The evolution of 5′ UTR length was shaped by natural selection for its functional impact on gene expression

Yu Zhan[1,2], Zhenbin Hu[1], Zhaolian Lu[1] and Zhenguo Lin[1*]

1. Department of Biology, Saint Louis University, St. Louis, MO 63103, USA
2. Department of Experimental Research, Bioinformatics Platform, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, P. R. China

[*]To whom correspondence should be addressed
Zhenguo Lin: zhenguo.lin@slu.edu

Keywords: 5′ UTR length, evolution, gene expression, TSS, CAGE

Running Title: Evolution of 5′ UTR length

ABSTRACT

The 5′ UTR of mRNA transcripts differs in length by three orders of magnitude among genes within a eukaryotic genome. Many genes also generate transcript isoforms with heterogeneous lengths of 5′ UTR by transcription from distantly located transcription start sites (TSSs). A null model was proposed that the diversification of 5′ UTR length is largely stochastic attributing to random gain or loss of TSS and 5′ UTR length evolved neutrally by genetic drift ("the null model"). Yet, multiple studies reported observations that are inconsistent with the null model. We conducted integrative analyses of multi-omics data to revisit the functional significance and evolution of 5′ UTR length. Our results reveal a strong negative correlation between transcripts' 5′ UTR lengths and their expression levels. Regulatory and evolutionary changes of 5′ UTR lengths are also negatively associated with changes in expression levels. We proposed that the fixation of a 5′ UTR length variant has been shaped by nature selection to better meet different expression demands among genes due to a negative impact of long 5′ UTR on translation efficiency. As the expression demands are highly diverse among genes and dynamic in response to environmental changes, different lengths of 5'UTR were favored by natural selection during evolution, creating a highly diverse landscape of 5'UTR lengths. Our model improves understanding of the evolution of gene structure in eukaryotes, and supports that ATU might facilitate fine-tuning of gene expression outcome by producing different lengths of 5′ UTR to adjust translation efficiency.

INTRODUCTION

The 5′ untranslated region (5′ UTR) of eukaryotic mRNAs plays important roles in regulating gene expression by affecting mRNA's stability, location, and translation efficiency (van der Velden and Thomas 1999; Bashirullah, et al. 2001; Mignone, et al. 2002). It is generally believed that the functional impacts of 5′ UTR are mainly mediated by its sequence elements and structural features (Sobczak and Krzyzosiak 2002; Ding, et al. 2012; Wang, et al. 2016; Guo, et al. 2017; Leppek, et al. 2018; Uppala, et al. 2022). One well-known 5′ UTR sequence element is the internal ribosome entry site (IRES), which enables the direct recruitment of the ribosome to the AUG initiation codon (Kozak 2005; Hinnebusch, et al. 2016). Some sequence elements in 5′ UTR may negatively impact gene expression. For instance, the presence of translation start codon AUG in 5′ UTR (upstream AUG, or uAUG) may result in dysfunctional proteins and significantly attenuate protein production from the gene (Dvir, et al. 2013) or produce protein isoforms with an extra N-terminus (Hinnebusch, et al. 2016). If a stop codon is present following the uAUG in the same reading frame in the 5′ UTR region, it forms an upstream ORF (uORF) that can suppress the translation of the downstream main ORF (Calvo, et al. 2009; Hinnebusch, et al. 2016). The formation of thermodynamically secondary and tertiary structures, such as RNA G-quadruplexes, may inhibit the scanning process of ribosomes and translation initiation (Leppek, et al. 2018). Empirical studies supported that elongated 5′ UTRs reduce translational efficiency compared to a shorter isoform (Araujo, et al. 2012; Wang, et al. 2016).

Although a single nucleotide of 5′ UTR was shown to be sufficient for the initiation of translation in mammals based on an *in vitro* study (Hughes and Andrews 1997), many eukaryotic genes are found to produce much longer 5′ UTRs than expected. For example, the longest known 5′ UTR in invertebrates is 4,498 nucleotide (nt) (Mignone, et al. 2002). In the fission yeast *Schizosaccharomyces pombe*, which has a highly compact genome of 12 million base pairs (bp), the length of 5′ UTR can exceed 2,000 nt (Rhind, et al. 2011). An overly long 5′ UTR increases the probability of occurrence of sequence and structural elements that may negatively impact gene expression. Intriguingly, overly long 5′ UTRs have been maintained in many genes, resulting in three orders of magnitude differences in 5′ UTR lengths within a genome. A null model was proposed that the large variations of 5′ UTR lengths are primarily driven by random mutations that resulted in turnovers of transcription start site (TSS) and fixation of various lengths of 5′ UTR is determined by stochastic processes, such as genetic drift ("the null model") (Lynch, et al. 2005). According to the null model, 5′ UTR length is selectively

neutral and the main constraint factor that may limit the length of 5′ UTR is the increase of uAUG (Lynch, et al. 2005).

As the null model predicts that the distribution of 5′ UTR is largely stochastic, it provides a plausible explanation for the 1000-fold differences in 5′ UTR lengths among genes. If the distribution of 5′ UTR length is stochastic, we would expect that 5′ UTR lengths are independent of gene expression levels or functions. However, multiple observations have been reported contradicting such predictions. Rao et al. examined two datasets of expressed sequence tag (EST) for various tissues in chicken and observed a weak negative association ($r = -0.1356$) between gene expression levels and 5′ UTR lengths (Rao, et al. 2013). A study used a high-density oligonucleotide tiling array mapped UTR boundaries for 2,044 genes in a budding yeast *Saccharomyces cerevisiae* and showed large variations of 5′ UTR length among genes with different functional categories (David, et al. 2006). In addition, elongation in 5′ UTR length was linked to the increased gene expression plasticity under different growth conditions during the evolution of budding yeasts based on RNA-seq data (Lin, et al. 2010; Lin and Li 2012).

The null model also predicts that species with larger effective population sizes are expected to have shorter 5′ UTRs because selection more efficiently prevents the accumulation of deleterious elongations (Lynch, et al. 2005). On the contrary, Chen et al. have shown that fruit fly *Drosophila melanogaster* has notably longer 5′ UTRs than that in many vertebrates despite having a larger effective population size (Chen, et al. 2011). Furthermore, as a higher GC content reduces the probability of AUG codon, a positive correlation between GC content and 5′ UTR length should be expected according to the null model. However, Reuter et al. found a significant quantitative difference between empirical data and the prediction based on the null model, indicating that non-stochastic factors may play a role in determining the evolution and distribution of 5′ UTR lengths (Reuter, et al. 2008).

In addition to their heterogenicity among genes, 5′ UTR lengths may vary substantially among transcript isoforms generated by the same gene because transcription of a gene can be initiated from distantly located TSSs (Brown, et al. 1999; Mignone, et al. 2002; Steinthorsdottir, et al. 2004; Pozner, et al. 2007; Shephard, et al. 2007; Rhind, et al. 2011; Wang, et al. 2016; Zhang, et al. 2017; Lu and Lin 2019; Policastro and Zentner 2021). Nearby TSSs can be computationally grouped as a TSS cluster (TC), representing the location of a core promoter, where transcription preinitiation complex (PIC) assembles and initiates transcription (Carninci, et al. 2006; Valen, et al. 2009; Lu, et al. 2021). Many genes contain multiple TCs or core promoters which can be hundreds of bp apart, creating transcript isoforms with large differences

in 5′ UTR lengths (Forrest, et al. 2014; Lu and Lin 2019). Furthermore, the relative usage of different TSSs of a gene can be highly dynamic in different cell types, developmental stages, and growth conditions, resulting in prevalent alternative TSS usage (ATU) (Polack, et al. 1993; Steinthorsdottir, et al. 2004; Pozner, et al. 2007; Shephard, et al. 2007; Zhang, et al. 2017; Lu and Lin 2019). Appropriate transcription initiation activities from specific TSSs or core promoters are important for normal cellular functions, because dysregulation in TSS usages was found to be associated with many human diseases, such as breast cancer, neurological disorders, and Alzheimer's disease (Sobczak and Krzyzosiak 2002; Mihailovich, et al. 2007; Tan, et al. 2007; Anvar, et al. 2018). For instance, transcription initiation of a tumor repressor gene, BRCA1, was found to be mainly initiated from a distal TSS in breast cancer tissues, which generates transcript isoforms with a longer 5′ UTR that resulted in 10-fold reductions of BRCA1 protein expression (Sobczak and Krzyzosiak 2002). Many researchers believed that the prevalence of ATU expands the diversity of transcriptome, which is considered a consequence of adaptive evolution (Landry, et al. 2003; Carninci, et al. 2006; Davuluri, et al. 2008; Lenhard, et al. 2012; de Klerk and t Hoen 2015; Policastro and Zentner 2021). Xu and colleagues argued that ATUs reflect imprecise transcription initiations and could be deleterious (Xu, et al. 2019). Nevertheless, both the adaptive and nonadaptive models of ATU would suggest that non-stochastic factors have influenced the fixation of novel TSSs and the evolution of 5′ UTR length.

Inaccurate 5′ UTR length and expression data may be among the factors that lead to conflicting observations and views regarding the evolution and function of 5′ UTR length. For example, in Rao et al's study, the expression level of a gene was calculated from pooled EST counts from different libraries (Rao, et al. 2013). EST data may or may not accurately determine the 5'end of a gene depending on library preparation strategies (Nagaraj, et al. 2007), and EST counts are not ideal for quantifying transcripts approach for insufficient depth (Kukurba and Montgomery 2015). In addition, this study was based on 4,728 genes, which accounts for ~26% of chicken protein-coding genes. One may argue that the weak correlation between 5′ UTR length and gene expression could be due to artifacts because of the inaccurate data and small sample size. Microarray or RNA-seq data were also known to suffer from the drawback of imprecise identification of the TSS and 5′ UTR length (Batut and Gingeras 2013; Steijger, et al. 2013). The assembly of RNA-seq reads usually extends transcript contigs to the last nucleotide (Wang, et al. 2010; Grabherr, et al. 2011), which can be used to infer the longest 5′ UTR of a transcript isoform, but not necessarily the most abundant isoforms. Furthermore, due to prevalent ATU, 5′ UTR lengths of a gene's transcripts may change substantially depending on cell types and growth conditions. Another bottleneck for these studies is the heterogeneous

nature of 5′ UTR lengths due to transcription initiation from distantly located TSSs of a gene (Andersson, et al. 2014; Forrest, et al. 2014; Lu and Lin 2019, 2021). Therefore, it is necessary to systematically re-evaluate the function and evolution of 5′ UTR length using more accurate and complete omics data. Single-nucleotide resolution of 5′ UTR length can be obtained by sequencing 5′end of RNA transcripts and mapping their sequencing reads to reference genome. One of the most popular 5′end profiling techniques is Cap Analysis of Gene Expression (CAGE) (Murata, et al. 2014; Morioka, et al. 2020). As CAGE read counts are proportional to transcriptional abundance, 5′ UTR length data based on CAGE data can be used to simultaneously quantify 5′ UTR length and transcription abundance for each transcript isoform, which is ideal to study the relationship between transcript 5′ UTR length and expression levels.

This study sought to re-evaluate the relationship between 5′ UTR length and expression and functional features of genes and revise the evolutionary model of 5′ UTR length based on a large collection of CAGE data. We primarily focused on the model organism *Saccharomyces cerevisiae* for two main reasons. First, CAGE-based TSS maps have been generated in *S. cerevisiae* in various grown conditions, allowing us to determine the regulatory changes of 5′ UTR length in response to environmental stimuli (Lu and Lin 2019, 2021). Second, as only ~ 5% of *S. cerevisiae* genes contain introns (Parenteau, et al. 2008), the paucity of introns minimizes the impacts of alternative splicing on the accuracy of 5′ UTR length data. We present multiple lines of evidence supporting that 5′ UTR length has a significant impact on the gene expression outcome by influencing translation efficiency. The evolution of 5′ UTR lengths has been shaped by natural selection due to its functional impact. Our model provides plausible explanations for the high heterogenicity of 5′ UTR length across genes and within a gene, contributing to a better understanding of the evolutionary dynamics of gene structure.

RESULTS

Transcript 5′ UTR lengths negatively correlate with their expression levels

If gene expression level was not considered as a constraint factor for 5′ UTR length according to the null model (Lynch, et al. 2005), the distribution of 5′ UTR length should be independent of gene expression levels, but it contradicts the observation in chicken (Rao, et al. 2013). We first sought to re-evaluate whether 5′ UTR length is independent of gene expression levels by addressing issues related to data accuracy and sample size using CAGE sequencing data of *S. cerevisiae* grown in nine different growth conditions obtained from (Lu and Lin 2019).

We generated a TSS quantitative map for each of the nine *S. cerevisiae* samples by mapping raw CAGE reads to its reference genome (see Methods and Materials, Supplementary Table S1-2). We observed that most genes have multiple clusters of TSSs upstream of their coding regions, which is consistent with studies in other organisms (Carninci, et al. 2006; Hoskins, et al. 2011; Haberle, et al. 2014; Lu and Lin 2021). A TSS cluster reflects the location of a core promoter (Frith, et al. 2008), indicating that most genes are controlled by multiple core promoters. For instance, two major clusters of TSS signals are present upstream of the coding sequences (CDS) of the UTP5 gene (YDR398W) in *S. cerevisiae* cultured in the rich YPD medium (Fig. 1A). Here, we denoted the distal core promoter as C1, and the proximal one as C2. mRNA transcripts produced from C1 and C2 differ in 5′ UTR lengths by ~350 nt. Thus, neither C1's nor C2's 5′ UTR length accurately depicts the 5′ UTR length of UTP5's mRNAs. To address this issue, we used individual transcript isoforms, instead of genes, as examining units considering that transcription activity from one core promoter is independent of another. The 5′ UTR length of transcripts generated from a core promoter was calculated as the distance between the dominant TSS (the TSS with the strongest transcription initiation activities; i.e., the largest TPM value) and the start of CDS. As a single CAGE read/tag was generated for each RNA transcript, the expression level of each transcript isoform was calculated as the numbers of CAGE reads/tags mapped to a core promoter normalized by millions of total mapped reads in a sample (tag per million, TPM). Thus, the CAGE TPM value is equivalent to transcript per millions of transcripts (TPM) in RNA-Seq studies.

For yeast cells cultured in the YPD medium, we identified 6,777 core promoters for 5,212 protein-coding genes (Supplementary Dataset S1). By plotting all transcript 5′ UTR lengths against their $\log_2$TPM values, we observed a significant negative correlation between the two features (the Pearson correlation coefficient $r = -0.40$, $P = 4.67\times10^{-262}$, Fig. 1B), which is much stronger than observed in chicken (Rao, et al. 2013). Our results show that transcript isoform's 5′ UTR length is a strong indicator of its expression level, suggesting that the distribution of 5′ UTR lengths is unlikely stochastic.

Another factor that should be taken into consideration is that the expression level of a transcript isoform or gene is highly dynamic responding to environmental changes or in different cell types. In the case of UTP5, transcription initiation activities from both promoters vary from sample to sample (Fig. 1A). Such difference is usually attributed to changes in regulatory input in response to environmental stimuli by activating or repressing a specific set of transcription factors. To exclude the possibility that the observed negative correlation is due to specific

regulatory effects, we performed similar analyses for quantitative TSS maps obtained in other eight *S. cerevisiae* samples (Supplementary Table S2, Supplementary Dataset S1). We observed similar strong negative correlations in all samples examined, with *r* ranging from −0.4 to −0.48 (Supplementary Fig. S1). These results support that the negative correlation between transcript 5′ UTR length and expression level is independent of regulatory inputs, supporting that the negative correlation is likely an intrinsic feature.

To further exclude the probability of observing the strong negative correlation by chance, we conducted 10,000 random permutation tests. In each permutation test, the $\log_2$TPM value of a transcript isoform was randomly assigned to another one, while its 5′ UTR length remained unchanged. For each permutation dataset, the Pearson correlation coefficient *r* between 5′ UTR length and $\log_2$TPM of all transcripts was then calculated. No strong negative correlation was observed in any of these permutation datasets, as their *r* values are very close to 0, with a standard deviation $\sigma = 0.014$ (Supplementary Fig. S2). With a Z-Score = −33.36 ($P < 0.01$), we can confidently exclude the possibility of a stochastic effect, supporting non-random distribution of 5′ UTR length among genes.

We next sought to determine whether the negative correlation between 5′ UTR length and expression level can also be observed in other organisms. We conducted similar studies using CAGE data obtained from five other organisms, including two other budding yeasts (*Saccharomyces paradoxus* and *Naumovozyma castellii*), a fission yeast *Schizosaccharomyces pombe*, mouse *Mus musculus*, and human *Homo sapiens* (see Materials and Methods). The correlation coefficient *r* ranges from −0.3 to −0.44, supporting that the strong negative correlation between 5′ UTR length and transcription levels is likely a common feature of eukaryotic genes (Supplementary Fig. S3 and Supplementary Dataset S2).

Transcript 5′ UTR length is a stronger indicator for its peak expression level

Given that the transcription level of an mRNA isoform may differ significantly from one growth condition to another due to changes in regulatory inputs, it is necessary to mitigate the impacts of these regulatory inputs to more accurately portray the relationship between 5′ UTR length and transcription levels. One feasible way is to use the highest expression level (peak expression level) of a transcript isoform in all examined samples, representing its activated status or highest expression demand for a transcript to minimize the influence of different regulatory inputs. For instance, the C1 isoform of UTP5 has its peak expression level (10.8 TPM) under cell cycle arrest triggered by α factor, while its C2 isoform reached its peak expression level (22.3 TPM) after treating with H_2O_2 (Fig. 1A). In total, we obtained peak expression values

for 9,873 transcripts isoforms in *S. cerevisiae* based on their largest TPM values in the nine growth conditions (Supplementary Dataset S3). Based on the peak expression values, we observed a stronger negative correlation ($r = -0.5$, $P = 9.88\times10^{-324}$) between transcript 5′ UTR length and expression level than any individual samples examined (Fig. 1C), supporting that 5′ UTR length is indeed a strong indicator for transcript expression level, particularly its peak level.

If a shorter 5′ UTR indicates a higher expression level at a genome scale, one would predict that a gene's shorter isoform tends to have a higher expression level than its longer isoform. To test this prediction, we compared the peak expression level of transcript isoforms for 2,672 genes which have at least two core promoters in *S. cerevisiae* (Supplementary Dataset S4). If more than two core promoters are present in a gene, only the two with the largest TPM values were considered, as they usually account for most transcription signals. Among these genes, two-thirds (67.5%) of transcript isoforms with a shorter 5′ UTR have a higher peak TPM value than their longer isoforms ($P < 2.2\times10^{-16}$, binomial test, Supplementary Fig. S4A), further supporting that 5′ UTR length as an indicator for its expression level. It is also reasonable to predict that a larger difference in 5′ UTR length between transcript isoforms should be associated with a larger difference in their peak TPM values. By plotting the $\log_2$ ratio of 5′ UTR length [$\log_2$(5′ UTR_1/5′ UTR_2)] between two transcript isoforms to that of peak expression level [$\log_2(TPM_1/TPM_2)$], we observed a strong negative correlation ($r = -0.48$, $P = 5.69\times10^{-321}$, Supplementary Fig. S4B) between the two features. These results suggest that even though different genes may significantly differ in their expression levels, it is a general feature that their higher-expressed transcript isoform tends to have shorter 5′ UTR.

Non-random distribution of 5′ UTR lengths among genes of different functional groups

The null model would predict that the distribution of 5′ UTR length should be largely random among genes of different functions as gene function is not a constraint factor either. Our observations would predict otherwise because it has been well-documented that different functional groups of genes demonstrate highly diverse expression levels. To test this hypothesis, we examined the distributions of 5′ UTR lengths among genes of different functional categories in *S. cerevisiae*. Gene functional classifications were obtained based on annotation information from Gene Ontology (GO) (Ashburner, et al. 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, et al. 2002). To mitigate bias due to small sample sizes, only KEGG or GO groups with more than 50 gene members were used for subsequent analyses.

Considering that many genes produce multiple transcription isoforms, the 5′ UTR length of a gene was calculated as its weighted average of all isoforms' 5′ UTR lengths

(Supplementary Dataset S5). Our results show significant variations in 5′ UTR lengths among KEGG or GO groups (Fig. 2A, Supplementary Fig. S5, ANOVA test, P < 0.0001). For example, the median 5′ UTR length of genes in the KEGG group of “MAPK signaling pathway” is 128 nt, which is 4.57× as long as genes encoding ribosomal proteins (28 nt). To determine the probability of observing a difference in median 5′ UTR length between the two groups that is equal to or more extreme than what was observed, we conducted 10,000 permutation tests (see Methods and Materials). The ratio between the median 5′ UTR length between the two groups was calculated for each permutated dataset. As expected, most permutated datasets yielded a ratio that is very close to 1 (Supplementary Fig. S6). The probability of observing 4.57× differences in median 5′ UTR length by chance is extremely low based on the distribution of permutated ratios (Z-Score = 29.33, $P < 0.0001$), supporting a non-random distribution of 5′ UTR length among different functional groups of genes.

On the other hand, if the distribution of 5′ UTR lengths is random, genes with short or long 5′ UTR in a genome are not expected to be significantly enriched in different functional groups. To test this hypothesis, we performed GO enrichment analysis for *S. cerevisiae* genes with long 5′ UTR length (> 75th percentile) and short 5′ UTR (< 25th percentile). We found that genes with short 5′ UTR tend to be significantly enriched in genes encoding ribosomal proteins and methyltransferase activity (Fig. 2B). These genes are generally considered as housekeeping genes, as they are activated in most tissues or growth conditions (Rouleau, et al. 1992; Butte, et al. 2001). In contrast, genes with long 5′ UTR are significantly enriched in those encoding transcription factors or kinase (Supplementary Fig. S7A), which are known to have dynamic expression profiles (Li, et al. 2015).

To test if the non-random distribution of 5′ UTR length is also present in other organisms, we conducted similar studies in the fission yeast *S. pombe* using CAGE data obtained from (Lu and Lin 2021) (Supplementary Dataset S5). We observed similar patterns in *S. pombe* that genes related to cytoplasmic translation or ribosomal proteins tend to have shorter 5′ UTR than others, while genes encoding transcription factors are among those with the longest 5′ UTR (Supplementary Fig. S7B). These results support that the 5′ UTR lengths vary substantially among genes of different functional groups.

Changes in 5′ UTR length are associated with gene differential expression in response to environmental stimuli.

Our previous study demonstrated prevalent ATU in yeast in response to environmental stimuli (Lu and Lin 2019). Given that introns are almost depleted in *S. cerevisiae* and alternative

splicing is nearly non-existent, the only consequence of ATU is a change in 5′ UTR length in most genes. A shift of transcription initiation from a distal TSS to a proximal TSS generates transcripts with a shorter 5′ UTR, while a reverse shift elongates the transcript's 5′ UTR. For instance, transcription of MTD1 (YKR080W) predominantly initiates from its proximal promoter in yeast cells grown in the YPD medium (Fig. 3A). In contrast, with NaCl treatment, transcription initiation from the proximal core promoter was almost depleted, and transcription was mostly initiated from the distal promoter. As a result of ATU, the weighted 5′ UTR length MTD1 was increased from 77 nt to 150 nt with 2% NaCl treatment.

As our data suggests that 5′ UTR is a strong indicator for gene expression level and ATU changes 5′ UTR length, it is reasonable to predict that: 1) genes with ATU in response to environmental stimuli are more likely to be associated with gene differential expression; 2) shortening of 5′ UTR of a gene is more likely to be associated with upregulation of gene expression (negative correlation), while elongation of 5′ UTR are enriched with downregulation of gene expression (positive correlation). To test these hypotheses, we first identified genes with ATU by comparing each TSS map of *S. cerevisiae* cells with eight different treatments to samples to that of rich YPD medium (control). Here, ATU was defined as significant changes in the distributions of TSS signals in the promoter regions of a gene, and the dominant TSS of a gene shifted to another one that is at least 50 bp away in treatment compared to the control (see Materials and Methods). Genes with differential expression (DE) were identified by DESeq2 based on raw CAGE read counts associated with each gene.

Among 5,512 genes examined, the number of genes with ATU ranges from 109 to 188 in the eight samples and the number of DE genes ranges from 705 to 2,414 (Supplementary Dataset S6). As predicted, genes with ATU are much more likely to be associated with DE in each treatment. In all eight treatments, the proportion of genes with ATU associated with DE are significantly higher than those without ATU (Fig. 3B). For instance, with heat shock treatment, 141 of 188 genes with ATU (75%) are associated with DE, while only 33% of other genes are associated with DE, suggesting that ATU might function as a mechanism to facilitate regulation of gene expression in response to environmental changes. Furthermore, our results also support that shortening of 5′ UTR is more likely to be associated with gene upregulation (negative correlation) and vice versa (Fig. 3C). For instance, associated with elongation of 5′ UTR after treatment of 2%, the transcription level of MTD1 is significantly downregulated from 118.06 TPM to 31.95 TPM (Fig. 3A). Under oxidative stress (H_2O_2 treatment), 88% of genes with ATU demonstrate a negative correlation between 5′ UTR length and expression level, that

is a shortening of 5′ UTR is associated with gene upregulation and elongation of 5′ UTR is coupled with gene downregulation. Overall, there is a significant negative correlation between $\log_2$ fold changes of 5′ UTR length to $\log_2$ fold changes of TPM values in all comparisons (r = −0.39, $P = 1.2\times10^{-12}$, Fig. 3D).

To determine if ATU is also associated with gene differential expression in different cell types in multicellular organisms, we conducted similar analyses using CAGE data of two human cell lines K562 and HepG2. We identified 752 genes with ATU between the two cell lines (Supplementary Dataset S7). Differentially expressed genes were identified as fold changes ≥ 2 with FDR < 0.01. Similar to *S. cerevisiae*, genes with ATU are more likely to be associated with gene differential expression than other genes (*P*-value < 0.001, exact binomial test, Supplementary Fig. S8A). Furthermore, the proportion of genes with a longer weighted 5′ UTR length between the two cell lines is much higher to have a lower expression level than that with a high expression level (*P*-value < 0.001, exact binomial test, Supplementary Fig. S8B), which is consistent with the observation in yeasts. Moreover, a strong negative correlation between fold changes of 5′ UTR length and expression level between the two human cell lines K562 and HepG2 was also observed (Supplementary Fig. S8C). The consistent patterns observed in both yeast and human samples support that the association of ATU and gene differential expression is conserved between distantly related eukaryotic organisms. These findings also suggest that ATU might contribute to the fine-tuning of gene expression during post-transcriptional processes by changing 5′ UTR length.

Evolutionary changes in 5′ UTR length are associated with divergence in gene expression level during the evolution of yeast species.

If 5′ UTR length is a strong indicator of a gene's expression level, it is also reasonable to postulate that evolutionary change of 5′ UTR length via gain/loss of TSSs in promoter regions is likely to be associated with divergence of gene expression level. To test this hypothesis, we identified 4,872 one-to-one orthologous gene groups from two closely related yeast species, *S. cerevisiae* and *S. paradoxus* (Supplementary Dataset S5 and S8). The 5′ UTR length and expression level of each pair of orthologous genes were obtained based on the CAGE sequencing data from cells cultured in rich YPD medium in both species (Lu and Lin 2019, 2021). Orthologous genes that differ in 5′ UTR length by at least 50 nt and expression level with $\log_2$ FC ≥ 2 were identified for subsequent analysis.

We found that a gene with shorter 5′ UTR lengths in either species tends to higher expression level than its orthologous gene in the other species with longer 5′ UTR. For example,

the AAD10 gene (YJR155W) in *S. cerevisiae*, which encodes putative aryl-alcohol dehydrogenase, is expressed with 2.23 TPM, and has a weighted 5′ UTR length of 332 nt (Fig. 4A). Its orthologous gene in *S. paradoxus* has a much higher expression level (17.15 TPM), but their RNA transcripts have a much shorter 5′ UTR (54 nt) (Fig. 4A, Supplementary Dataset S8). At a genome scale, among all orthologous groups with a significant difference in gene expression levels, genes with a lower expression level are more likely to be associated with a longer 5′ UTR and vice versa (Fig. 4B, $P < 0.001$, t-test). Overall, there is a negative correlation ($r = -0.34$, $P = 6.71\times10^{-10}$) between evolutionary changes of 5′ UTR length and divergence of gene expression levels (Fig. 4C). All these observations support that changes in 5′ UTR length might have been associated with the evolutionary divergence of gene expression. These observations further support that 5′ UTR length is strongly associated with a gene's expression level.

A revised evolutionary model of 5′ UTR length provides a plausible explanation for the negative correlation between 5′ UTR length and expression level

Our studies provide multiple lines of evidence supporting a strong negative correlation between transcript 5′ UTR length and expression levels. As a product of transcription, it is unlikely that 5′ UTR in an mRNA transcript directly influences gene expression in the transcriptional process. As it is well-documented that long 5′ UTRs tend to have a negative impact on the translational process for its sequence elements (Brown, et al. 1999; Sobczak and Krzyzosiak 2002; Sonenberg and Hinnebusch 2009; Waern and Snyder 2013; Wang, et al. 2016; Chen, et al. 2017; Tresenrider, et al. 2021) and structure features (Tuller, et al. 2010; Pop, et al. 2014), it makes perfect sense from an evolutionary perspective that how 5′ UTR length negatively correlates with gene expression level. Here, we proposed a revised model depicting the evolution and divergence of 5′ UTR length to provide plausible explanations for all these observations (Fig. 5).

New TSSs may appear in the promoter region of a gene due to stochastic mutations. Our previous study suggested a single transversion mutation that demolishes the pyrimidine-purine (PyPu) at the TSS dinucleotides may abolish most transcription initiation activities from the TSS (Lu and Lin 2021). Therefore, gain or loss of a TSS may occur due to a single nucleotide substitution. Based on the observations above, it appears that the fixation of novel TSSs is largely non-stochastic. We argue that natural selection has played an important role in determining the evolutionary fate of a novel TSS due to its impact on the translation processes. Depending on the expression demand of a gene, a specific 5′ UTR length has different fitness

impacts. Specifically, if a new TSS is located downstream of the existing TSS, it generates a new transcript isoform with shorter 5′ UTRs, which tends to produce protein more efficiently because it reduces the probability of forming a more complex or stable secondary structure. If the gene encodes proteins with high demand, such as ribosomal proteins, an increase in translation efficiency provides selective advantages and subsequently increases the fitness of the organism. The selective advantage would be more significant if the transcription level of such genes is already saturated, as it provides a new way to increase protein production to meet the demand of cells. As a result, the frequency of the allele that generates short 5′ UTR would be subsequently increased in the population and eventually fixed. In contrast, a mutation that creates a distal TSS would be favored for genes encoding proteins with low demand, as it reduces translation efficiency, which is likely beneficial as it facilitates further repression of gene expression under most conditions. Meanwhile, existing TSS or other non-favorable TSSs may be eventually lost during evolution due to the accumulation of mutations. In summary, our model proposes that natural selection plays an important role in shaping the landscape of 5′ UTR length in a genome.

Multiple lines of evidence support the selection model of 5′ UTR length evolution

Our model assumes that 5′ UTR length may influence gene expression outcome due to its impact on translation efficiency. A longer 5′ UTR increases the number of uAUG and uORF, as well as the probability of forming a more complex secondary or tertiary structure. All these features may reduce translation efficiency and production of normal protein products. This assumption has its theoretical ground and is also supported by empirical data based on high-throughput studies. We examined the relationship between 5′ UTR length of transcript isoforms and transcription efficiency using published datasets obtained from mouse fibroblasts (Wang, et al. 2016) (see Materials and Methods). Our results show that 80% of transcript isoforms (925 out of 1162) with shorter 5′ UTR have a higher translation efficiency (see Methods and Materials, Fig. 6A and supplementary Dataset S9). Furthermore, a strong negative correlation ($r = -0.60$, $P = 5.12 \times 10^{-113}$) between the differences in transcript 5′ UTR length and their translation efficiency (Fig. 6B and Supplementary Dataset S9) supports that the effect of 5′ UTR length on translation efficiency is largely linear.

Our model proposes that natural selection favored TSSs that generate an optimal 5′ UTR length for a gene depending on expression demands. The scatter plots between 5′ UTR length and transcription level show that most genes in eukaryotic genomes are likely equipped with an optimal or near-optimal length of 5′ UTR (Fig. 1B-C). However, the scatter plots also

showed that some genes may not have a 5′ UTR length that best meets their expression level based on the linear regression model (outlier genes). For example, some genes may have a short 5′ UTR, but have a much lower expression level than other genes with a similar 5′ UTR length. This could be due to the absence of activation signals for short 5′ UTR genes in the growth conditions examined. Another factor could be the lack of mutations in the promoter region to produce an appropriate length of 5′ UTR, because the availability of desired mutations was considered a limited factor for evolution (Nei 2014). Therefore, we predict that genomic variants are less likely to accumulate in the promoter regions of genes with optimal 5′ UTR length, while more genomic variants can be found in genes with non-optimal 5′ UTR length.

To test this prediction, we used a simple linear regression model to identify genes with non-optimal 5′ UTR length (outlier genes). The linear relationship between transcript 5′ UTR lengths and their peak expression levels ($\log_2$TPM) can be described as $\log_2\text{TPM} = -0.0095\times L+6.46$, where L is the 5′ UTR length (Supplementary Fig. S9, and Supplementary Dataset S10). Genes without an optimal 5′ UTR length are defined as those outside the 95% prediction interval for the predicted $\log_2$TPM values given a 5′ UTR length based on the linear model. We identified a total number of 97 outlier genes that were considered without an optimal 5′ UTR length. Among them, only 3 genes have a much longer 5′ UTR given their high expression level, while 94 of them have a much shorter 5′ UTR given their low expression level. If a lack of optimal 5′ UTR length is solely due to lack of needed mutations, we would expect similar numbers of genes with overly long or short 5′ UTR. As 97% of genes appear to be equipped with a shorter 5′ UTR length given their observed expression level, the observed expression level may not represent their true peak expression level because only limited numbers of growth conditions were examined.

We then retrieved genomic variant data from whole genome sequencing for over 1000 *S. cerevisiae* isolates (Peter, et al. 2018). We identified the number of single nucleotide polymorphisms (SNPs) in the promoter regions for each gene, which was defined as 750 bp upstream of the main ORF. Considering that many genes contain multiple core promoters, and most of them are separated by less than 300 bp, it will result in overlapping regions. To exclude potential bias due to overlapping promoters, we only used genes with a single core promoter for subsequent studies. On average, 62.4 SNPs are found in the promoters of outlier genes, which is significantly more than 52.4 SNPs in the other genes ($P = 3.13\times10^{-5}$, t-test, Fig. 6C). Assuming stochastic mutations occur at a similar rate across the yeast genome, a higher density of SNP in the promoter regions of outlier genes suggests that variants in these genes

are less likely to be eliminated by natural selection compared to genes with optimal 5′ UTR lengths, indicates a lower selection force acting on the promoters of these outlier genes. To further test our conclusion, we compared nucleotide diversity (π) of promoter regions between the two types of genes. As expected, the outlier genes have a much higher average nucleotide diversity (π = 4.00±2.88) than that of other genes (π = 3.11±1.97, $P = 9.85\times10^{-6}$, t-test, Fig. 6D), supporting a more relaxed selection force acting on the promoter regions of outlier genes to allow a higher degree of polymorphism.

DISCUSSION

In this study, we systematically examined the relationship between 5′ UTR length and gene expression level based on high-resolution quantitative TSS maps from various species. We found a strong negative correlation between transcript 5′ UTR length and expression level in both unicellular and complex eukaryotic organisms. In addition, changes in 5′ UTR lengths in responding to environmental stimuli and during the evolution of yeast are also coupled with changes in gene expression level. A negative correlation also exists between 5′ UTR length and translation efficiency. These observations support an impact of 5′ UTR length on gene expression, and consequently, the evolution of 5′ UTR length is unlikely selectively neutral. We proposed that the evolution of 5′ UTR length was shaped by nature selection due to its functional impact on translational processes. This model is also supported by the simulation study based on GC content (Reuter, et al. 2008).

The natural selection model of 5′ UTR length provides a plausible explanation for our observations. A shorter 5′ UTR provides selective advantages for the expression of high-demand genes, such as those encoding ribosomal proteins, by increasing its translation efficiency (Fig. 5). In contrast, for genes with low demand, a longer 5′ UTR is advantageous as it reduces translation efficiency, further decreasing its protein production. During evolution, different lengths of 5′ UTR were favored by natural selection depending on the expression demand of genes. Our results also support that ATU might serve as a mechanism to fine-tune the expression of genes by increasing or reducing translation efficiency via changing 5′ UTR length, providing new insights into understanding the complexity of gene regulatory mechanisms.

Similar to the null model, our selection model of 5′ UTR length also well explains the 1,000-fold range of 5’ UTR lengths of genes in a genome. The demand for protein production varies substantially among genes within a genome. Some housekeeping genes are in high

demand in most growth conditions or cell types (David, et al. 2006). In contrast, the expression demand for many other genes, such as transcription factor or stress-response genes, are low in most conditions, a longer 5′ UTR would be beneficial for maintaining the low abundance of protein products for these genes. As natural selection has shaped 5′ UTR lengths of genes with different expression demands, it created a landscape of highly diverse 5′ UTR lengths in a genome.

It is worth noting that our model provides explanations for some observations that cannot be explained by the null model. First, our model well explains the negative correlation between the 5′ UTR length and transcription levels. Our model suggests that 5′ UTR length reflects different demands of gene products. Although 5′ UTR length, as a structural feature of transcription product, is not involved in the control of transcription activities, it contributes to fine control of the translation process to further increase or decrease protein production of specific genes. As a result of natural selection, proper 5′ UTR lengths have been selected to better meet gene expression demand. Because genes of different functional groups have very different expression demands, our model explains why the 5′ UTR length varies substantially among functional groups of genes (David, et al. 2006). In addition, due to a lack of suitable mutations (Nei 2014) or other compensatory factors, it is also expected that many genes are not equipped with a 5′ UTR length that best fits their expression level.

Furthermore, our model suggests natural selection, instead of genetic drift, plays a major role in shaping the landscape of 5′ UTR length due to its functional impact on the translation process. Therefore, the functional consequence of a variant that creates a new TSS and 5′ UTR length, rather than the population size, is a leading factor that determines the evolutionary fate of a new TSS. It explains why some organisms with a large population size have longer 5′ UTR than those with a smaller population size (Chen, et al. 2011). In addition, our model also explains why the positive relationship between GC content and 5′ UTR length predicted by the null model is not consistent with empirical data (Reuter, et al. 2008). This is because the expression demand of a gene, instead of uAUG, is likely the most important constraint factor.

We obtained a stronger negative correlation with 5′ UTR length by using the peak expression levels of transcripts based on CAGE data obtained in nine growth conditions. Considering that the nine growth conditions only represent very small portions of growth environments, the peak expression levels we obtained may not necessarily be accurate for all transcript isoforms. With a significant increase of examined samples, we expect to obtain true peak expression levels for more transcripts and thus the negative correlation between the 5′

UTR length and expression level could be even stronger than observed in this study. Therefore, we may still underestimate the impact of 5′ UTR length on expression outcomes of genes and its roles in the evolution of gene structure.

MATERIALS AND METHODS

Data sources

Raw CAGE sequencing reads of *S. cerevisiae* (Lu and Lin 2019), *S. paradoxus*, *N. castellii,* and *S. pombe* (Lu and Lin 2021) were downloaded from the NCBI SRA database (Supplementary Table S1). CAGE data of human K562 and HepG2 cell lines each contains three replicates and are highly consistent with each other (r ≥ 0.99) and thus were used to study the relationships between 5′ UTR length and expression levels in human. Mapped CAGE reads in bam format for human K562 and HepG2 cell lines were downloaded from FANTOM5 (Lizio, et al. 2015). 5′ UTR length, mRNA abundance, and translation efficiency of transcripts from mouse fibroblasts (NIH3T3) were obtained from (Wang, et al. 2016).

Generation of quantitative TSS maps based on CAGE data

Quantitative TSS maps were generated by aligning raw CAGE sequencing data to the reference genome of each species (Assembly version: sacCer3 for *S. cerevisiae*, N17 for *S. paradoxus*, S. pombe for *S. pombe*). HISAT2 (Kim, et al. 2019) was used for sliced read alignment by disabling its '–no-softclip' option to avoid false-positive TSSs. Only uniquely mapped reads (mapping quality score MAPQ > 20) were used for TSS identification. We used the TSSr package (Lu, et al. 2021) to call TSSs from mapped read files. TSS read counts from biological replicates of the same sample were merged for quantification of TSS signals, which were quantified as the numbers of CAGE tags (reads) supporting the TSS per million mapped reads (TPM). Nearby TSSs were grouped as a TSS cluster (TC), representing a core promoter using the 'peakclu' function of TSSr (parameters are 'peakDistance = 50, extensionDistance = 25, localThreshold = 0.1, clusterThreshold = 1'). A set of consensus TCs were generated for the nine samples of *S. cerevisiae*, and the two cell lines of human, respectively, by using the 'consensusCluster' function in TSSr with an option of 'dis = 50'. The transcription level of a TC was calculated as the sum of TPM values of all eligible TSSs within it. Only TCs with TPM ≥1 are considered as biologically significant and used for further analyses. For yeast species, consensus TCs were then assigned to their downstream genes if they are within 750 bp upstream and 50 bp downstream of the start codon of annotated ORFs. TCs associated with

non-coding RNA genes in *S. cerevisiae* (e.g., SUTs, XUTs, CUTs, tRNA, rRNA and snoRNAs, Supplementary Dataset S11) were identified using parameters “upstream: 100, upstreamOverlap: 50, downstream: 25”. Consensus TCs in human were assigned to their downstream genes if they were within 100 bp upstream and 50 bp downstream of the annotated TSSs.

Calculation of 5′ UTR length and identification of genes with non-optimal 5′ UTR length

The 5′ UTR length of a transcript isoform was calculated as the distance between its dominant TSS, which is the TSS with the largest TPM values within a TC, and the translation start codon of the associated gene (Supplementary Dataset S1). Transcripts with introns in 5′ UTR were excluded from our analysis to avoid potential biases. If a TC’s dominant TSS is located within a gene coding region, which creates a negative 5′ UTR length, the TC was excluded for subsequent analysis. The 5′ UTR length for a gene was calculated as the weighted average of the 5′ UTR length of all transcript isoforms as described in (Lu and Lin 2019) (Supplemental Dataset S5). To infer a list of genes with non-optimal 5′ UTR length (outlier gene), a linear regression model for weighted 5′ UTR length and their TPM values was first constructed. Genes without an optimal 5′ UTR length are defined as those outside the 95% prediction interval for the predicted $\log_2$TPM values given a 5′ UTR length based on the linear model.

Random simulation analyses

We carried out random simulation analysis to calculate the probability of observing a strong negative correlation between 5′ UTR length and gene expression levels. In each random simulation, the TPM value of a transcript isoform was randomly assigned to another transcript, while its 5′ UTR length remained unchanged, to generate a pseudoreplicate dataset. A total number of 10,000 pseudoreplicate datasets were generated. Pearson correlation coefficient (r) was then calculated for each pseudoreplicate dataset. The mean and standard deviation (σ) of r values obtained from the 10,000 datasets were calculated. As a normal distribution was observed from the 10,000 simulated r values, a Z-score and p-value were calculated to infer the probability of obtaining the observed r value.

Random simulation analysis was also carried out to infer the probability of observing large differences in median 5’UTR lengths between two groups of genes with a size of 109 (Group M) and 159 (Group R) respectively. In each simulation, 109 and 159 genes were randomly selected from a pool of 6929 *S. cerevisiae* genes and assigned to Group M, and

Group R, respectively. The median lengths of their 5′ UTR were calculated for each group based on the pseudoreplicate data. The ratio of the median 5′ UTR length of the two groups (Group M/Group R) was then calculated. The process was repeated 10,000 times to generate 10,000 pseudoreplicate ratios. The mean and standard deviation (σ) of the 10,000 ratios were calculated to infer the Z-score and P values for the observed ratio (4.57).

Gene functional classification and gene set enrichment analysis.

Functional classifications of genes were obtained from KEGG and Gene Ontology (GO) annotations. KEGG pathways and associated genes in *S. cerevisiae* and *S. pombe* were extracted from the KEGG REST server using the getGenesets function from KEGGREST (https://bioconductor.org/packages/KEGGREST). KEGG pathways with at least 50 genes were selected for further analysis. GO annotation file (biological process) of *S. cerevisiae* was downloaded from the Gene Ontology Resource (Ashburner, et al. 2000; Gene Ontology, et al. 2023). GO terms with gene numbers more than 50, but less than 500, were selected for subsequent analysis. Redundant GO groups were filtered using REVIGO with *C* = 0.4 (Supek, et al. 2011). R packages ggplot2 (Wickham 2009) and multcompView (https://rdrr.io/cran/multcompView/) were used for data visualization. GO enrichment analysis was performed using GO Term Finder (Boyle, et al. 2004).

Identification of alternative TSS usage (ATU) and differentially expressed genes (DEG)

An ATU event between two CAGE samples was defined as significant changes in the distribution of CAGE signals in the promoter regions of a gene and the distance of dominant TSS differed by more than 50 bp. Different distribution of CAGE signals was determined by paired samples Wilcoxon test, with a cutoff of $P < 0.05$ adjusted by the Benjamini-Hochberg procedure. Differentially expressed genes between two samples were identified using “deGene” function of TSSr (Lu, et al. 2021), which calculated the sum of CAGE read counts mapped to all core promoters of each gene, serving as input data for statistical inference of differentially expressed genes by DESeq2 (Love, et al. 2014).

Evolutionary analysis of population genomics data in *S. cerevisiae*

The single nucleotide polymorphism (SNP) dataset was retrieved from whole genome sequencing data of 1011 *S, cerevisiae* strains (Peter, et al. 2018). To ensure the integrity of the SNP dataset, multiallelic SNPs were removed using "bcftools" version 1.13 (Danecek, et al. 2021). In our analysis, a promoter SNP was defined as a TSS located within a 750 nucleotide (nt) region upstream of an ORF. The nucleotide diversity (π) for each SNP was computed

utilizing "vcftools" version 0.1.16, as outlined in (Danecek, et al. 2011). SNPs with a nucleotide diversity (π) exceeding the threshold of 0.004 were selected for further downstream analysis.

Statistical analysis

All statistical analyses were performed in R v.4.2.1. Statistical analysis information was included in the figure legends. For correlation analysis, the Pearson correlation coefficients (r) were performed. Binomial tests were employed to determine whether equal numbers of genes in two different patterns. A one-tailed t-test was employed to compare SNP number and π diversity in different gene groups. Scatter plots with density information and boxplots were generated using R packages ggplot2 (Wickham 2009) and ggpointdensity (https://github.com/LKremer/ggpointdensity).

ACKNOWLEDGEMENTS

This work was supported by the U.S. National Science Foundation (NSF 1951332) and the 2022 President's Research Fund from Saint Louis University to ZL.

REFERENCES

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507:455-461.

Anvar SY, Allard G, Tseng E, Sheynkman GM, de Klerk E, Vermaat M, Yin RH, Johansson HE, Ariyurek Y, den Dunnen JT, et al. 2018. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. Genome Biol 19:46.

Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, Burns SC, Penalva LO. 2012. Before It Gets Started: Regulating Translation at the 5' UTR. Comp Funct Genomics 2012:475731.

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

Bashirullah A, Cooperstock RL, Lipshitz HD. 2001. Spatial and temporal control of RNA stability. Proc Natl Acad Sci U S A 98:7025-7028.

Batut P, Gingeras TR. 2013. Rampage: Promoter Activity Profiling by Paired-end Sequencing of 5′-complete cDNAs. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] 104:Unit.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics (Oxford, England) 20:3710-3715.

Brown CY, Mize GJ, Pineda M, George DL, Morris DR. 1999. Role of two upstream open reading frames in the translational control of oncogene mdm2. Oncogene 18:5631-5637.

Butte AJ, Dzau VJ, Glueck SB. 2001. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". Physiol Genomics 7:95-96.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A 106:7507-7512.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626-635.

Chen C-H, Lin H-Y, Pan C-L, Chen F-C. 2011. The plausible reason why the length of 5' untranslated region is unrelated to organismal complexity. BMC Research Notes 4:312.

Chen J, Tresenrider A, Chia M, McSwiggen DT, Spedale G, Jorgensen V, Liao H, van Werven FJ, Ünal E. 2017. Kinetochore inactivation by expression of a repressive mRNA. Elife 6:e27417.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences 103:5320-5325.

Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang THM. 2008. The functional consequences of alternative promoter use in mammalian genomes. Trends in Genetics 24:167-177.

de Klerk E, t Hoen PA. 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet 31:128-139.

Ding Y, Shah P, Plotkin JB. 2012. Weak 5′-mRNA Secondary Structures in Short Eukaryotic Genes. Genome Biology and Evolution 4:1046-1053.

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A 110:E2792-2801.

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, et al. 2014. A promoter-level mammalian expression atlas. Nature 507:462-470.

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. 2008. A code for transcription initiation in mammalian genomes. Genome Res 18:1-12.

Gene Ontology C, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, et al. 2023. The Gene Ontology knowledgebase in 2023. Genetics 224.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644-652.

Guo N, Qian Y, Zhang Q, Chen X, Zeng G, Zhang X, Mi W, Xu C, St. Leger RJ, Fang W. 2017. Alternative transcription start site selection in Mr-OPY2 controls lifestyle transitions in the fungus Metarhizium robertsii. Nature Communications 8:1565.

Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, et al. 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. Nature 507:381-385.

Hinnebusch AG, Ivanov IP, Sonenberg N. 2016. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science 352:1413-1416.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21:182-192.

Hughes MJ, Andrews DW. 1997. A single nucleotide is a sufficient 5' untranslated region for translation in an eukaryotic in vitro system. FEBS Lett 414:19-22.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res 30:42-46.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907-915.

Kozak M. 2005. A second look at cellular mRNA sequences said to function as internal ribosome entry sites. Nucleic Acids Res 33:6593-6602.

Kukurba KR, Montgomery SB. 2015. RNA Sequencing and Analysis. Cold Spring Harb Protoc 2015:951-969.

Landry JR, Mager DL, Wilhelm BT. 2003. Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet 19:640-648.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13:233-245.

Leppek K, Das R, Barna M. 2018. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nature Reviews Molecular Cell Biology 19:158-174.

Li Y, Varala K, Coruzzi GM. 2015. From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks. Trends Genet 31:509-515.

Lin Z, Li W-H. 2012. Evolution of 5′ Untranslated Region Length and Gene Expression Reprogramming in Yeasts. Mol Biol Evol 29:81-89.

Lin Z, Wu W-S, Liang H, Woo Y, Li W-H. 2010. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. BMC Genomics 11:581.

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16:22.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

Lu Z, Berry K, Hu Z, Zhan Y, Ahn T-H, Lin Z. 2021. TSSr: an R package for comprehensive analyses of TSS sequencing data. NAR Genomics and Bioinformatics 3:lqab108.

Lu Z, Lin Z. 2021. The origin and evolution of a distinct mechanism of transcription initiation in yeasts. Genome Res 31:51-63.

Lu Z, Lin Z. 2019. Pervasive and dynamic transcription initiation in Saccharomyces cerevisiae. Genome Res 29:1198-1210.

Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. Mol Biol Evol 22:1137-1146.

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biol 3:REVIEWS0004.

Mihailovich M, Thermann R, Grohovaz F, Hentze MW, Zacchetti D. 2007. Complex translational regulation of BACE1 involves upstream AUGs and stimulatory elements within the 5' untranslated region. Nucleic Acids Res 35:2975-2985.

Morioka MS, Kawaji H, Nishiyori-Sueki H, Murata M, Kojima-Ishiyama M, Carninci P, Itoh M. 2020. Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites. Methods Mol Biol 2120:277-301.

Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, Itoh M. 2014. Detecting expressed genes using CAGE. Methods Mol Biol 1164:67-85.

Nagaraj SH, Gasser RB, Ranganathan S. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8:6-21.

Nei M. 2014. Mutation-driven evolution. Oxford: Oxford University Press.

Parenteau J, Durand M, Véronneau S, Lacombe A-A, Morin G, Guérin V, Cecez B, Gervais-Bird J, Koh C-S, Brunelle D, et al. 2008. Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. Mol Biol Cell 19:1932-1941.

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, et al. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature 556:339-344.

Polack A, Feederle R, Klobeck G, Hörtnagel K. 1993. Regulatory elements in the immunoglobulin kappa locus induce c-myc activation and the promoter shift in Burkitt's lymphoma cells. Embo J 12:3913-3920.

Policastro RA, Zentner GE. 2021. Global approaches for profiling transcription initiation. Cell Reports Methods 1:100081.

Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol Syst Biol 10:770.

Pozner A, Lotem J, Xiao C, Goldenberg D, Brenner O, Negreanu V, Levanon D, Groner Y. 2007. Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. BMC Dev Biol 7:84.

Rao YS, Wang ZF, Chai XW, Nie QH, Zhang XQ. 2013. Relationship between 5′ UTR length and gene expression pattern in chicken. Genetica 141:311-318.

Reuter M, Engelstädter J, Fontanillas P, Hurst LD. 2008. A Test of the Null Model for 5′ UTR Evolution Based on GC Content. Mol Biol Evol 25:801-804.

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI, et al. 2011. Comparative functional genomics of the fission yeasts. Science 332:930-936.

Rouleau J, Tanigawa G, Szyf M. 1992. The mouse DNA methyltransferase 5'-region. A unique housekeeping gene promoter. J Biol Chem 267:7368-7377.

Shephard EA, Chandan P, Stevanovic-Walker M, Edwards M, Phillips IR. 2007. Alternative promoters and repetitive DNA elements define the species-dependent tissue-specific expression of the FMO1 genes of human and mouse. The Biochemical Journal 406:491-499.

Sobczak K, Krzyzosiak WJ. 2002. Structural determinants of BRCA1 translational regulation. J Biol Chem 277:17349-17358.

Sonenberg N, Hinnebusch AG. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell 136:731-745.

Steijger T, Abril JF, Engström PG, Kokocinski F, Consortium R, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 10:1177-1184.

Steinthorsdottir V, Stefansson H, Ghosh S, Birgisdottir B, Bjornsdottir S, Fasquel AC, Olafsson O, Stefansson K, Gulcher JR. 2004. Multiple novel transcription initiation sites for NRG1. Gene 342:97-105.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLOS ONE 6:e21800.

Tan W, Wang Y, Gold B, Chen J, Dean M, Harrison PJ, Weinberger DR, Law AJ. 2007. Molecular cloning of a brain-specific, developmentally regulated neuregulin 1 (NRG1) isoform and identification of a functional promoter variant associated with schizophrenia. J Biol Chem 282:24343-24351.

Tresenrider A, Morse K, Jorgensen V, Chia M, Liao H, van Werven FJ, Ünal E. 2021. Integrated genomic analysis reveals key features of long undecoded transcript isoform-based gene repression. Mol Cell 81:2231-2245.e2211.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107:3645-3650.

Uppala JK, Sathe L, Chakraborty A, Bhattacharjee S, Pulvino AT, Dey M. 2022. The cap-proximal RNA secondary structure inhibits preinitiation complex formation on HAC1 mRNA. Journal of Biological Chemistry 298:101648.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19:255-265.

van der Velden AW, Thomas AA. 1999. The role of the 5' untranslated region of an mRNA in translation regulation during development. Int J Biochem Cell Biol 31:87-106.

Waern K, Snyder M. 2013. Extensive transcript diversity and novel upstream open reading frame regulation in yeast. G3 (Bethesda) 3:343-352.

Wang B, Guo G, Wang C, Lin Y, Wang X, Zhao M, Guo Y, He M, Zhang Y, Pan L. 2010. Survey of the transcriptome of Aspergillus oryzae via massively parallel mRNA sequencing. Nucleic Acids Research 38:5075-5087.

Wang X, Hou J, Quedenau C, Chen W. 2016. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol 12:875.

Wickham H. 2009. ggplot2 Elegant Graphics for Data Analysis.

Xu C, Park JK, Zhang J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. PLoS Biol 17:e3000197.

Zhang P, Dimont E, Ha T, Swanson DJ, Hide W, Goldowitz D. 2017. Relatively frequent switching of transcription start sites during cerebellar development. BMC Genomics 18:461.

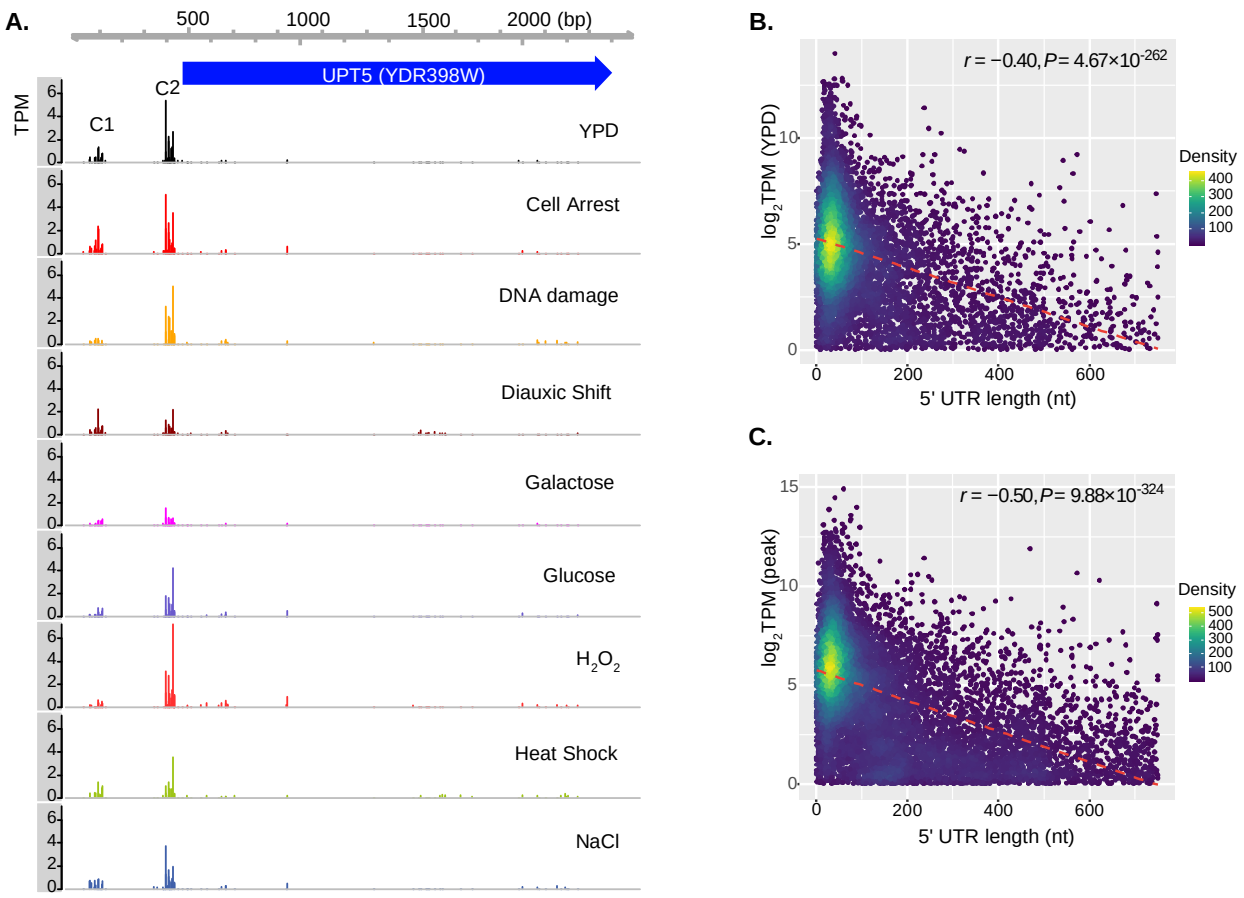


Figure 1. Transcript's 5′ UTR length is a strong indicator of its expression level. (A) An example of TSS maps of the UPT5 gene obtained in *S. cerevisiae* cultured in nine different growth conditions. The CDS of UPT5 is shown as a blue arrow. Descriptions of each growth condition are provided in Table S1. (B) Negative correlation between transcript 5' UTR lengths and their expression levels in *S. cerevisiae* cultured in rich YPD medium. (C) Negative correlation between transcript 5' UTR lengths and their peak expression levels in *S. cerevisiae*. A transcript isoform's peak expression level is the highest expression level among samples cultured in nine growth conditions. Each dot represents a transcript isoform. Colors show the density of the transcripts. The best-fit line based on linear regression is shown as a red dashed line.

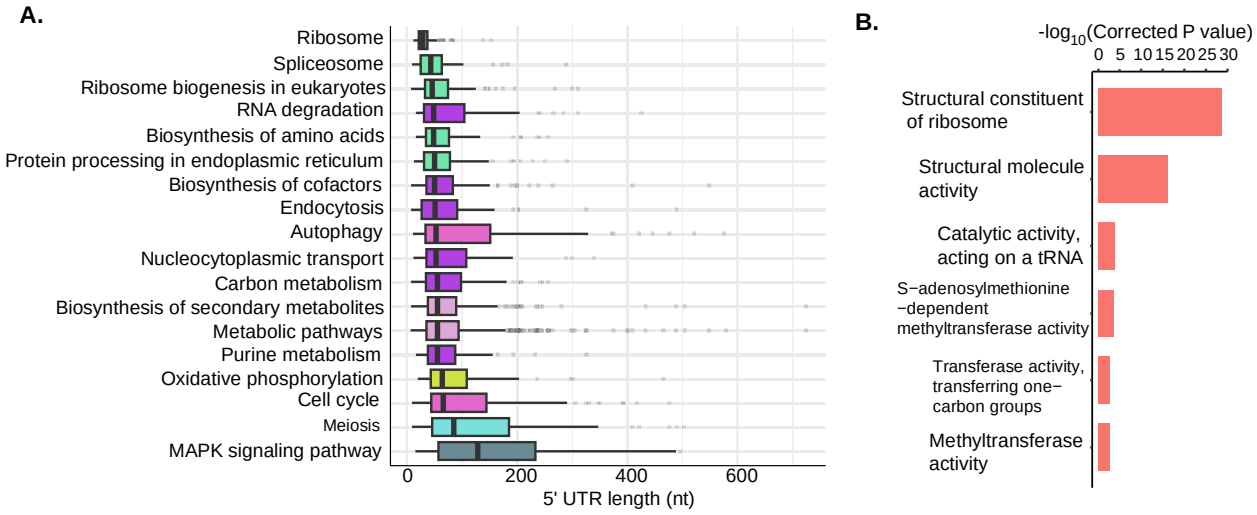


Figure 2. Non-random distributions of 5′ UTR length among genes of different functional groups. (A) Box plots showing large variations in 5′ UTR lengths among genes among KEGG pathways. (B) GO term enrichment analysis for genes with short 5′ UTR (< 25th percentile). GO groups of molecular function were sorted by decreasing order of evidence, based on adjusted P-value.

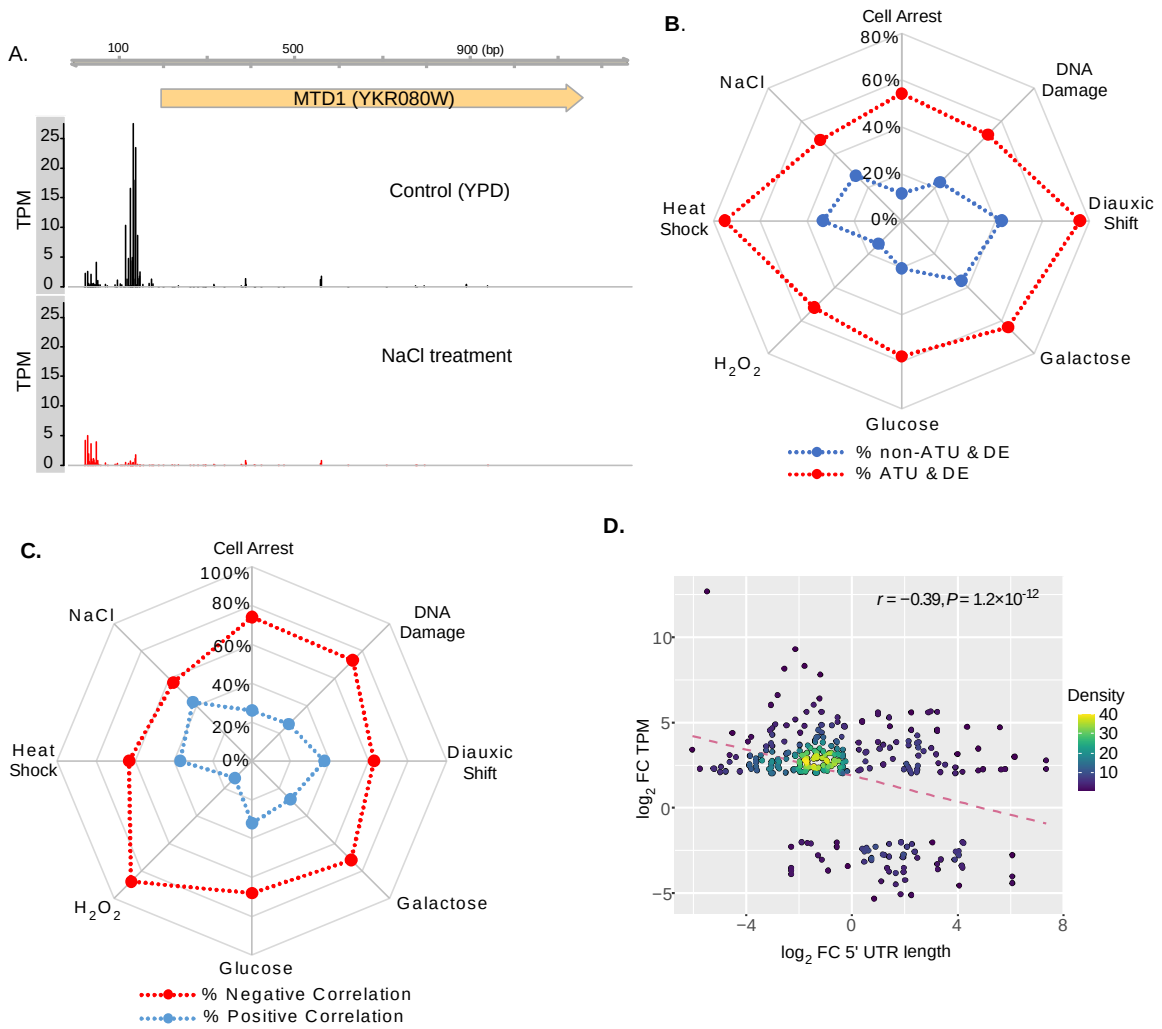


Figure 3. Changes in 5′ UTR length are associated with gene differential expression in response to environmental stimuli. (A) An example of changing 5′ UTR length by the MTD1 gene with treatment of 1M NaCl due to alternative TSS usage (ATU). (B) Genes with ATU are more likely to be associated with differential expression (DE) in response to environmental stimuli. The proportions of genes with ATU and DE (ATU & DE) are shown as red dots, while the proportions of genes without ATU but with DE (non-ATU & DE) are shown as blue dots. Specific treatments are indicated. (C) Changes in 5′ UTR length are more likely to be negatively correlated with changes in expression levels. The proportions of genes with a negative correlation between changes in 5′ UTR length and expression levels are shown as red dots and proportions of genes with a positive correlation are shown as blue dots. (D) A strong negative correlation between fold changes (FC) of weighted 5' UTR length and gene expression levels in the eight treatments. The colors show the data density. The best-fit line based on linear regression is shown as a red dashed line.

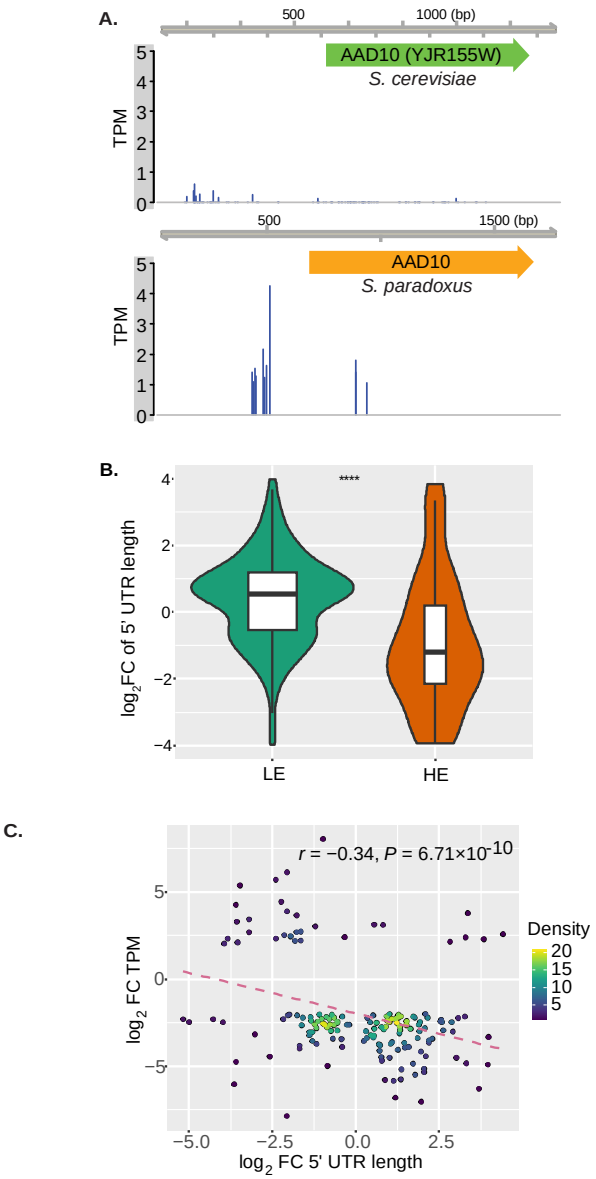


Figure 4 Changes in 5' UTR length negatively correlate with the divergence of gene expression levels during evolution of yeast species. (A) An example of evolutionary divergence of 5' UTR length and gene expression levels. The mRNA transcripts of *S. cerevisiae* AAD10 gene and its orthologs gene in *S. paradoxus* differ in 5' UTR length due to divergence of TSSs. The expression levels between the two orthologous genes have also diverged during the evolution of the two species. (B) Orthologous genes with a lower expression level (LE) tend to have longer 5'UTR ($\log_2$ FC > 0), while orthologous genes with a higher expression level (HE) tend to have shorter 5'UTR ($\log_2$ FC < 0). (C) A negative correlation between evolutionary changes in 5'UTR length and divergence in gene expression level between *S. cerevisiae* and *S. paradoxus*. The colors show the density of transcripts. The best-fit line based on linear regression is shown as a red dashed line.

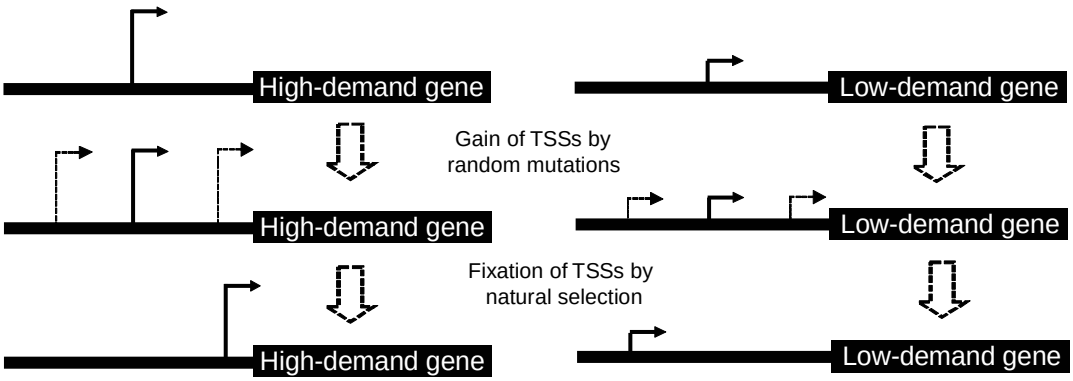


Figure 5. A revised evolutionary model of 5' UTR length. A Gene may gain new or lose a TSS due to stochastic mutations. The evolutionary fate of a new TSS is largely determined by its locations and expression demand of a gene. As a shorter transcript isoform has a better translation efficiency, the TSS that generates a shorter 5'UTR would be favored by nature selection and fixed for a high-demand gene. In contrast, a TSS that generates longer 5'UTR would be favored by a gene with a lower expression demand for lower translation efficiency to further reduce protein production.

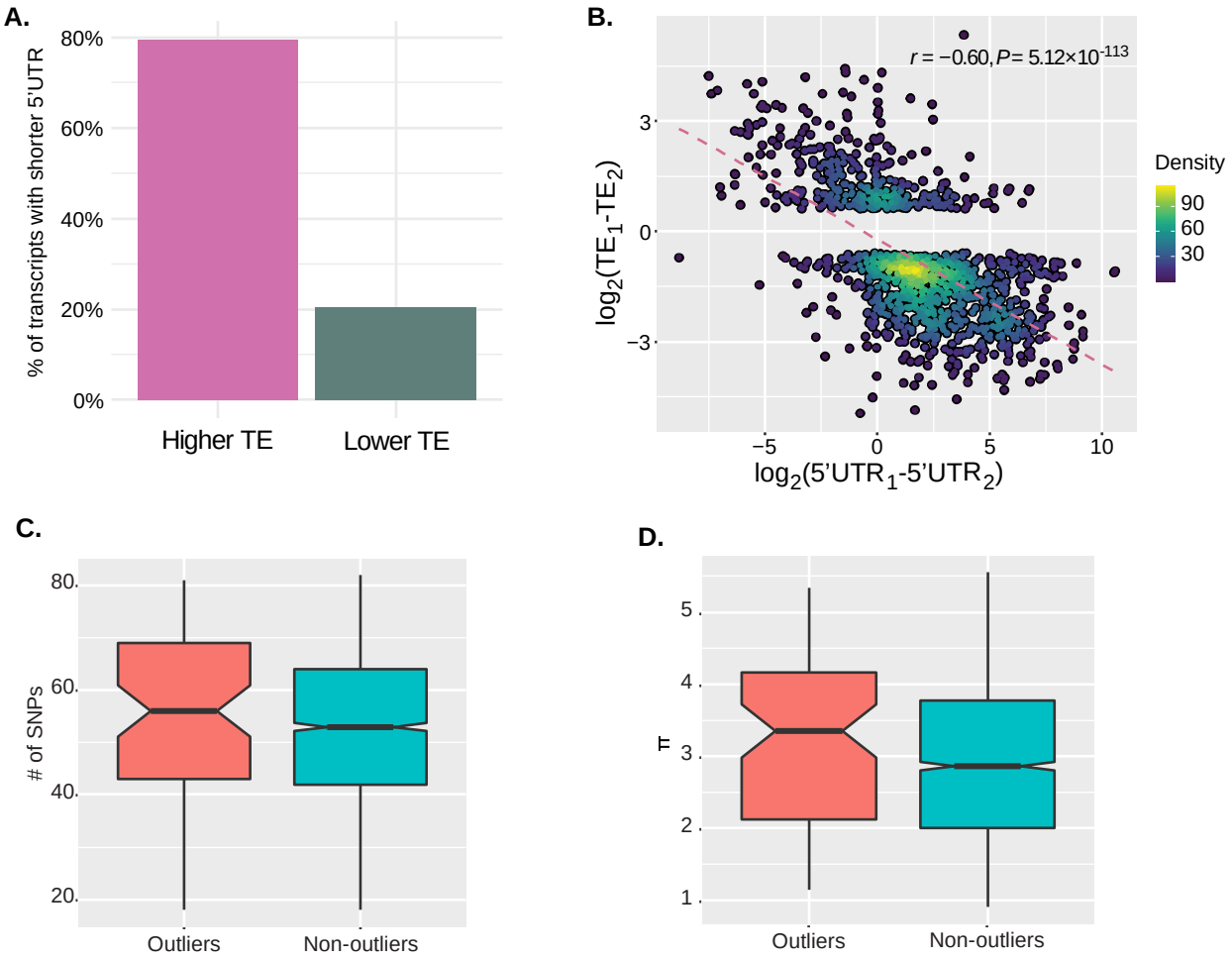
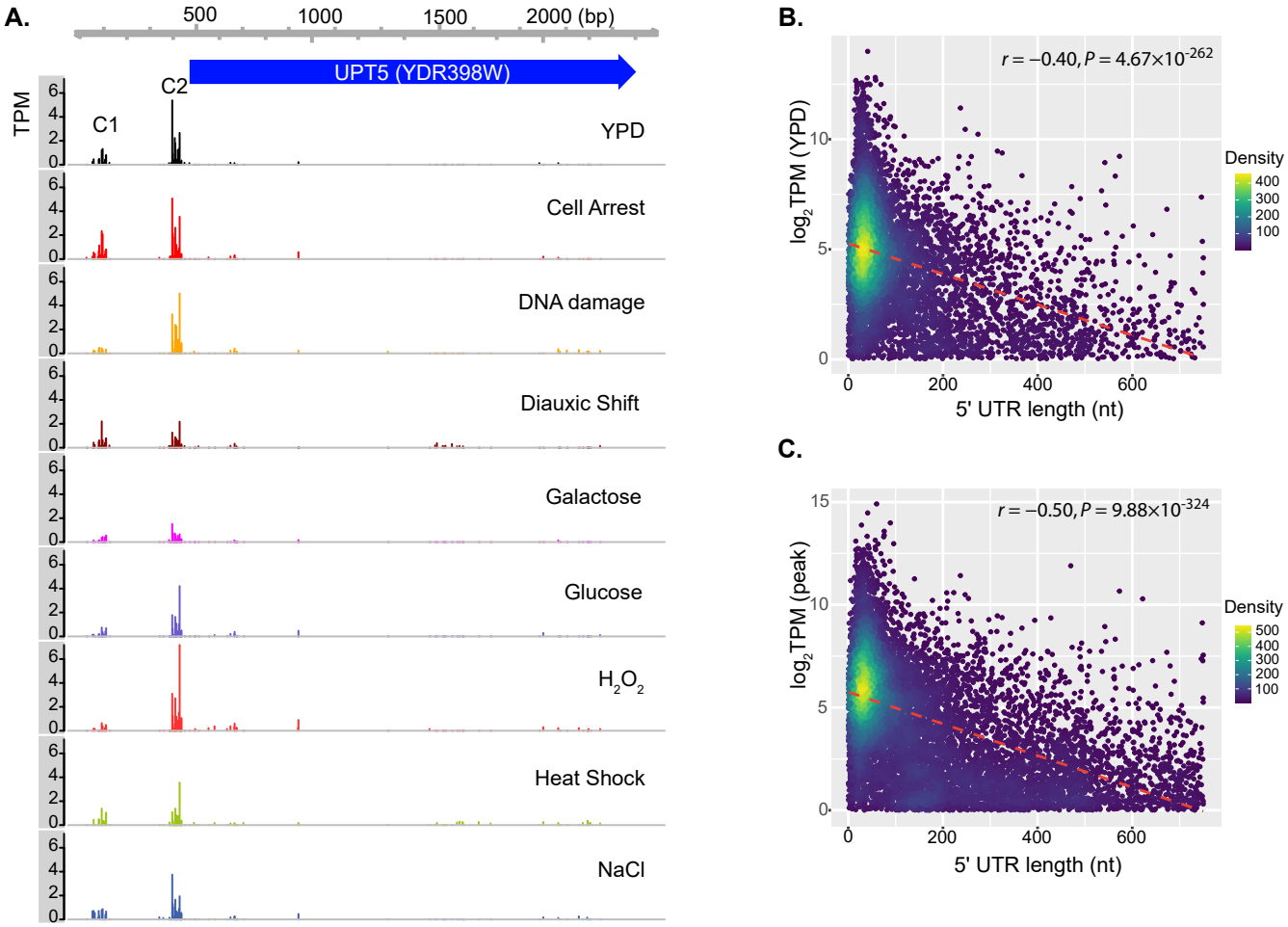
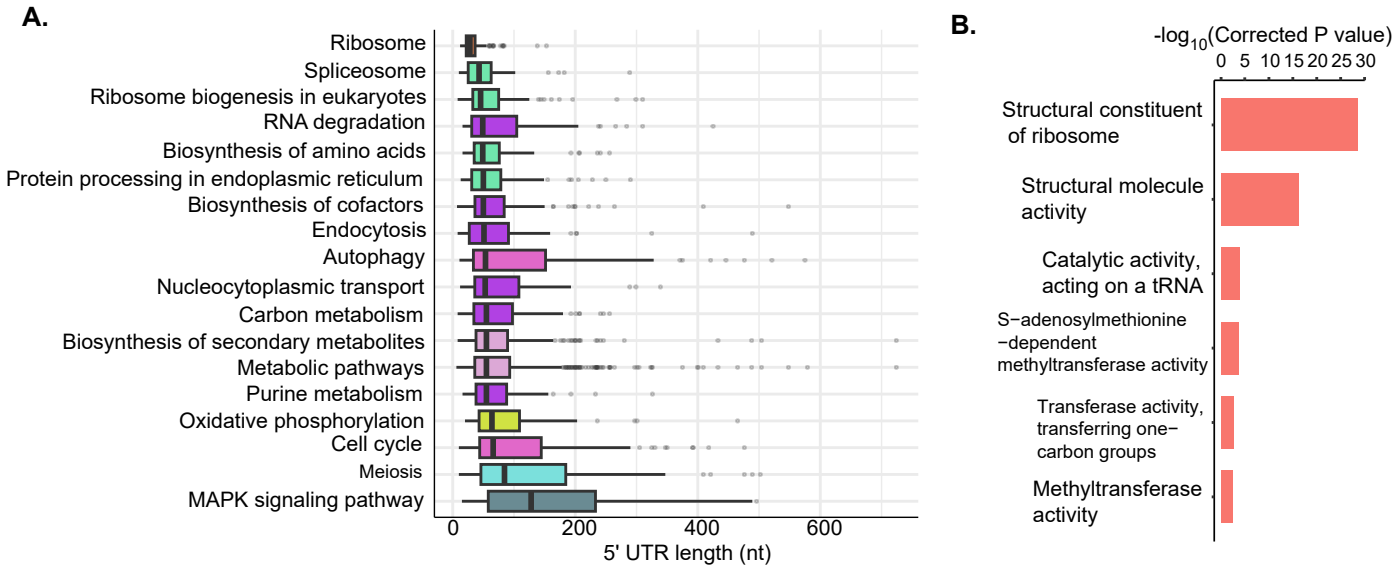
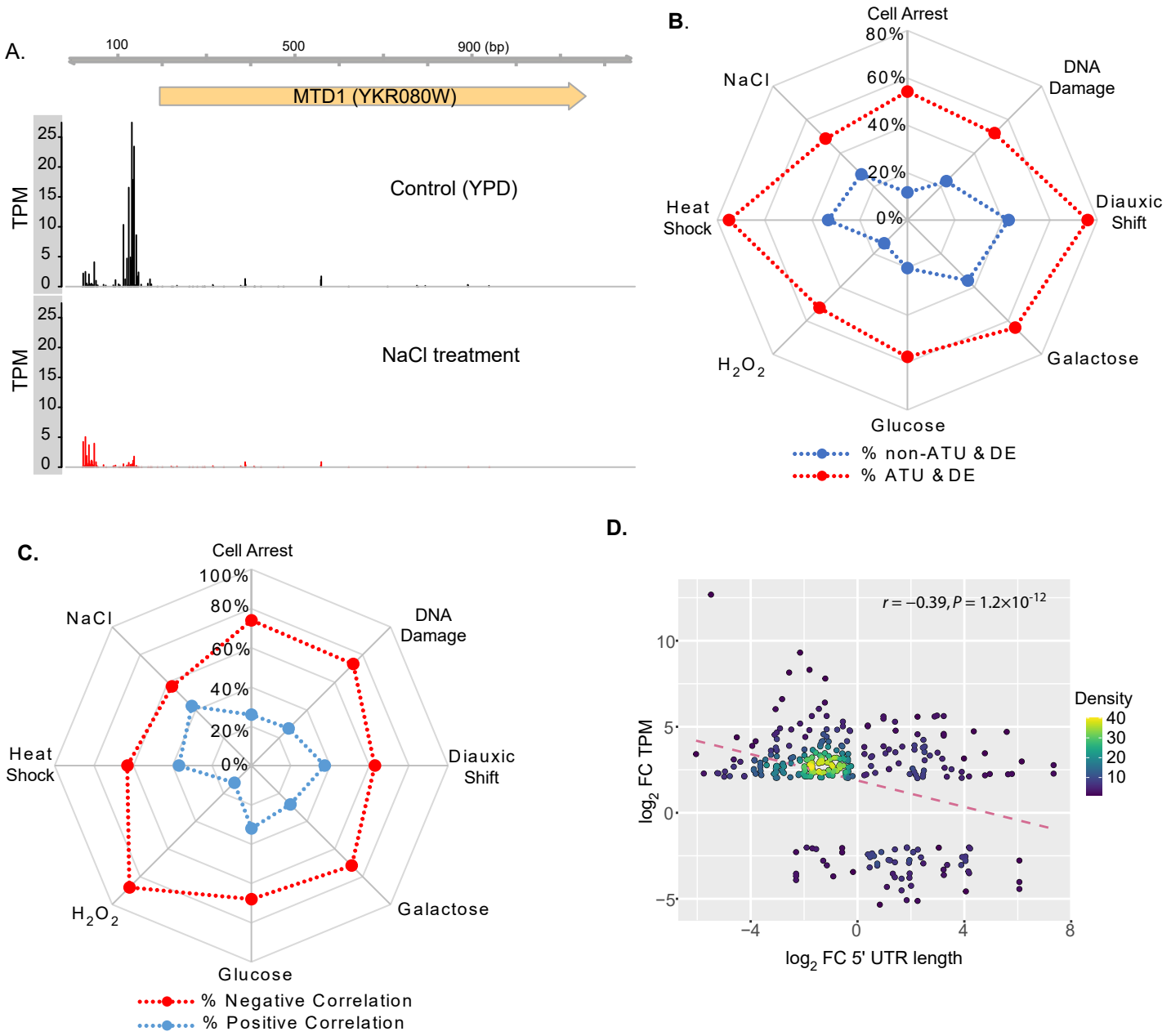
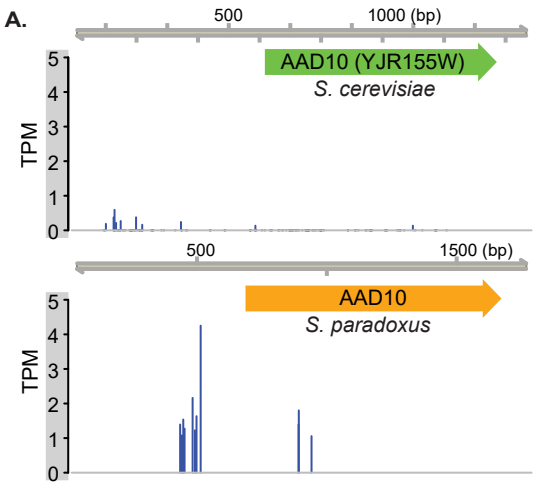


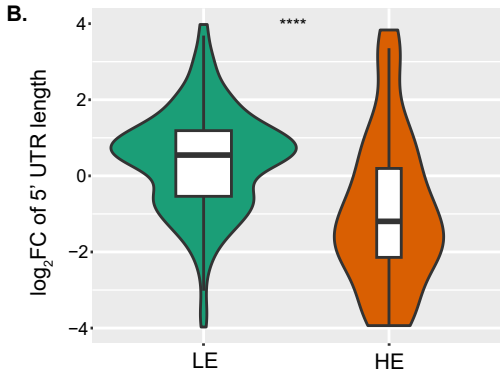
Figure 6. Evidence supporting the evolution of 5′ UTR length was shaped by natural selection. (A) Most transcripts (~80%) with a higher translation efficiency (TE) have a shorter 5′ UTR than its longer isoforms produced by the same gene. This number is only 20% in transcripts with a higher TE. (B) Negative correlations between difference in 5′ UTR lengths and translation efficiency (TE) between transcript isoform pairs in mouse. Transcript isoform pairs with differences in TE with at least 1.5-fold differences were employed in this study. (C) Box plots showing more SNPs identified in the promoters of genes with non-optimal 5′ UTR lengths (outlier genes) than others. (D) Box plots showing a higher nucleotide diversity (π) in the promoters of genes with non-optimal 5′ UTR lengths (outlier genes) than others.

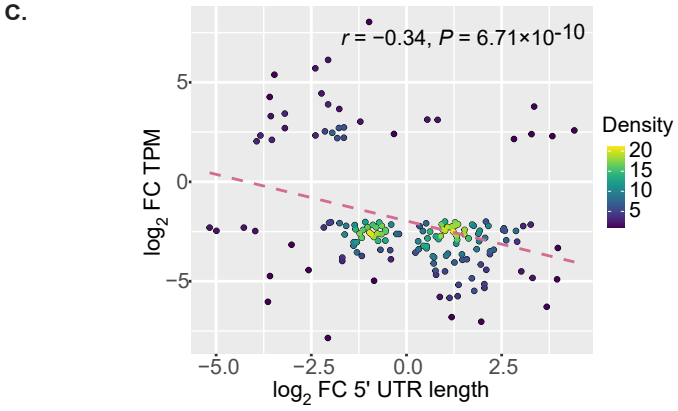

A.
500
1000
1500
2000 (bp)
UPT5 (YDR398W)
TPM
C1
C2
YPD
Cell Arrest
DNA damage
Diauxic Shift
Galactose
Glucose
H2O2
Heat Shock
NaCl
B.
r = −0.40, P = 4.67×10-262
log2TPM (YPD)
5' UTR length (nt)
Density
400
300
200
100
C.
r = −0.50, P = 9.88×10-324
log2TPM (peak)
5' UTR length (nt)
Density
500
400
300
200
100

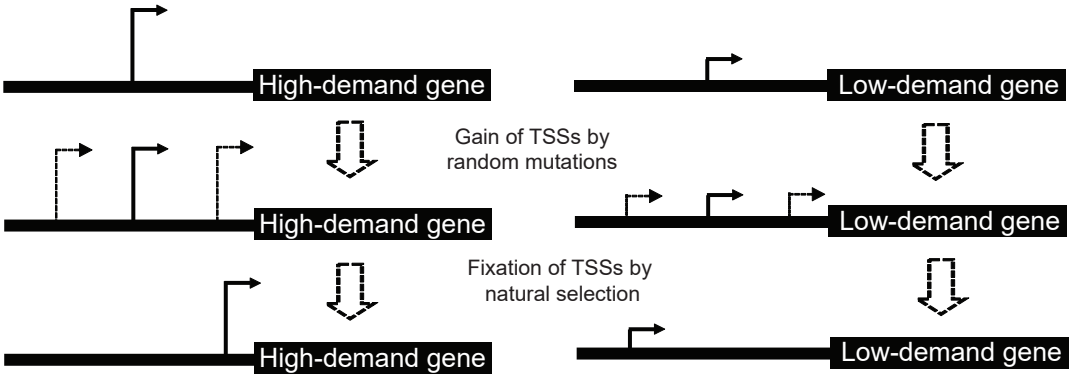

A.
Ribosome
Spliceosome
Ribosome biogenesis in eukaryotes
RNA degradation
Biosynthesis of amino acids
Protein processing in endoplasmic reticulum
Biosynthesis of cofactors
Endocytosis
Autophagy
Nucleocytoplasmic transport
Carbon metabolism
Biosynthesis of secondary metabolites
Metabolic pathways
Purine metabolism
Oxidative phosphorylation
Cell cycle
Meiosis
MAPK signaling pathway
0
200
400
600
5' UTR length (nt)
B.
-log10(Corrected P value)
0 5 10 15 20 25 30
Structural constituent of ribosome
Structural molecule activity
Catalytic activity, acting on a tRNA
S−adenosylmethionine −dependent methyltransferase activity
Transferase activity, transferring one−carbon groups
Methyltransferase activity

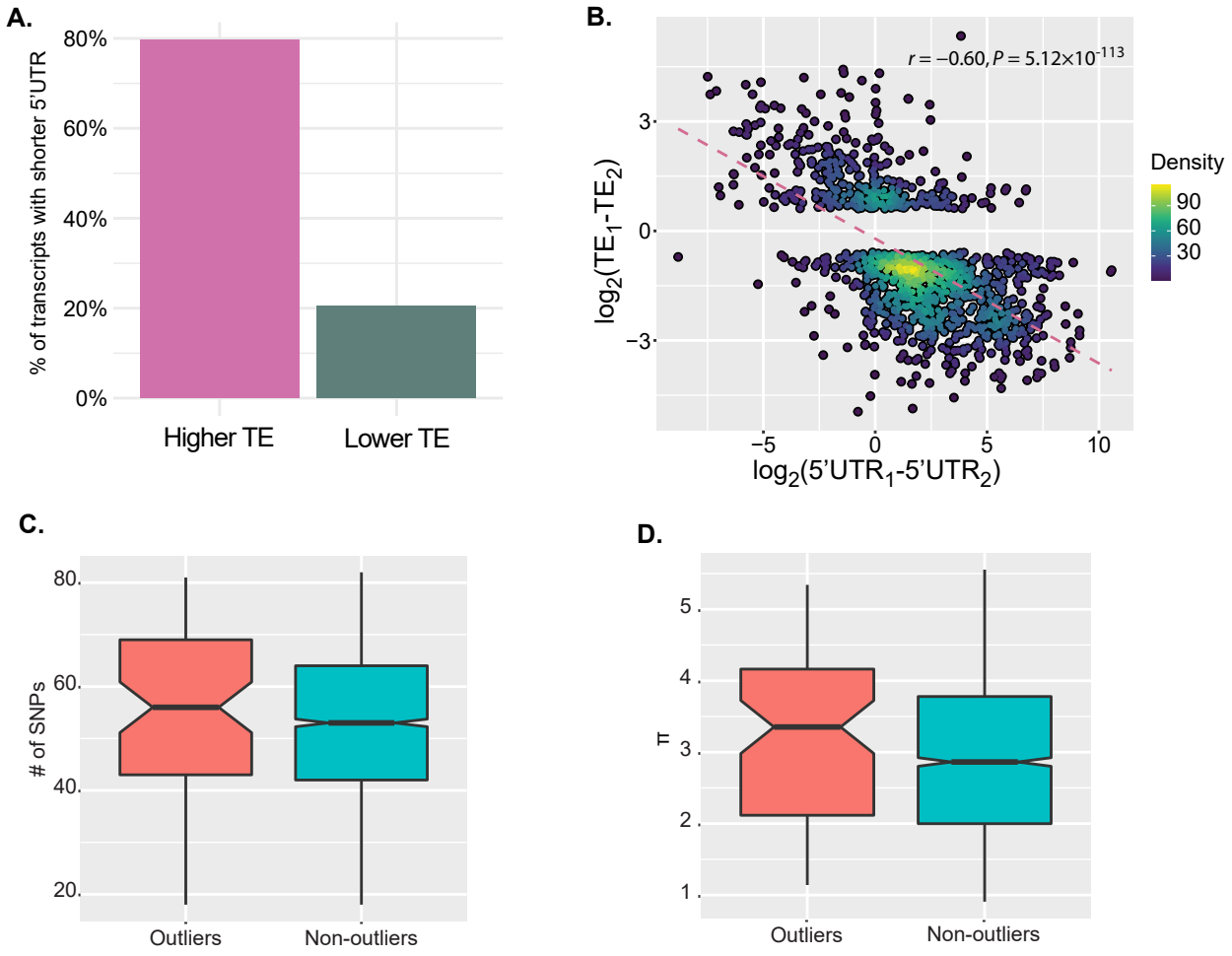

A.
100
500
900 (bp)
MTD1 (YKR080W)
TPM
Control (YPD)
NaCl treatment
B.
Cell Arrest
DNA Damage
Diauxic Shift
Galactose
Glucose
H_2O_2
Heat Shock
NaCl
80%
60%
40%
20%
0%
% non-ATU & DE
% ATU & DE
C.
Cell Arrest
DNA Damage
Diauxic Shift
Galactose
Glucose
H_2O_2
Heat Shock
NaCl
100%
80%
60%
40%
20%
0%
% Negative Correlation
% Positive Correlation
D.
$r = -0.39, P = 1.2\times10^{-12}$
Density
$\log_2$ FC TPM
$\log_2$ FC 5' UTR length


A.
500
1000 (bp)
AAD10 (YJR155W)
S. cerevisiae
TPM
500
1500 (bp)
AAD10
S. paradoxus
TPM


B.
****
log2FC of 5' UTR length
LE
HE


C.
r = −0.34, P = 6.71×10-10
Density
log2 FC TPM
log2 FC 5' UTR length


High-demand gene
Low-demand gene
Gain of TSSs by
random mutations
High-demand gene
Low-demand gene
Fixation of TSSs by
natural selection
High-demand gene
Low-demand gene


A.
% of transcripts with shorter 5'UTR
80%
60%
40%
20%
0%
Higher TE
Lower TE
B.
r = −0.60, P = 5.12×10^-113
log2(TE1-TE2)
log2(5'UTR1-5'UTR2)
Density
90
60
30
C.
# of SNPs
Outliers
Non-outliers
D.
π
Outliers
Non-outliers