5**, 765–768 (2014).

8. Wethmar, K., Smink, J. J. & Leutz, A. Upstream open reading frames: molecular switches in (patho)physiology. *Bioessays* **32**, 885–893 (2010).
9. Medenbach, J., Seiler, M. & Hentze, Matthias W. Translational control via protein-regulated upstream open reading frames. *Cell* **145**, 902–913 (2011).
10. Orr, M. W., Mao, Y., Storz, G. & Qian, S.-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.* **48**, 1029–1042 (2020).
11. Calviello, L. et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
12. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).
13. Whiffin, N. et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 2523 (2020).
14. Brar, G. A. et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557 (2012).
15. Aspden, J. L. et al. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife* **3**, e03528 (2014).
16. Chew, G. L., Pauli, A. & Schier, A. F. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.* **7**, 11663 (2016).
17. McGillivray, P. et al. A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.* **46**, 3326–3338 (2018).
18. Niu, R. et al. uORFlight: a vehicle toward uORF-mediated translational regulation mechanisms in eukaryotes. Database 2020, <https://doi.org/10.1093/database/baaa007> (2020).
19. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl Acad. Sci. USA* **106**, 7507–7512 (2009).
20. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
21. Benitez-Cantos, M. S. et al. Translation initiation downstream from annotated start codons in human mRNAs coevolves with the Kozak context. *Genome Res.* **30**, 974–984 (2020).
22. Calviello, L. & Ohler, U. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.* **33**, 728–744 (2017).
23. Andreev, D. E. et al. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.* **45**, 513–526 (2017).
24. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16**, 651–664 (2015).
25. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).
26. Ingolia, N. T. Ribosome footprint profiling of translation throughout the genome. *Cell* **165**, 22–33 (2016).
27. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
28. Guenther, U. P. et al. The helicase Ded1p controls use of near-cognate translation initiation codons in 5' UTRs. *Nature* **559**, 130–134 (2018).
29. Lei, L. et al. Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J.: Cell Mol. Biol.* **84**, 1206–1218 (2015).
30. Hsu, P. Y. et al. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc. Natl Acad. Sci. USA* **113**, E7126–e7135 (2016).
31. Bazin, J. et al. Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl Acad. Sci. USA* **114**, E10018–E10027 (2017).
32. Zhang, H. et al. Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. *PLoS Biol.* **16**, e2003903 (2018).
33. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife* **2**, e01179 (2013).
34. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
35. Stumpf, Craig R. et al. The translational landscape of the mammalian cell cycle. *Mol. Cell* **52**, 574–582 (2013).
36. Fritsch, C. et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22**, 2208–2218 (2012).
37. Wang, Y., Zhang, H. & Lu, J. Recent advances in ribosome profiling for deciphering translational regulation. *Methods* <https://doi.org/10.1016/j.ymeth.2019.05.011> (2019).
38. Lee, S. et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl Acad. Sci. USA* **109**, E2424–E2432 (2012).
39. Garreau de Loubresse, N. et al. Structural basis for the inhibition of the eukaryotic ribosome. *Nature* **513**, 517–522 (2014).
40. Gao, X. et al. Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12**, 147–153 (2015).
41. Resch, A. M., Ogurtsov, A. Y., Rogozin, I. B., Shabalina, S. A. & Koonin, E. V. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genom.* **10**, 162 (2009).
42. Chen, J. et al. Kinetochores inactivation by expression of a repressive mRNA. *Elife* **6**, <https://doi.org/10.7554/eLife.27417> (2017).
43. Cheng, Z. et al. Pervasive, coordinated protein-level changes driven by transcript isoform switching during meiosis. *Cell* **172**, 910–923.e916 (2018).
44. Kurihara, Y. et al. Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in Arabidopsis. *Proc. Natl Acad. Sci. USA* **115**, 7831–7836 (2018).
45. Yang, Y. F. et al. Trans-splicing enhances translational efficiency in C. elegans. *Genome Res.* **27**, 1525–1535 (2017).
46. Sidrauskis, C., McGeachy, A. M., Ingolia, N. T. & Walter, P. The small molecule ISRIB reverses the effects of eIF2 $\alpha$  phosphorylation on translation and stress granule assembly. *eLife* **4**, e05033 (2015).
47. Andreev, D. E. et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4**, e03971 (2015).
48. Hinnebusch, A. G. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.* **59**, 407–450 (2005).
49. Young, S. K., Willy, J. A., Wu, C., Sachs, M. S. & Wek, R. C. Ribosome reinitiation directs gene-specific translation and regulates the integrated stress response. *J. Biol. Chem.* **290**, 28257–28271 (2015).
50. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl Acad. Sci. USA* **101**, 11269–11274 (2004).
51. Xu, G. et al. Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature* **545**, 487–490 (2017).
52. Andreev, D. E. et al. Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biol.* **16**, 90 (2015).
53. Andreev, D. E. et al. TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *Elife* **7**, <https://doi.org/10.7554/eLife.32563> (2018).
54. Gaba, A., Jacobson, A. & Sachs, M. S. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol. Cell* **20**, 449–460 (2005).
55. Gerashchenko, M. V., Lobanov, A. V. & Gladyshev, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl Acad. Sci. USA* **109**, 17394–17399 (2012).
56. Kozak, M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* **9**, 5233–5252 (1981).
57. Lynch, M., Scofield, D. G. & Hong, X. The evolution of transcription-initiation sites. *Mol. Biol. Evol.* **22**, 1137–1146 (2005).
58. Neafsey, D. E. & Galagan, J. E. Dual modes of natural selection on upstream open reading frames. *Mol. Biol. Evol.* **24**, 1744–1751 (2007).
59. Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V. & Milanese, L. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics* **17**, 890–900 (2001).
60. Churbanov, A., Rogozin, I. B., Babenko, V. N., Ali, H. & Koonin, E. V. Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes. *Nucleic Acids Res.* **33**, 5512–5520 (2005).
61. von Bohlen, A. E. et al. A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Mol. Genet. Genom. Med.* **5**, 261–268 (2017).
62. Schulz, J. et al. Loss-of-function uORF mutations in human malignancies. *Sci. Rep.* **8**, 2395 (2018).
63. Barbosa, C., Peixeiro, I. & Romao, L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* **9**, e1003529 (2013).
64. Cenik, C. et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25**, 1610–1621 (2015).
65. Wiestner, A., Schlemper, R. J., van der Maas, A. P. & Skoda, R. C. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat. Genet.* **18**, 49–52 (1998).
66. Liu, L. et al. Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat. Genet.* **21**, 128–132 (1999).
67. Zhang, H. et al. The annotation of upstream open reading frames and N-terminal extensions in 478 eukaryotes. [figshare.9980441.v4](https://doi.org/10.6084/m9.figshare.9980441.v4) (2020).
68. Sengupta, S. & Higgs, P. G. Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol.* **80**, 229–243 (2015).



69. Baranov, P. V., Atkins, J. F. & Yordanova, M. M. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nat. Rev. Genet.* **16**, 517–529 (2015).
70. Heaphy, S. M., Mariotti, M., Gladyshev, V. N., Atkins, J. F. & Baranov, P. V. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Mol. Biol. evolution* **33**, 2885–2889 (2016).
71. Swart, E. C., Serra, V., Petroni, G. & Nowacki, M. Genetic codes with no dedicated stop codon: context-dependent translation termination. *Cell* **166**, 691–702 (2016).
72. Záhonová, K., Kostygov, A. Y., Ševčíková, T., Yurchenko, V. & Eliáš, M. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* **26**, 2364–2369 (2016).
73. Bachvaroff, T. R. A predated nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean *Amoebophrya* sp. ex *Karlodinium veneficum*. *PLoS One* **14**, e0212912 (2019).
74. Lobanov, A. V. et al. Position-dependent termination and widespread obligatory frameshifting in *Euplotes* translation. *Nat. Struct. Mol. Biol.* **24**, 61–68 (2017).
75. Kumar, S., Stecher, G., Suleski, M. & Heddes, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
76. Nilsen, T. W. Trans-splicing of nematode premessenger RNA. *Annu. Rev. Microbiol.* **47**, 413–440 (1993).
77. Reuter, M., Engelstädter, J., Fontanillas, P. & Hurst, L. D. A test of the null model for 5' UTR evolution based on GC content. *Mol. Biol. Evolution* **25**, 801–804 (2008).
78. Clote, P., Ferré, F., Kranakis, E. & Krizanc, D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* **11**, 578–591 (2005).
79. Workman, C. & Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27**, 4816–4822 (1999).
80. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
81. Meisel, R. P. & Connallon, T. The faster-X effect: integrating theory and data. *Trends Genet.* **TIG 29**, 537–544 (2013).
82. Lu, J. & Wu, C.-I. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc. Natl Acad. Sci. USA* **102**, 4063–4067 (2005).
83. Mank, J. E., Axelsson, E. & Ellegren, H. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. *Genome Res.* **17**, 618–624 (2007).
84. Ye, Y. et al. Analysis of human upstream open reading frames and impact on gene expression. *Hum. Genet.* **134**, 605–612 (2015).
85. Michel, A. M. et al. GWIPS-viz: 2018 update. *Nucleic Acids Res.* **46**, D823–D830 (2017).
86. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
87. dos Reis, M. & Wernisch, L. Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.* **26**, 451–461 (2009).
88. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).
89. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
90. Zhang, H. et al. Combinatorial regulation of gene expression by uORFs and microRNAs in *Drosophila*. *Sci. Bull.* <https://doi.org/10.1016/j.scib.2020.10.012> (2020).
91. Felsenstein, J. Phylogenies and the comparative method. *Am. Naturalist* **125**, 1–15 (1985).
92. Haller, B. C. & Messer, P. W. asymptoticMK: a web-based tool for the asymptotic McDonald-Kreitman test. *G3* **7**, 1569–1575 (2017).
93. Messer, P. W. & Petrov, D. A. Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl Acad. Sci. USA* **110**, 8615–8620 (2013).
94. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local determinants of the mutational landscape of the human genome. *Cell* **177**, 101–114 (2019).
95. Kitano, S., Kurasawa, H. & Aizawa, Y. Transposable elements shape the human proteome landscape via formation of cis-acting upstream open reading frames. *Genes Cells* **23**, 274–284 (2018).
96. Stark, A. et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219 (2007).
97. Hayashi, N. et al. Identification of *Arabidopsis thaliana* upstream open reading frames encoding peptide sequences that cause ribosomal arrest. *Nucleic Acids Res.* **45**, 8844–8858 (2017).
98. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
99. Mackay, T. F. et al. The *Drosophila melanogaster* genetic reference panel. *Nature* **482**, 173–178 (2012).
100. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
101. Xing, X. et al. Qualitative and quantitative analysis of the adult *Drosophila melanogaster* proteome. *Proteomics* **14**, 286–290 (2014).
102. Casas-Vila, N. et al. The developmental proteome of *Drosophila melanogaster*. *Genome Res.* **27**, 1273–1285 (2017).
103. Ashley, J. et al. Retrovirus-like Gag protein Arc1 binds RNA and traffics across Synaptic Boutons. *Cell* **172**, 262–274 e211 (2018).
104. Kuznetsova, K. G. et al. Proteogenomics of Adenosine-to-Inosine RNA Editing in the Fruit Fly. *J. Proteome Res.* **17**, 3889–3903 (2018).
105. Sabbadin, F. et al. An ancient family of lytic polysaccharide monoxygenases with roles in arthropod development and biomass digestion. *Nat. Commun.* **9**, 756 (2018).
106. Sample, P. J. et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).
107. Noderer, W. L. et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748–748 (2014).
108. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genom. Hum. Genet.* **10**, 285–311 (2009).
109. Katju, V. & Bergthorsson, U. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol. Evol.* **11**, 136–165 (2019).
110. Gentles, A. J. & Karlin, S. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11**, 540–546 (2001).
111. Cavener, D. R. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* **15**, 1353–1361 (1987).
112. Hernandez, G., Osnaya, V. G. & Perez-Martinez, X. Conservation and variability of the AUG initiation codon context in eukaryotes. *Trends Biochem. Sci.* <https://doi.org/10.1016/j.tibs.2019.07.001> (2019).
113. Schleich, S. et al. DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth. *Nature* **512**, 208–212 (2014).
114. Spealman, P. et al. Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res.* **28**, 214–222 (2018).
115. Battle, A. et al. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
116. Signor, S. A. & Nuzhdin, S. V. The evolution of gene expression in cis and trans. *Trends Genet.* **34**, 532–544 (2018).
117. McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24**, 422–430 (2014).
118. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Res.* **24**, 411–421 (2014).
119. Wang, S. H., Hsiao, C. J., Khan, Z. & Pritchard, J. K. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol.* **19**, 83–83 (2018).
120. Khan, Z. et al. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100 (2013).
121. Lin, Y. et al. Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Res.* (2019).
122. Ivanov, I. P. et al. Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. *Mol. Cell* **70**, 254–264.e256 (2018).
123. Mackowiak, S. D. et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16**, 179 (2015).
124. van der Horst, S., Snel, B., Hanson, J. & Smeeckens, S. Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid sequences in the 5' leader of mRNAs *Arabidopsis thaliana*. *Rna* **25**, 292–304 (2019).
125. Kim, J. H., Park, S. M., Park, J. H., Keum, S. J. & Jang, S. K. eIF2A mediates translation of hepatitis C viral mRNA under stress conditions. *EMBO J.* **30**, 2454–2464 (2011).
126. Starck, S. R. et al. Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**, aad3867 (2016).
127. Sandoel, A. et al. Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**, 494–499 (2017).
128. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
129. Xu, G. et al. uORF-mediated translation allows engineered plant disease resistance without fitness costs. *Nature* **545**, 491–494 (2017).
130. Zhang, H. et al. Genome editing of upstream open reading frames enables translational control in plants. *Nat. Biotechnol.* **36**, 894–898 (2018).
131. Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proc. Natl Acad. Sci.* **110**, 11284 (2013).



132. Aken, B. L. et al. The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
133. Park, D., Morris, A. R., Battenhouse, A. & Iyer, V. R. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res.* **42**, 3736–3749 (2014).
134. Coyne, R. S. et al. Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. *Genome Biol.* **12**, R100 (2011).
135. Jiang, M., Anderson, J., Gillespie, J. & Mayne, M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinform.* **9**, 192 (2008).
136. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
137. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
138. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
139. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
140. Hinchliff, C. E. et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl Acad. Sci. USA* **112**, 12764–12769 (2015).
141. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
142. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
143. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
144. Haussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–d858 (2019).
145. Kim, M. S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
146. Kimura, M. Diffusion models in population genetics. *J. Appl. Probab* (1964).
147. Kronja, I. et al. Widespread changes in the posttranscriptional landscape at the *Drosophila* oocyte-to-embryo transition. *Cell Rep.* **7**, 1495–1508 (2014).
148. Dunn, J. G. & Weissman, J. S. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genom.* **17**, 958 (2016).
149. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
150. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. R package (2019).
151. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
152. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
153. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 31 (2005).
154. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
155. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
156. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
157. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
158. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat Biotechnol.* <https://doi.org/10.1038/nbt.4236> (2018).
159. Chen, C., Li, Z., Huang, H., Suzek, B. E. & Wu, C. H. A fast Peptide Match service for UniProt Knowledgebase. *Bioinformatics* **29**, 2808–2809 (2013).
160. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
161. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
162. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
163. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

## Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (No. 91731301) and the Ministry of Science and Technology of the People's Republic of China (2016YFA0500800) awarded to J.L. and the China Postdoctoral Science Foundation (2019M650003) to Y.W. H.Z. and Y.W. are supported by grants from the National Postdoctoral Innovative Talents Supporting Program. Some of the analyses were performed on the High-Performance Computing Platform of the Center for Life Sciences. The authors thank the National Center for Protein Sciences at Peking University in Beijing, China, for the assistance with the analysis of mass spectrometry data.

## Author contributions

J.L. supervised the entire project and conceived and designed the research. H.Z., Y.W., X.W., X.T., and C.W. contributed to the data analyses. X.W. and X.T. performed the experiments. J.L., H.Z., and Y.W. wrote the manuscript.

## 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/s41467-021-21394-y>.

**Correspondence** and requests for materials should be addressed to J.L.

**Peer review information** *Nature Communications* thanks Pavel Baranov and the other anonymous reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission 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 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021, corrected publication 2021