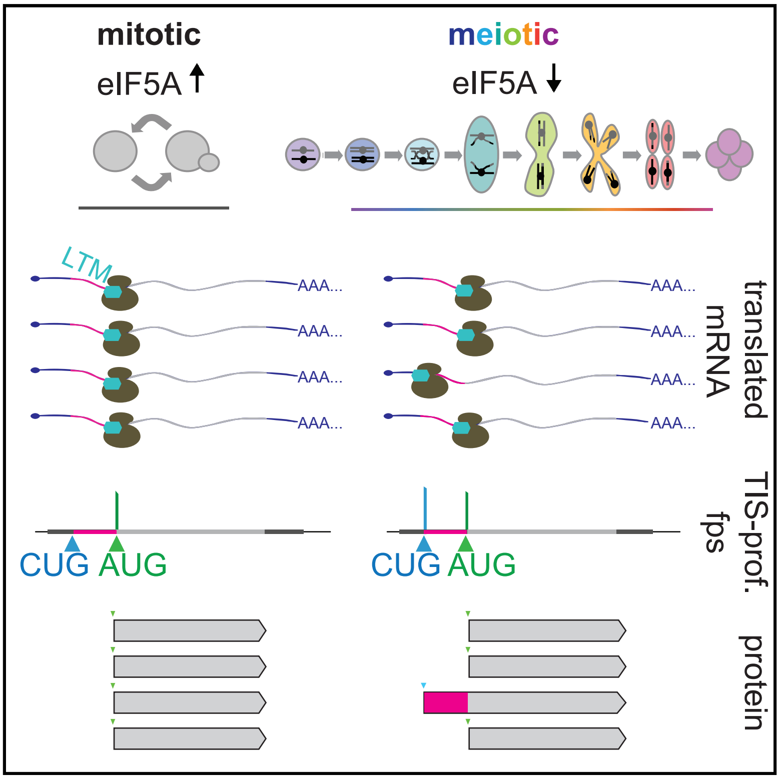
Cell Systems

Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast

Graphical Abstract



Authors

Amy R. Eisenberg, Andrea L. Higdon, Ina Hollerer, ..., Manolis Kellis, Marko Jovanovic, Gloria A. Brar

Correspondence

gabrar@berkeley.edu

In Brief

Eisenberg et al. identify translation initiation sites genome-wide in budding yeast. They define a class of 149 genes with alternate extended isoforms that initiate at non-AUG start codons. These isoforms are produced in concert with their corresponding canonical isoforms but typically at lower abundance, and are selectively induced during meiosis.

Highlights

- TIS-profiling reveals widespread translation of non-canonical ORFs in budding yeast
- Production of non-AUG-initiated extended isoforms is prevalent and inefficient
- A small subset of possible near-cognate sites is used for translation initiation
- eIF5A-based regulation allows conditional unmasking of non-AUG initiation in meiosis

Eisenberg et al., 2020, Cell Systems *11*, 1–16
August 26, 2020 © 2020 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.06.011

Cell Systems

CellPress
OPEN ACCESS

Article

Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast

Amy R. Eisenberg,[1] Andrea L. Higdon,[1,2] Ina Hollerer,[1] Alexander P. Fields,[3,7] Irwin Jungreis,[4,5] Paige D. Diamond,[1] Manolis Kellis,[4,5] Marko Jovanovic,[6] and Gloria A. Brar[1,2,8,*]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Center for Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[3]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA
[4]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
[5]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[6]Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[7]Present address: GRAIL, Menlo Park, CA 94025, USA
[8]Lead Contact
*Correspondence: gabrar@berkeley.edu

https://doi.org/10.1016/j.cels.2020.06.011

SUMMARY

Genomic analyses in budding yeast have helped define the foundational principles of eukaryotic gene expression. However, in the absence of empirical methods for defining coding regions, these analyses have historically excluded specific classes of possible coding regions, such as those initiating at non-AUG start codons. Here, we applied an experimental approach to globally annotate translation initiation sites in yeast and identified 149 genes with alternative N-terminally extended protein isoforms initiating from near-cognate codons upstream of annotated AUG start codons. These isoforms are produced in concert with canonical isoforms and translated with high specificity, resulting from initiation at only a small subset of possible start codons. The non-AUG initiation driving their production is enriched during meiosis and induced by low eIF5A, which is seen in this context. These findings reveal widespread production of non-canonical protein isoforms and unexpected complexity to the rules by which even a simple eukaryotic genome is decoded.

INTRODUCTION

Our understanding of cell function has been advanced by genome annotations that comprehensively predict the repertoire of protein products within the cell. Genes were historically annotated computationally based on a set of rules that were informed by existing knowledge of the mechanism of translation and the features shared by most well-studied genes (Brent, 2005). Open reading frames (ORFs), for example, have been defined as starting at an AUG and stopping at the next in-frame stop codon because this reflects characterized properties of translation of an mRNA by the ribosome (reviewed in Aitken and Lorsch, 2012). Development of experimental approaches to globally define translated regions has now made it possible to determine the prevalence of translated ORFs that do not follow these rules. Additionally, such approaches enable identification of condition-specific changes in ORF identity, such as during stress or developmental progression, which cannot be predicted from sequence-based annotation alone.

Ribosome profiling was the first method to allow genome-wide experimental identification of translated regions *in vivo*. This method involves isolating and sequencing the short (~30 nt) regions of mRNA that are protected from nuclease digestion by translating ribosomes (Ingolia et al., 2009). We previously used ribosome profiling to assess changes in translation as yeast cells progress through meiosis (Brar et al., 2012), the highly conserved cellular differentiation program that leads to gamete formation. We observed pervasive and condition-specific non-canonical translation, including spans of translation that initiated at near-cognate start codons (which differ from AUG by one nucleotide) and translation of uORFs (upstream ORFs) in 5′ leader regions. However, the prevalence of overlapping translated ORFs in 5′ leader regions in meiotic cells made it challenging to unambiguously assign ribosome footprints, complicating our goal of achieving high-confidence annotations of all translated ORFs.

A modified ribosome profiling strategy, in which cells are pretreated with drugs that inhibit post-initiation ribosomes, yields footprint reads that map primarily to translation initiation sites (TISs), aiding in the detection and annotation of ORFs (Ingolia et al., 2011; Lee et al., 2012). Global TIS mapping has been performed under several conditions (Fields et al., 2015; Fritsch et al., 2012; Ingolia et al., 2011; Lee et al., 2012; Machkovech et al., 2019; Sapkota et al., 2019; Stern-Ginossar et al., 2012) but thus far only in mammals and viruses, which have complex gene structures. Budding yeast (*Saccharomyces cerevisiae*) has relatively simple transcript architectures with far fewer known cases of complexity, such as from alternative splicing, despite extensive analyses of its transcriptome (Davis et al., 2000; Hossain et al., 2011; Juneau et al., 2009; Kim Guisbert et al., 2012; Yassour et al., 2009). This simple architecture allows for investigation of TISs to be more directly informative, as identification of the start codon alone can generally be used to define an ORF.

We developed a TIS identification approach for budding yeast, both in vegetative and meiotic conditions, with the goal of characterizing ORF types that were previously challenging to identify systematically by standard ribosome profiling. The class of ORFs that we were most interested in assessing, due to their potential to modulate the function of well-characterized genes, were those encoding alternate protein isoforms that result from translation initiation at non-AUG codons upstream of the characterized start codon (see Table S1 for a summary of prior studies of this class of proteins). Several individual examples of N-terminally extended proteins isoforms have been identified in an ad hoc manner using classical approaches (Chang and Wang, 2004; Heublein et al., 2019; Kearse and Wilusz, 2017; Kritsiligkou et al., 2017; Monteuuis et al., 2019; Suomi et al., 2014; Tang et al., 2004; Touriol et al., 2003), and a recent computational study predicted the existence of many additional cases (Monteuuis et al., 2019). However, it was not previously possible to directly experimentally evaluate the prevalence of this class of translation products comprehensively in yeast. Our approach allowed us to determine that condition-specific translation of non-AUG-initiated protein isoforms is common, reflecting regulated induction of a pool of alternative proteins that is facilitated by low eIF5A levels. More broadly, this study revealed surprising complexity to translation—even at characterized loci—in this widely studied organism.

RESULTS

TIS-Profiling in Yeast Globally Defines Translation Initiation Sites

We sought to perform TIS identification in yeast by using ribosome profiling following pre-treatment with harringtonine or lactimidomycin (LTM), two established drugs that preferentially inhibit post-initiation ribosomes but allow elongating ribosomes to run off, resulting in ribosome footprint enrichment at TISs (Figure 1A; Fresno et al., 1977; Ingolia et al., 2011; Lee et al., 2012; Sugawara et al., 1992). Initial testing of both drugs under the conditions used for this purpose in mammalian contexts was unsuccessful in yeast. Even treatment with extremely high concentrations of harringtonine (10-fold higher than used in mammalian cells; Ingolia et al., 2011) did not result in a growth defect, suggesting that this drug does not effectively inhibit translation in yeast. Harringtonine treatment did inhibit the growth of a yeast strain that lacks ABC transporter efflux pumps, pointing to active drug efflux as the mechanism of harringtonine resistance in wild-type yeast (Figure S1A; Suzuki et al., 2011). However, this strain could not efficiently undergo meiosis, precluding its use for our experiments (data not shown).

Testing of previously used LTM treatment conditions resulted in ribosome profiling reads throughout ORFs in yeast, consistent with LTM inhibiting both post-initiation and elongating ribosomes at high concentrations (Figure S1B; Schneider-Poetsch et al., 2010). LTM concentrations 25-fold less than those used for TIS mapping in mammalian cells (Lee et al., 2012) still caused a growth defect in yeast (Figure S1C) and resulted in strong TIS enrichment of ribosome footprints (Figure S1D). This suggests that post-initiation ribosomes are more sensitive to LTM-based inhibition than elongating ribosomes. We selected an LTM concentration of 3 μM and a 20 minute incubation prior to harvesting to allow sufficient run-off time for elongating ribosomes. We performed translation initiation site profiling (TIS-profiling) for eight meiotic time points to assess translation initiation globally during meiosis (Figure 1B). For comparison, we also included samples from vegetative cells during either exponential growth or stationary phase, as well as diploid cells that cannot undergo meiosis grown in media matched to meiotic samples (*MAT*a/a). Metagene analysis of the regions surrounding annotated start codons revealed a strong peak at the TIS and a low level of background reads in ORF bodies, suggesting that TISs were indeed being highly efficiently captured by our approach (Figure 1C). This is in contrast to the expected distribution of ribosome footprint reads across the entirety of the ORF seen for standard ribosome profiling, which is also seen for a representative gene, *TUB2* (Figures 1C and 1D).

We confirmed that our data accurately reported the expected positions and condition specificity of both canonical and non-canonical start codons through analysis of several well-studied genes. For example, at the locus of a meiotic gene, *REC8*, a single abundant peak was observed at the known TIS during time points when Rec8 is normally expressed (Figure 1E). TIS-profiling also revealed peaks at known non-canonical TISs, including the four AUG-initiated uORFs known to regulate *GCN4* (Figure 1F). Finally, peaks at near-cognate codons were detected in our dataset, consistent with mammalian experiments using LTM or harringtonine (Ingolia et al., 2011; Lee et al., 2012). One of the few characterized examples of productive near-cognate translation initiation in yeast is for the tRNA synthetase gene *ALA1*, which encodes two functionally characterized protein isoforms (Tang et al., 2004). Translation of the canonical isoform initiates at an AUG, while translation of an N-terminally extended isoform initiates from an ACG in the 5′ leader. This upstream initiation event appends a mitochondrial targeting sequence to the canonical protein, which localizes this isoform to the mitochondria. We observed strong and specific peaks for both the upstream near-cognate start codon as well as the annotated AUG for *ALA1* in our dataset (Figure 1G) and concluded that our TIS-profiling protocol could capture both known canonical and non-canonical TISs.

TIS-Profiling Reveals Thousands of Non-canonical ORFs

To systematically annotate translation products, including those that were challenging to assess by traditional ribosome profiling, like alternate protein isoforms, we used ORF-RATER, a linear regression algorithm (Fields et al., 2015). ORF-RATER integrates both standard and TIS-profiling data to evaluate read patterns over ORFs within annotated transcripts. It then assigns scores to detected peaks based on the similarity of their read patterns to annotated ORFs, with scores closest to 1 being the most similar. This method was particularly well suited to our goal of

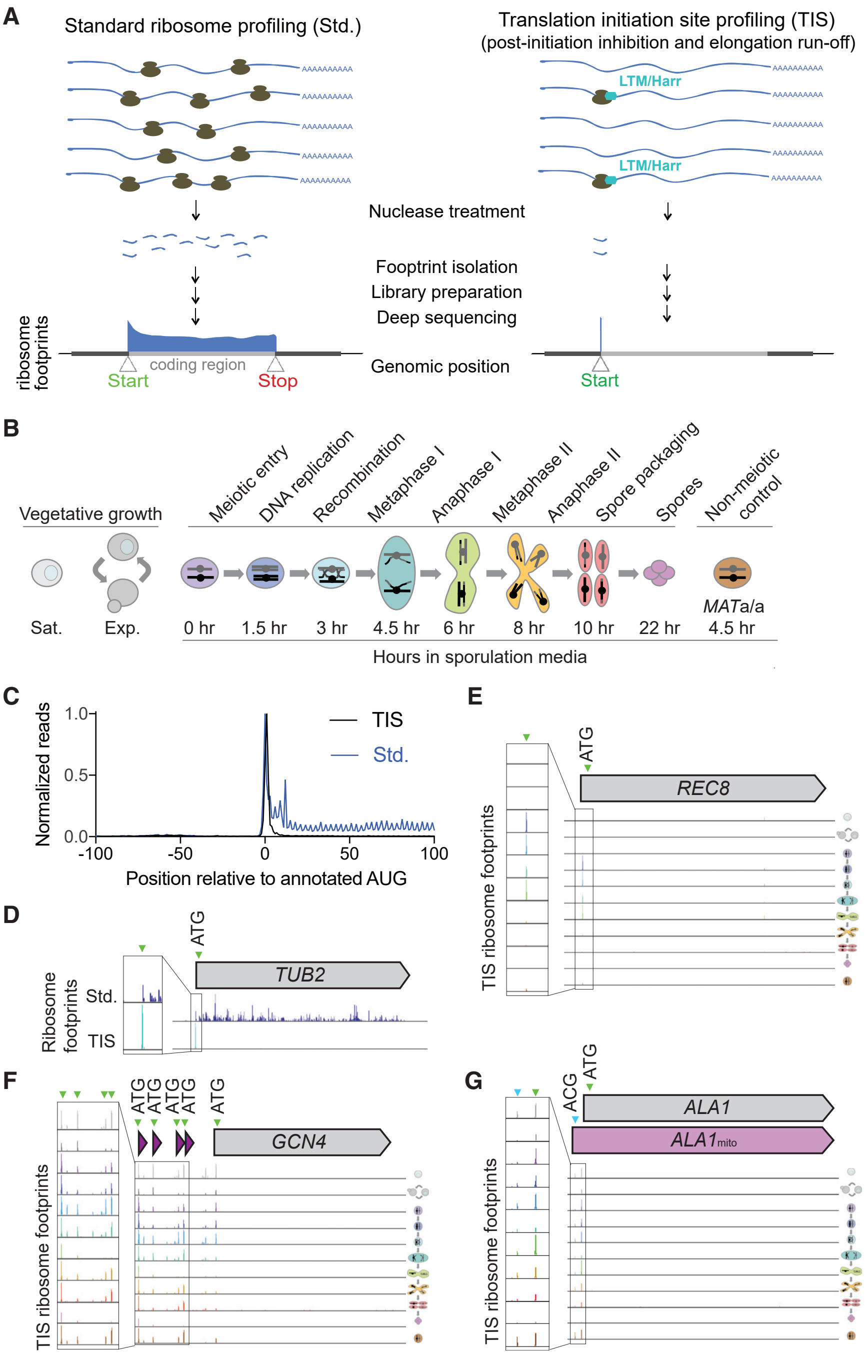


Figure 1. TIS Ribosome Profiling in Mitotic and Meiotic Yeast Cells

(A) Cartoon comparing standard (Std., left) and TIS (right) ribosome profiling, with representative ribosome footprint profiles for a typical ORF.

(B) Schematic of yeast cell stages and samples collected for TIS-profiling, including vegetative saturated (sat.), vegetative exponential (exp.), 0, 1.5, 3, 4.5, 6, 8, 10, and 22 h after addition to sporulation media, and a *MAT*a/a non-meiotic control taken at 4.5 h in sporulation media.

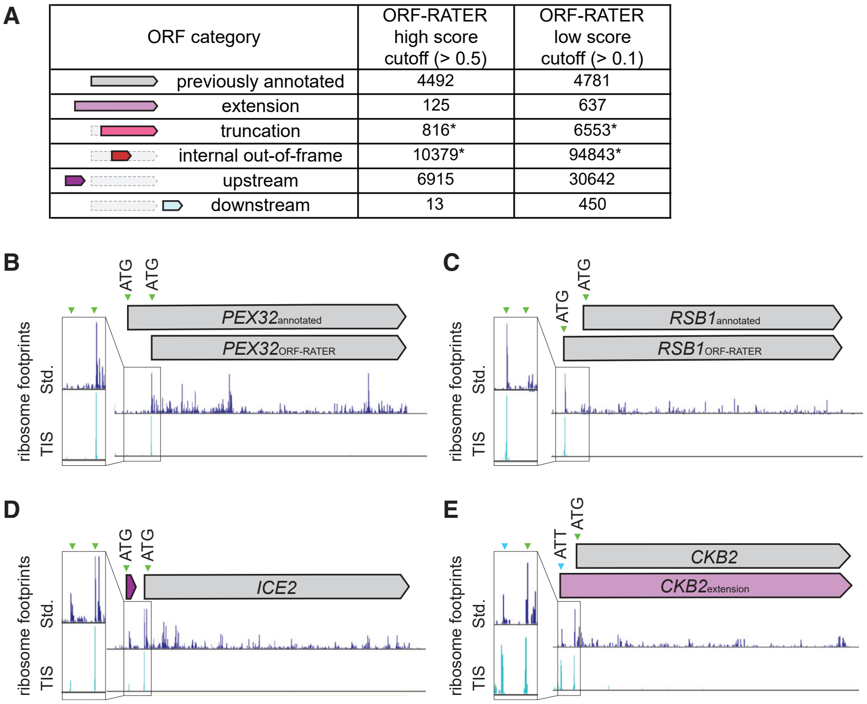
(C) Metagene plot of normalized reads from standard ribosome profiling (blue) and TIS-profiling (black), 100 nucleotides upstream and downstream of annotated AUG start codons. Values are normalized to the peak at position zero.

(D) Comparison of standard and TIS-profiling for *TUB2*, a representative gene, from all time points combined. Green arrowheads indicate peaks at ATGs, and inset shows close-up view of region around predicted initiation site.

(E–G) TIS-profiling of *REC8* (E), *GCN4* (F), and *ALA1* (G), showing ribosome footprints at the time points indicated in Figure 1B. Green arrowheads indicate peaks at ATGs, and blue arrowheads indicate non-ATG peaks.

identifying uORFs and ORFs that overlap annotated ORFs, which were the most difficult to annotate from standard ribosome profiling data since they are often obscured by signal from elongating ribosomes.

ORF-RATER successfully called most previously annotated canonical coding regions using the TIS-profiling dataset and a time-point-matched standard ribosome profiling dataset (Cheng et al., 2018). Of annotated ORFs in our yeast reference dataset, ORF-RATER identified 67% at a high score cutoff (>0.5; Figure 2A; Table S2). Of those that were not called by ORF-RATER, 45.8% are expressed at low abundance under the conditions tested (fewer than 5 mean reads per kilobase million [RPKM]; Figures S2A and S2B). An interesting category of uncalled annotated ORFs includes cases of apparent misannotation, such as *PEX32* and *RSB1*, for which the likely predominant initiation site based on TIS-profiling and ORF-RATER analysis is upstream or downstream of the annotated TIS (Figures 2B and 2C). In these cases, the previously annotated TIS does not show evidence of initiation in our dataset, indicating that the alternate TIS that is called is likely to be the correct one for these genes. This category represents approximately 39% of "uncalled" annotated ORFs, as these are instead erroneously called as extensions or truncations. This includes cases for which the previous annotation was based on the assumption that the predominant TIS is the one that produces the longest possible ORF at a given locus and also includes cases in which the original reference genome annotation for the ORF was incorrect based on sequencing errors or sequence differences between yeast strains. An example of the latter is *DEP1*, which has a stop codon upstream of the annotated stop codon in our strain background (SK1; Figure S2C). Finally, we estimate that approximately 15% of uncalled canonical annotated ORFs (representing 5% of total annotated ORFs) are false negatives, like *RIM11*, for which ORF-RATER did not call an ORF despite an observable peak at the annotated start codon in the TIS-profiling data (Figure S2D).



ORF category	ORF-RATER high score cutoff (> 0.5)	ORF-RATER low score cutoff (> 0.1)
previously annotated	4492	4781
extension	125	637
truncation	816*	6553*
internal out-of-frame	10379*	94843*
upstream	6915	30642
downstream	13	450

Figure 2. ORF-RATER-Based Annotations of TIS-Profiling Data

(A) Numbers of different types of ORFs called by ORF-RATER at two different score cutoffs—a high score cutoff (>0.5) and a low score cutoff (>0.1). Truncation and internal out-of-frame numbers are likely high overestimates due to high rates of false positives, indicated with a *.

(B–E) Comparison of standard and TIS-profiling for (B) *PEX32*, which has a likely incorrect TIS annotation. The likely correct (downstream) TIS was called by ORF-RATER, while the previously annotated site was not called. (C) *RSB1*, for which the likely correct TIS is upstream of the previously annotated site. (D) *ICE2*, which has a previously uncalled uORF identified by ORF-RATER. (E) *CKB2*, which has a previously uncalled 5′ extended ORF with a non-AUG TIS identified by ORF-RATER.

It is not surprising that ORF-RATER was generally successful at calling annotated canonical ORFs because the approach trains on this set. To assess its success in identifying unconventional translation products from our dataset, we examined ORF-RATER calls for the few previously well-characterized non-canonical ORFs, which includes 17 AUG-initiated uORFs, 6 near-cognate-initiated extensions, and 6 AUG-initiated alternate isoforms (Table S3). Among this set, the high score cutoff (>0.5) was sufficiently sensitive to detect 71% (12/17) of the known AUG-initiated uORFs and 67% (4/6) of AUG-initiated alternate ORF isoforms but failed to detect 3 of the 6 (50%) known near-cognate-initiated 5′ extended ORFs. We could detect all but one of these cases (83%) when using a lower ORF-RATER score cutoff (>0.1), which also slightly increased the detection of known AUG-initiated uORFs to 77% and AUG-initiated alternate ORFs to 83%. To increase the likelihood of detection of non-canonical ORFs, we used the lower score cutoff for further analyses, which resulted in the provisional annotation of 133,125 non-canonical ORFs in several classes (Figure 2A). This number was much higher than we expected to represent true translated regions, and so we investigated each class in more detail.

Case-by-case investigation of read patterns in the TIS-profiling and standard ribosome profiling data revealed substantial variability in apparent false-positive calls between different ORF categories. A very high proportion of newly called internal ORFs (both truncations and out-of-frame; Figure 2A) are likely to be false positives, based on visual analysis of the LTM data (such as for *SIN3* and *CDC15*; Figures S2E and S2F), and the fact that there were a median of 16 internal ORFs called per annotated gene (score > 0.1; Figure S2G). This high rate of apparent false positives is likely due to residual translation elongation inhibition at the concentration of LTM used in our method, resulting in background ribosome footprints within translated ORFs that erroneously result in internal TIS calls. While real internal initiation sites are expected to exist within these calls, the experimental and detection conditions here were not able to systematically separate true from false positives. In contrast to internally initiated ORFs, manual visual analysis of the data for extensions and downstream ORFs called by ORF-RATER suggested that ORF-RATER calls of these classes of non-canonical ORFs are highly specific. We concluded that our analytical conditions are suitable to detect both canonical and non-canonical ORFs, with the exception of internal ORFs. We therefore excluded both out-of-frame internal ORFs and in-frame internally initiated truncations from further analyses, and the ORF-RATER calls from these categories should be interpreted cautiously.

The remaining non-canonical ORFs that were confidently called at the low score cutoff included 637 N-terminal extensions (akin to *ALA1*; Figure 1G), 30,642 uORFs, and 450 downstream ORFs in which translation initiates within predicted 3′UTR regions (Figure 2A). Traditional ribosome profiling had previously predicted translation from some of these unannotated ORFs, but as expected, some were sensitively detected only with analysis incorporating the TIS-profiling data. Newly identified non-canonical ORFs included uORFs (for example, *ICE2*; Figure 2D), N-terminal extensions (for example, *CKB2*; Figure 2E), and downstream ORFs. We further refined the N-terminal extension class based on length. A cutoff of greater than 10 amino acids was chosen based on the minimum length predicted for function, such as for a targeting signal or binding domain (Figure S3A; Almagro Armenteros et al., 2019; Fukasawa et al., 2015). Excluding AUG-initiated extensions, many of which are likely to represent misannotations (as for *RSB1;* Figure 2C), left 231 extensions, representing 160 unique genes, as some genes contained multiple predicted extensions (Figure S3B; Table S4; this number was ultimately adjusted to 149 based on misannotations discovered through conservation analysis).

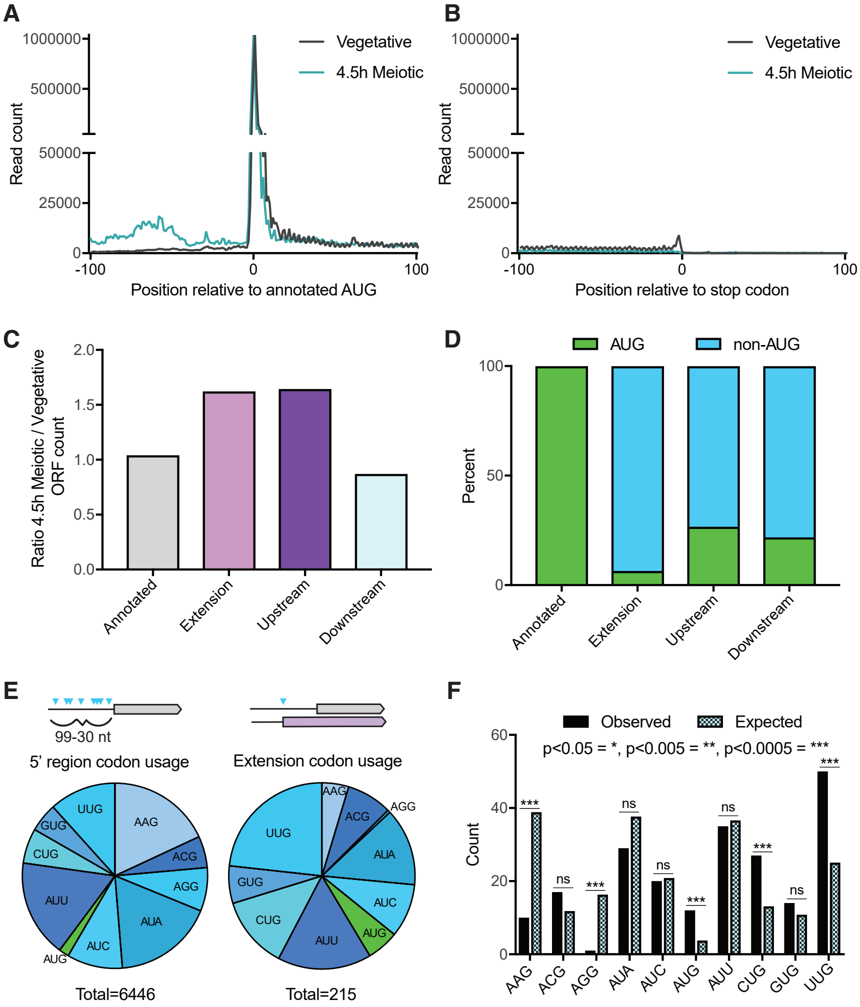


Figure 3. Specificity of uORF and N-Terminal Extension Translation Is Influenced by Condition and Start Codon Identity

(A) Metagene plot of read counts from vegetative exponential and 4.5 h time points, 100 nucleotides upstream and downstream of annotated AUG start codons. Reads are normalized to aligned reads for that time point. Increased read density is observed for the meiotic time point upstream of annotated start codons, but not after.

(B) Metagene plot of read counts from vegetative exponential and 4.5 h time points, 100 nucleotides upstream and downstream of annotated stop codons. Reads are normalized to aligned reads for that time point.

(C) Relative numbers of ORFs from different ORF categories, comparing the 4.5 h meiotic time point with vegetative exponential. More 5′ extensions and upstream ORFs are called in the meiotic time point, while annotated and downstream ORFs are similar between the two conditions.

(D) Percent of AUG versus non-AUG TISs for different ORF types. Annotated ORFs all have AUG start codons, while 5′ extensions, upstream, and downstream ORFs have primarily non-AUG TISs.

(E) Distribution of AUG and non-AUG start codon usage 99–30 nucleotides (nt) upstream of annotated AUG start codons for all possible TISs (left) and called 5′ extensions (right). Of the 6,446 sites possible in 5′ regions, 215 are observed to initiate translation of 5′ extensions called by ORF-RATER.

(F) Near-cognate codon usage for called extensions (observed) compared with relative abundance of all possible near-cognate codons within upstream regions (expected). Expected distribution is derived from counts of all possible TISs in the 99–30 nt upstream of annotated AUG start codons. p values calculated by Fisher's exact test, with *p < 0.05; **p < 0.005; ***p < 0.0005; and ns = not significant.

Translation of uORFs and 5′ Extended ORFs Is Enriched in Meiosis

Increased ribosome footprints within 5′ leader regions were previously observed in meiosis in yeast (Brar et al., 2012). To determine whether TIS-profiling detected increased meiotic translation initiation within 5′ leaders, we compared metagene profiles surrounding annotated start codons for vegetative exponentially growing cells to a representative mid-meiotic time point (4.5 h). This, indeed, revealed a meiosis-specific increase in translation initiation 5′ of annotated start codons (Figure 3A) but no difference between the vegetative and meiotic LTM-based ribosome footprints in regions surrounding annotated stop codons (Figure 3B). The increased read density in 5′ leaders during meiosis could reflect an increase in translation of either uORFs or 5′ extended ORFs. To investigate this, we compared the types of ORFs called in the vegetative exponential time point to the mid-meiotic time point. The calls for both uORFs and 5′ extensions are increased in meiosis, while the number of annotated and downstream ORFs are similar between the two conditions (Figure 3C). Although annotated ORFs all begin with an AUG start codon, extensions and uORFs initiate at near-cognate start codons in 93.6% and 73.3% of cases, respectively (Figure 3D). The translation of both uORFs and N-terminally extended proteins results from increased translation initiation within 5' leaders, but the consequences of these two classes of non-canonical translation are fundamentally different. Translation initiation at the start codon of a uORF may regulate the translation of the downstream canonical ORF or produce a small peptide, whereas translation initiation at the start codon for an N-terminal extension generates a modified protein product with potentially distinct function (Hood et al., 2009; Morris and Geballe, 2000). For example, the extended isoform of Ala1 is targeted to the mitochondria rather than the cytosol, providing alanyl-charged tRNAs for mitochondrial translation (Tang et al., 2004). Our TIS-profiling data identified translation of the known extensions at the *ALA1*, *YMR31/KGD4*, *HYR1/GPX3*, *TRZ1*, and *HFA1* loci, as well as 155 other genes, which we proceeded to evaluate in more detail (Table S4; Heublein et al., 2019; Kritsiligkou et al., 2017; Monteuuis et al., 2019; Suomi et al., 2014; Tang et al., 2004).

Non-AUG-Initiated Isoform Translation Is Specific and Does Not Preclude Canonical Isoform Translation

The low number of AUG-initiated N-terminal extensions identified here (Figure 3D) likely reflects the fact that traditional genome annotations selected the longest AUG-initiated ORF at a locus as the one most likely to be translated. We wondered whether these extended ORFs generally represented an additional translated ORF or whether these were the sole translated ORF at these loci. Consistent with the former, 85% (136/160) of genes encoding extended ORFs had a corresponding annotated ORF that was called by ORF-RATER. Of the 24 that were not called, 17 show evidence of translation initiation at the annotated AUG-initiation site in our TIS-profiling data but were not called by ORF-RATER (Table S4). Four of the remaining seven are misannotations, similar to *RIM11* (Figure S2D), and one (*YPL034W*) includes a likely frameshifting event (see note in Table S4). This leaves only 2 cases in which the near-cognate-initiated extension is the sole or predominant translation product: *HFA1*, which is indeed the only characterized gene in yeast in which a non-AUG-initiated product is thought to be the primary translation product (Suomi et al., 2014) and *YNL187W*, a poorly characterized gene. We concluded from these analyses that loci that encode near-cognate-initiated extended protein isoforms generally express them in concert with the canonical AUG-initiated isoform.

Given the prevalence of translation initiation within 5′ leaders in meiosis, most of which is at near-cognate start codons, we wondered if generally less stringent TIS selection in meiotic conditions might produce 5′ extended ORFs non-specifically. To estimate the number of theoretically possible N-terminal extensions based on non-specific “sloppy” initiation, we calculated the number of in-frame cognate and near-cognate start codons that fall between 99–30 nucleotides upstream of annotated start codons and do not have an in-frame stop codon before the canonical start codon. We chose this region to account for the average length of yeast 5′ UTRs and to include only the potential ORF extensions that would be expected to be long enough to confer new biological function (>10 additional amino acids; David et al., 2006; Nagalakshmi et al., 2008). We found 6,446 possible sites, only 3.3% of which have evidence of being used to initiate translation in our TIS-profiling dataset. This indicates highly stringent selection of certain near-cognate TISs to produce N-terminal extensions.

Some of this specificity resulted from preferential initiation at certain near-cognate codons (Figures 3E and 3F). The codons that we found to be enriched for initiation of 5′ extended ORFs, including CUG and UUG, have been previously shown through *in vitro* assays to be the most efficiently initiated near-cognate codons (Kolitz et al., 2009). The preference for specific near-cognate codons alone could not explain the small percentage of potential start codons in 5′ leaders used to translate extended ORFs, so we also searched for evidence that start codon context influenced the set of used versus theoretically possible TISs. We found only weak enrichment for the optimal (Kozak-like) motif found around annotated AUG-initiated ORFs (Figure S4A; Kozak, 2002, 1999, 1984, 1978), which is consistent with previous reports of differences between optimal contexts around near-cognate and AUG start codons (Chang et al., 2010). We were unable to identify any simple context cues that were enriched specifically in the translated near-cognate TISs (data not shown), suggesting that other, yet-to-be-determined features define the specific start codons used for translation initiation of extended isoforms.

Predicted N-Terminal Extensions Can Be Detected by Mass Spectrometry

To determine whether the identified N-terminally extended protein isoforms are abundant in meiosis, we re-analyzed a previously generated quantitative mass spectrometry dataset, searching for peptides that uniquely arise from the N-terminally extended regions (Cheng et al., 2018). Our search set contained all extensions with an ORF-RATER score of 0.1 or higher, an extension length greater than ten amino acids, and initiation at a near-cognate start codon (Figure S3A). Of the 160 unique genes searched in this way, seven showed at least one peptide originating from the extension. Three of the seven had ORF-RATER scores well below the high score cutoff of 0.5 (Figure 4A), suggesting that our choice of the lower cutoff to define extended isoforms is appropriate. For the majority (69%), the annotated isoform was quantifiable, but we detected extension-derived peptides for only 6.25% of those searched (average extension length of 25 amino acids). By comparison, a parallel search for peptides within the first 25 amino acids of annotated proteins identified 43.2% of cases. The high degree of discrepancy in detection between these two classes, and the fact that we only identified two of the six established extensions (*HYR1* and *YMR31*), suggests that near-cognate-initiated extended proteins, as a class, may be lowly expressed relative to canonical proteins.

Extended Protein Isoform Levels Are Lower than Expected Based on TIS-Profiling Peak Height

To probe the relative levels of near-cognate-initiated and canonical protein isoforms, we characterized in more detail the expression of Ymr31, a subunit of the mitochondrial alpha-ketoglutarate dehydrogenase recently found to be produced from both a canonical AUG and upstream UUG start codon (Heublein et al., 2019). We chose Ymr31 for this analysis for three reasons. First, mass spectrometry had detected multiple peptides from this extension, indicating that the extended protein isoform was likely to be abundant in our conditions. Second, it was the highest scoring extension called by ORF-RATER. Lastly, the discrepancy in size between the GFP-tagged small canonical protein (41 kDa) and the relatively large extended protein (44 kDa) made the two isoforms readily distinguishable by western blot. This last property, which was rare among genes with extended isoforms, was especially valuable in enabling *in vivo* analyses of isoform regulation.

To evaluate relative expression levels of the two *YMR31*-encoded isoforms, a C-terminally GFP-tagged version of this protein was expressed with either the wild-type (*WT*) start codon, the annotated ATG start-codon-encoding site mutated to an alanine-encoding codon (*M1A*), or a stop codon inserted directly upstream of this ATG (*ustop*). In *M1A* cells, the extension is expected to be the only isoform translated, and cells carrying the *ustop* construct are expected to only produce the canonical AUG-initiated isoform (Figure 4B). Samples were collected in vegetative cells, and at 3 and 6 h after inducing meiosis. In *YMR31-M1A* and *YMR31-ustop* cells, only the extended or canonical forms were observed, respectively, confirming our predicted *YMR31* ORF annotations (Figures 4C and S5A). The extended form of Ymr31 was 10 times lower in abundance than the canonical form in *WT* cells by western blot analysis

A

Extended gene	Peptide Sequence	Start position	End position	ORF-RATER score
CAF17	GSCLAFPLD**M**FISR	7	20	0.223
CKB2	QI**M**GSRSENVGTVTR	11	25	0.835
FOL1	**V**IASSLINK	1	9	0.664
HYR1	FI**M**SEFYK	10	17	0.121
IMA3	**I**CSIHQSK	1	8	0.588
SMT3	LPPVAYRTK	2	10	0.113
YMR31	**M**NSLKQTIK	1	9	0.960
	IKVQEDRE**M**IATPIR	23	37	
	VQEDRE**M**IATPIR	25	37	
	E**M**IATPIR	30	37	

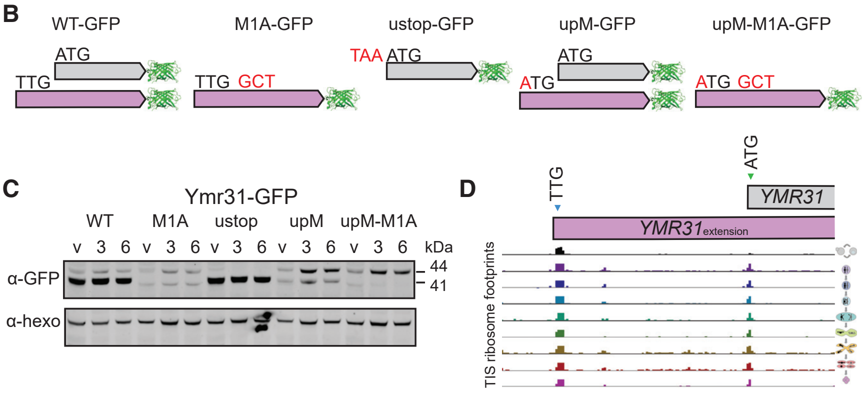


Figure 4. The Abundance of Near-Cognate-Initiated Isoforms Is Not Reflective of TIS-Profiling Peak Height

(A) Extensions with peptides identified that match to the extension-specific region of the protein from a meiotic mass spectrometry dataset. The annotated methionine is highlighted in green, and the extension start codon is highlighted in blue where relevant.

(B) Cartoon of tagging and mutagenesis strategy for validation of extensions. All constructs include a C-terminal GFP tag. Mutations include: *M1A* to mutate the annotated methionine to alanine, *ustop* to mutate the codon upstream of the annotated start codon to a stop codon, and *upM* to mutate the extension's upstream non-AUG start codon to a methionine.

(C) Western blot of Ymr31-GFP showing the *WT* construct with two bands corresponding to translation of the extension (44 kDa) and annotated ORF (41 kDa). *M1A* and *ustop* constructs show translation of the extension and annotated ORF individually, respectively. *upM* and *upM-M1A* constructs show an increase in the extension isoform. Samples were taken in vegetative exponentially growing cells (v) and at 3 and 6 h after addition to sporulation media. Anti-hexokinase (α-hexo) is a loading control. The band around 40 kDa visible in the *M1A* construct is of unknown identity and may represent translation from a downstream AUG.

(D) TIS-profiling of *YMR31*, showing ribosome footprints at the time points indicated in Figure 1-B, with the extension (TTG) and annotated (ATG) start-codon-encoding sites indicated.

(Figure S5A), which is in marked contrast with the TIS-profiling data showing over eight times higher ribosome footprint read density at the near-cognate initiation site than at the canonical start codon (Figures 4D and S6A).

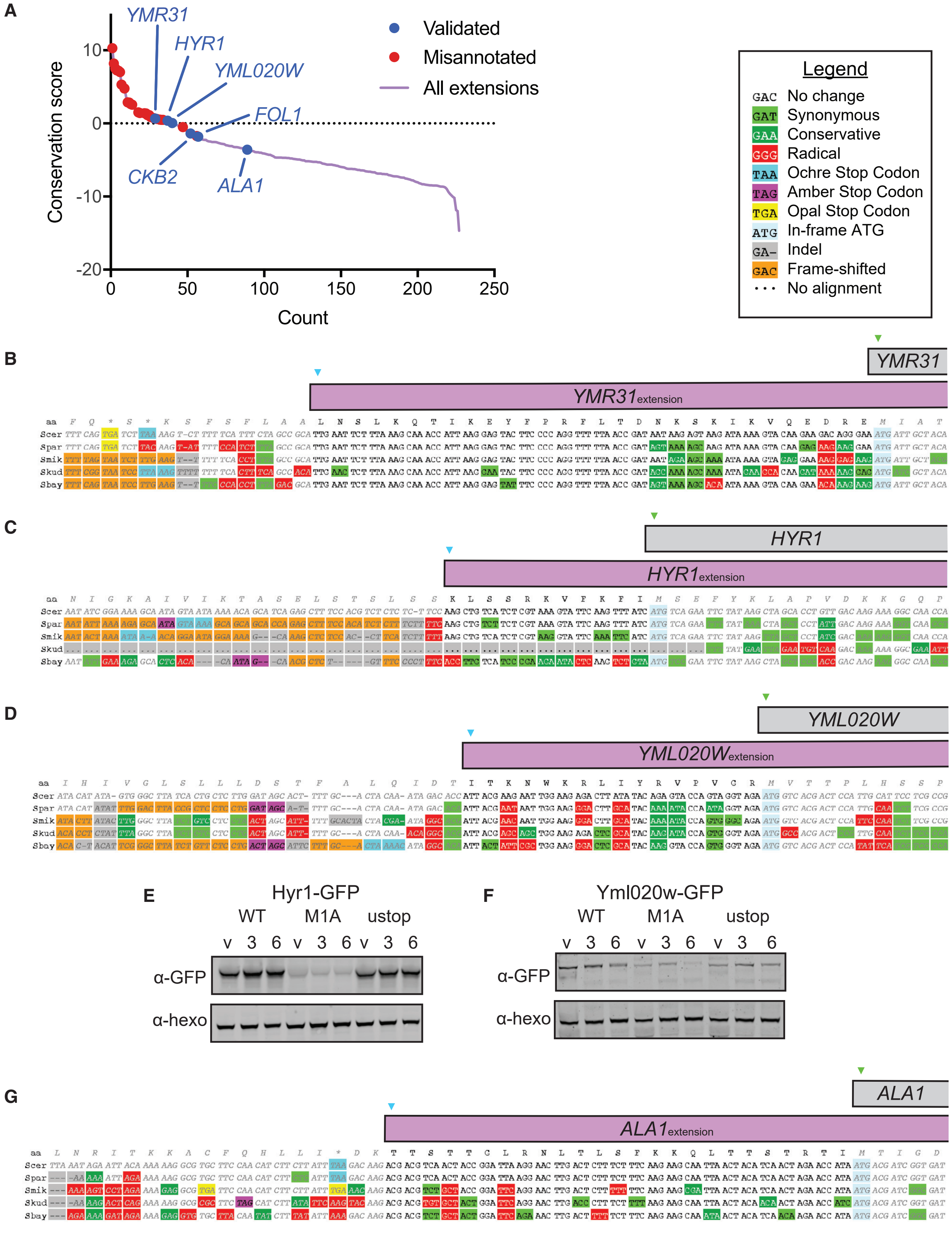
Mutation of the near-cognate-encoding initiation codon to ATG resulted in higher levels of the N-terminally extended Ymr31 isoform, either with (*upM-M1A*) or without (*upM*) mutation of the canonical start codon (Figure 4C). This suggested that the native near-cognate TIS is used inefficiently for translation initiation relative to AUG, consistent with *in vitro* and mutagenesis experiments comparing AUG and near-cognate initiation (Chen et al., 2008; Kolitz et al., 2009). This result also suggested that the peak height observed by TIS-profiling at near-cognate and AUG codons may not be comparable. This may be due to differences in the ability of LTM to inhibit the two different types of post-initiation ribosome complexes or in their timespan of initiation. We also considered the possibility that near-cognate-initiated proteins might be subject to proteasome-mediated degradation, but at least for Ymr31, we did not observe an increase in the alternate isoform in cells in which proteasome activity was inhibited by MG132 (Figures S6B and S6C).

We further investigated whether the discrepancy between protein levels and TIS peak height indicated that TIS-profiling peaks were not quantitatively predictive of translation levels. This was not generally true, at least for AUG-initiated ORFs, as the height of TIS peaks appeared to reflect known regulation patterns during meiosis for characterized genes. Across annotated ORFs, there was a positive association between the read count at the TIS for TIS-profiling and the density of ribosome footprints over ORFs for standard ribosome profiling (Figures S6D and S6E). This was seen by comparisons of individual time points (Figure S6E), as well as by calculating correlation scores for each gene across all time points (Figure S6D). Individual examples, such as Rec8 (Figure 1E), showed a strong correlation between TIS-profiling peaks and standard profiling reads (Pearson correlation coefficient = 0.833), and correlations were significantly enriched for positive values compared with a random distribution of genes (Figure S6D). This is consistent with a study using a similar approach in mammalian cells that suggested ribosome footprint peaks at AUG start codons following LTM treatment quantitatively reflect translation initiation levels (Lee et al., 2012). We concluded that our TIS-profiling protocol reports at least weakly quantitative values for translation initiation levels at AUG start codons but that TIS-profiling peak heights at near-cognate start codons are much higher than expected based on our poor detection of near-cognate-initiated peptides by mass spectrometry, as well as the inferred translation levels from western blotting analysis of the two Ymr31 isoforms.

5′ Extensions Are Poorly Conserved as a Class

To probe the likelihood that the N-terminally extended protein isoforms have conserved functionality within *Saccharomyces*, we analyzed the evolutionary protein-coding potential of the extensions using PhyloCSF, which reports a score indicating whether the local alignment of a region is more likely under coding or non-coding models of evolution (Lin et al., 2011). Positive

Please cite this article in press as: Eisenberg et al., Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.011



(legend on next page)

scores are more likely in conserved coding regions (Figure 5A). We noted that among the highest scoring cases were 11 in which the putative extension was a misannotation resulting from sequencing errors or strain-specific stop codons or indels, leaving 149 genes with apparent true near-cognate-initiated extensions (Table S4). Alignments of individual true extensions illustrate the degree of conservation, which for Ymr31 is high, reflected in its high PhyloCSF score (Figure 5B). We further evaluated two true extensions with high PhyloCSF scores, for the genes *HYR1* and *YML020W* (Figures 5C and 5D). In these cases, as well as for nearly every other extension-encoding gene we examined, the size difference between the extended and canonical isoform was too small to detect by western blot for the *WT* construct, making the *M1A* construct critical in confirming the expression of the extended isoform. For *HYR1*, using the tagging strategy previously described, we observed a lowly expressed band corresponding to the extended isoform in extract from cells carrying the *HYR1-M1A* mutant construct (Figures 5E and S5B). Similarly, we detect an N-terminally extended isoform of Yml020w in cells carrying the *YML020W-M1A* construct (Figures 5F and S5C).

The majority of extensions analyzed had scores below zero, suggesting a lack of conserved functionality (Figure 5A). In some cases, however, the extension might have conserved function but nonetheless have a negative PhyloCSF score because the amino acid sequence is under only weak purifying selection or is subject to an atypical constraint. An example of the latter is *ALA1*, where the ACG start codon and the reading frame are conserved in five species but the extension itself had a negative PhyloCSF score (−3.587; Figures 5A and 5G). A possible explanation is that the mitochondrial targeting function of the extension is present in the other species but imposes a constraint that PhyloCSF is not able to detect.

Transcripts with Canonical Start Codon Mutations Are Nonsense-Mediated Decay Targets

The length of the extended Ala1 protein relative to the canonical isoform was too small to allow both versions to be detected by western blotting and, because the start codon at the endogenous locus could not be manipulated to isolate production of the extended isoform without affecting cell fitness, GFP reporters (*ALA1^{GFP}*) were constructed to further investigate translation from this gene (Figure 6A). When the canonical start codon was present in the reporter (*ALA1^{GFP}-WT*), both Ala1 reporter isoforms were observed (Figures 6B, 6C, and S5D). The canonical Ala1 reporter isoform could be detected alone in extract from cells carrying the *ALA1^{GFP}-ustop* construct (Figures 6B and 6C). Surprisingly, in cells carrying the *ALA1^{GFP}-M1A* construct, however, we could not detect production of either protein isoform (Figures 6B and 6C). The dramatic difference in production of the extended reporter with and without the canonical start codon mutation cannot be explained by inefficient near-cognate usage alone. The difference we observed exceeded even the ~10–100-fold decrease we would expect based on inefficient near-cognate usage (Chen et al., 2008; Clements et al., 1988; Kolitz et al., 2009). We further found that the mRNA levels of GFP from the *ALA1^{GFP}-M1A* construct were dramatically decreased relative to the *ALA1^{GFP}-WT* construct (Figure 6D). This led us to explore the possibility that the nonsense-mediated decay (NMD) pathway degrades transcripts from mutated constructs lacking the canonical in-frame start codon, likely due to efficient translation initiation at a downstream out-of-frame AUG that results in early translation termination (Figure S7A). Consistent with this hypothesis, we observed that both mRNA and protein levels of the *ALA1^{GFP}-M1A* reporter construct increased in an NMD-deficient mutant background (*upf1Δ*), although not to the level of the extended isoform in the *ALA1^{GFP}-WT* reporter construct (Figures 6B–6D).

In addition to the *ALA1* reporters, several other *M1A* constructs showed little to no tagged protein in otherwise wild-type cells. This was consistent with our findings for the extended isoform of Hyr1, which was detected in our mass spectrometry dataset (Figure 4A) but was detected at extremely low levels in cells carrying the *HYR1-M1A* construct (Figure 5E; Kritsiligkou et al., 2017). Analysis of the *HYR1-M1A* construct in *upf1Δ* cells revealed increased levels of the N-terminally extended protein and *HYR1* mRNA (Figures 6E-6G, and S5E), consistent with NMD targeting of the mutant transcript. Analyses in the *upf1Δ* background allowed validation of additional N-terminally extended isoforms predicted by TIS-profiling-based annotation. These include *CKB2*, encoding the casein kinase beta subunit, and *FOL1*, which encodes a folic acid synthesis pathway enzyme. For these genes, like *ALA1* and *HYR1*, the mutant construct lacking the AUG start codon(s) (*M1A* for *CKB2*; *M1A M20A* for *FOL1*, see below) was not detected with *UPF1* present but was in *upf1Δ* cells (Figures 6E-6G).

For the two examples that were robustly detected in a *WT* background, Ymr31 and Yml020w, little increase in protein levels from *M1A* constructs in *upf1Δ* cells was seen for the extended versions (Figures 4C and 5F). Consistently, *YMR31-M1A* and *YML020W-M1A* mRNA levels were not dramatically decreased in *WT* cells relative to unmutated constructs (Figure 6G). The difference between cases like *CKB2*, *FOL1*, *ALA1,* and *HYR1*, in which mutation of the canonical start codon leads to high mRNA degradation by NMD, and *YMR31* and *YML020W*, in which it does not, is intriguing, as all loci produce the extended

Figure 5. Most 5′ ORF Extensions Are Poorly Conserved

(A) Plot of PhyloCSF conservation scores for 5′ extended ORFs. Misannotated extensions are shown with red dots, and validated extensions are shown with blue dots, including three previously validated extensions (*YMR31*, *HYR1,* and *ALA1*). The additional "validated" extensions (*YML020W*, *CKB2,* and *FOL1*) were validated in this study.

(B–D) Alignments showing level of conservation for *YMR31* (B)*, HYR1* (C), and *YML020W* (D), all of which have positive conservation scores.

(E) Western blot of Hyr1-GFP including *WT*, *M1A,* and *ustop* constructs. Samples were taken in vegetative exponentially growing cells (v) and at 3 and 6 h after addition to sporulation media.

(F) Western blot of Yml020w-GFP including *WT*, *M1A*, and *ustop* constructs. Samples were taken in vegetative exponentially growing cells (v) and at 3 and 6 h after addition to sporulation media.

(G) Alignment showing level of conservation for *ALA1*, which has a negative conservation score.

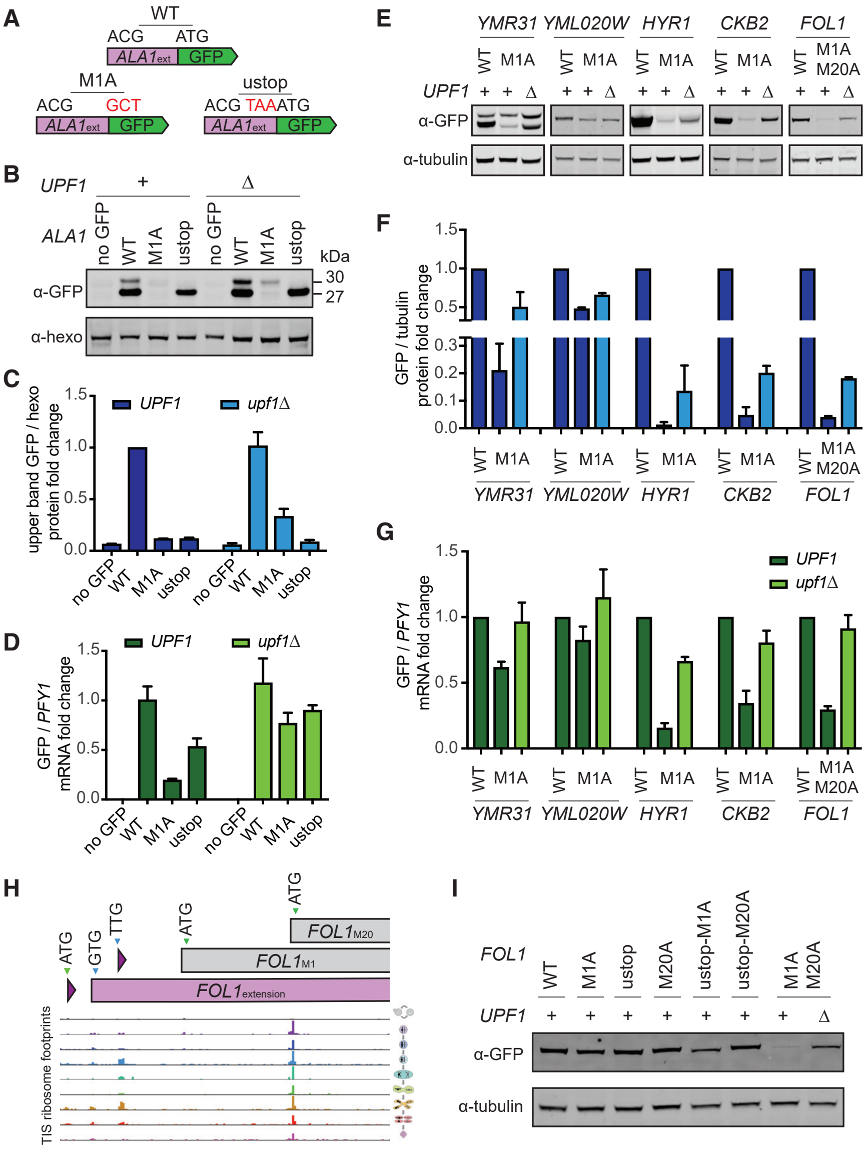


Figure 6. Extended ORF-Encoding Transcripts Lacking Canonical AUG Start Codons Are Degraded by NMD

(A) Schematic for *ALA1* tagging strategy, using a reporter including the region upstream of the ATG, and either including (*WT*) or not including (*M1A*) the in-frame ATG in front of the GFP, and a mutant with a stop codon upstream of the in-frame ATG (*ustop*).

(B) Western blot for Ala1^{GFP} reporters in *UPF1* and *upf1*Δ vegetative cells. The band corresponding to the extension (30 kDa) can be seen in the *WT* construct, but is not seen in the *M1A* construct in *UPF1* cells. In a *upf1*Δ background, the *M1A* construct now shows the extended protein.

(C) Western blot quantification of Ala1-GFP upper band intensity from Figure 6B normalized to hexokinase, for 3 replicates.

(D) qPCR fold change of *ALA1-GFP* transcript relative to *PFY1*, for 3 replicates. The level of the *M1A* mRNA in *UPF1* cells is low due to NMD acting on this transcript, and this effect is lessened in the *upf1*Δ background.

(E) Western blot analysis of Ymr31-GFP, Hyr1-GFP, Fol1-GFP, Ckb2-GFP, and Yml020w-GFP for the *WT* and *M1A* constructs in *UPF1* cells and the *M1A* construct in *upf1*Δ cells at 4.5 h in meiosis.

(F) Western blot quantification of GFP-tagged proteins from Figure 6E normalized to tubulin, for 3 replicates.

(G) qPCR fold change of *GFP* transcripts relative to *PFY1* for 3 replicates from strains from Figure 6E.

(H) TIS-profiling of *FOL1*, showing ribosome footprints at the time points indicated in Figure 1-B, with the positions of the extension (GTG), M1 (ATG), and M20 (ATG) start-codon-encoding sites indicated.

(I) Western blot analysis of Fol1-GFP for constructs including mutations at the annotated methionine (M1) as well as a methionine at position 20 (M20), indicating that translation can begin at three in-frame start codons.

proteins at lower levels than the canonical protein, and all *M1A* constructs are expected to result in translation of a short out-of-frame ORF that should trigger NMD. Among this group, there is no correlation between the distance from the new presumptive out-of-frame stop codon to the end of the transcript and the strength of NMD, as measured by the percent abundance of *M1A* relative to *WT* mRNA (Figures S7A–S7C), although this distance is thought to be a key factor in specifying yeast NMD substrates (reviewed in Hug et al., 2016). We did, however, observe a moderately positive association between the distance of the transcription start site to the location of the first downstream AUG (which is out of frame) in the *M1A* constructs and the degree of NMD (Figure S7B).

The *FOL1* Locus Encodes Three Protein Isoforms

Among the 149 genes identified as having alternate N-terminally extended isoforms by our TIS-profiling analysis, several cases appeared to have more than two alternate TISs. At the *FOL1* locus, for example, our data reveal translation initiation at two uORF start codons, an upstream in-frame GUG start codon (producing an N-terminally extended isoform), the annotated AUG start codon, and an AUG 19 codons downstream of the annotated AUG (Figure 6H). The relative usage of these start codons, as gauged by TIS-profiling peak height, differed among the conditions that we assayed. The three GFP-tagged Fol1 isoforms predicted based on these data could not be resolved by western blotting, but high Fol1 protein levels were observed in cells carrying either a *ustop-M1A* or *ustop-M20A* construct, confirming protein production from the downstream AUG (M20) alone and the canonical AUG (M1) alone, respectively (Figure 6I). *FOL1-M1A-M20A* cells showed a drastic decrease in *FOL1* mRNA and protein levels that were partially rescued in *upf1Δ* cells, confirming translation from the upstream GUG identified by TIS-profiling (Figures 6H, 6I, and S5F). Such coding complexity is

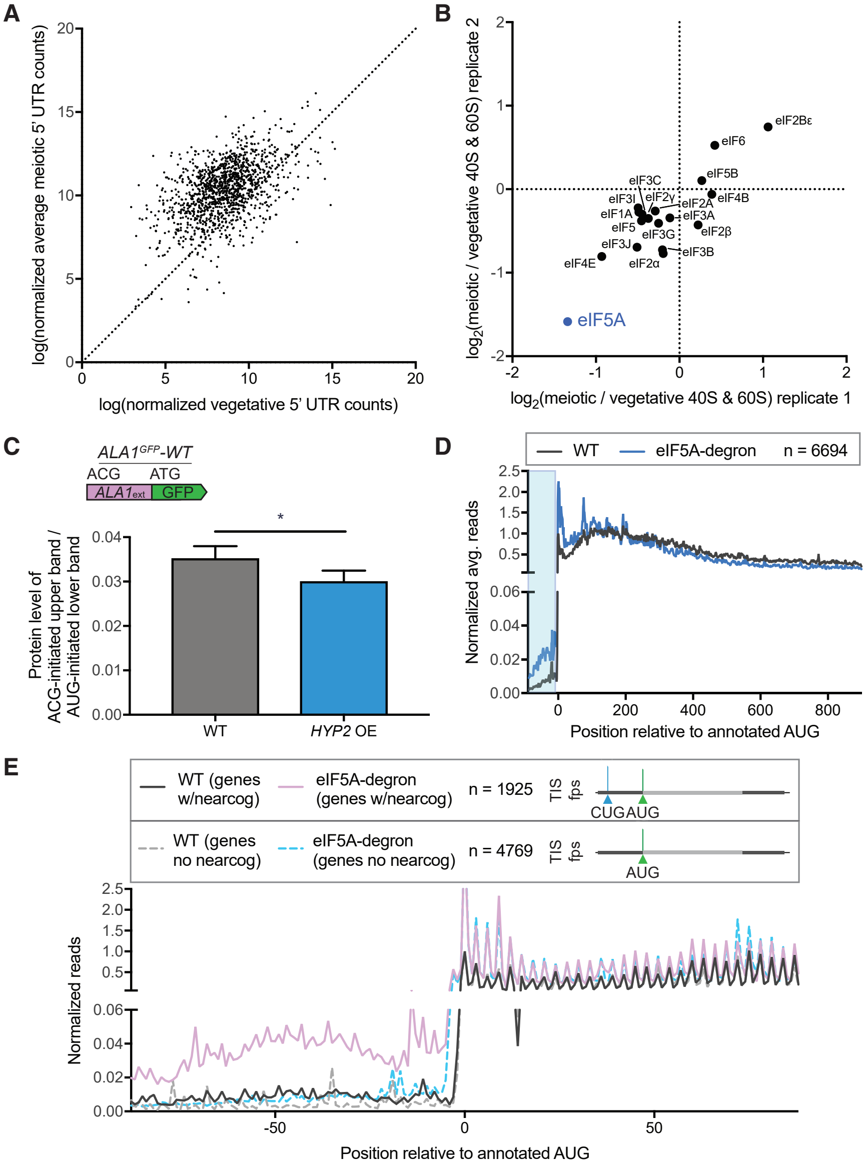


Figure 7. eIF5A Levels Regulate Pervasive Non-AUG-Initiated Translation

(A) Comparison of vegetative and average meiotic 5′ read density measurements.

(B) Enrichment of translation factors, comparing meiotic and vegetative samples for two replicates, determined by quantitative mass spectrometry of 40S and 60S ribosomal subunits isolated by sucrose gradient centrifugation of cell extract from meiotic and vegetative cells.

(C) Western blot quantification of *ALA1^{GFP}-WT* reporter in meiosis with copper induction in strains containing or lacking a copper-inducible overexpression (OE) *HYP2* allele. Non-AUG-initiated GFP upper band protein is normalized to AUG-initiated GFP lower band protein, which runs as a doublet. Both bands were used for quantification. The decrease seen with *HYP2* OE is significant ($p < 0.0146$, 4 replicates).

(D) Metagene plot of normalized average reads from *WT* (black) and eIF5A-degron (blue) samples, 100 nt upstream and 900 nt downstream of annotated AUG start codons for all genes (n = 6,694). Reads are normalized to *WT* at position zero and averaged across three nucleotides. Ribosome profiling data were re-analyzed from a previous study (Schuller et al., 2017). The area boxed in blue highlights the increased reads seen for the eIF5A-degron relative to *WT* in 5′ leader regions.

(E) Metagene plot around annotated start codons comparing genes with 5′ near-cognate-initiated ORFs annotated by ORF-RATER (n = 1,925, *WT* [genes w/nearcog]: solid black and eIF5A-degron [genes w/nearcog]: solid purple) and genes that do not contain 5′ near-cognate-initiated ORFs (n = 4,769, *WT* [genes no nearcog]: dotted gray and eIF5A-degron [genes no nearcog]: dotted light blue). Increased reads in the 5′ region are seen only in the eIF5A-degron samples for genes containing ORFs translated from near-cognate start codons in 5′ regions, based on TIS-profiling.

surprising to find in a eukaryote as simple as budding yeast and would not have been readily identifiable without TIS-profiling data.

eIF5A Levels Alter Non-AUG TIS Usage in Yeast Meiosis

The preferential translation of non-AUG-initiated ORFs in meiotic cells (Figure 3C) and the increase in TIS-profiling reads in 5′ leader regions in meiotic time points relative to vegetative cells suggest condition-specific modulation of translation initiation (Figure 7A). To identify candidates for this regulation, we performed quantitative mass spectrometry of 40S and 60S ribosomal subunits isolated by sucrose gradient centrifugation of cell extract from meiotic and vegetative cells. We found that eIF5A (encoded by *HYP2* in yeast) was strongly and reproducibly disenriched in meiotic relative to vegetative samples, indicating decreased ribosome association of this factor in meiotic cells (Figure 7B). Many of the initiation factors found to associate with the 40S and 60S subunits have lower overall levels in meiotic cells, but the disenrichment of ribosome association seen for eIF5A is greater than could be explained by its overall decrease in abundance relative to vegetative cells (Figure S8). eIF5A has recently been shown to influence translation elongation and termination (Gregio et al., 2009; Henderson and Hershey, 2011; Saini et al., 2009; Schuller et al., 2017) but was initially identified for activity in promoting a late stage of translation initiation *in vitro* (Benne and Hershey, 1978; Kemper et al., 1976; Lopo et al., 1986; Schreier et al., 1977). A CRISPRi screen in human cell lines identified eIF5A as a factor that enhanced translation of the CUG-initiated, N-terminally extended isoform of MYC when transcriptionally repressed (Manjunath et al., 2019). In this context, low eIF5A levels are thought to impair translation elongation, leading to ribosome queuing, which contributes to increased initiation at upstream near-cognate sites (Ivanov et al., 2018; Manjunath et al., 2019).

To test whether increased expression of eIF5A might alter the high near-cognate TIS selection that we observe in meiosis, we placed *HYP2* under a copper-inducible promoter and quantified the change in the non-AUG-initiated form of *ALA1^{GFP}-WT* in meiotic cells upon Hyp2 induction. We see a small but significant decrease in non-AUG-initiated translation, dependent on increased levels of *HYP2* (Figures 7C, S5G, and S5H), suggesting that lower eIF5A is at least partly responsible for the increased translation from near-cognate codons seen in meiotic cells. The small effect seen here is not surprising, as simply overexpressing eIF5A may not increase the relevant functional pool of this factor, which not only has multiple characterized roles as noted above but is also regulated by hypusine modification (Hershey et al., 1990). Indeed, mass spectrometry data show that Lia1, one of the enzymes responsible for Hyp2 hypusination, is dramatically decreased in meiotic cells, which would be expected to lead to low Hyp2 activity (Figure S8). Moreover, our data suggest that meiotic ribosomes show changes in association with multiple translation initiation factors relative to vegetative cells, some of which are known to be involved in TIS selection (Figure 7B; reviewed in Hinnebusch, 2011; Kearse and Wilusz, 2017). It may be that multiple changes in concert mediate the large increase in near-cognate initiation seen during meiosis.

A previously published vegetative ribosome profiling dataset (Schuller et al., 2017) was examined for evidence that the loss of eIF5A in a non-meiotic context mimicked the high near-cognate initiation we observe in meiosis. Metagene analysis of ribosome footprint reads over all genes was consistent with the elongation defect previously reported within ORFs (Schuller et al., 2017) and also revealed enrichment in 5′ leader reads in cells depleted for eIF5A relative to *WT* controls, supporting the reported role for this factor in repressing translation from 5′ leader TISs (Figure 7D; Manjunath et al., 2019). When the set of genes we identified as having near-cognate-initiated translation in 5′ leaders in our TIS-profiling data was separated from the set that do not, a dramatic difference was evident. The set that we identified as having near-cognate initiation in 5′ leaders in meiosis (n = 1,925) are enriched for ribosome footprint reads upstream of canonical start codons in eIF5A-depleted mitotic cells, but there was *no difference* seen for the set that we did not identify as having near-cognate initiation in 5′ leaders (n = 4,769), relative to *WT* cells (Figure 7E). This shows that low eIF5A levels alone can lead to selective enhanced near-cognate-initiated translation in *the specific subset* of genes with this non-canonical type of initiation in meiosis. Together, our data point to eIF5A as a factor that contributes to the condition-specific unmasking of near-cognate-initiated alternate protein isoforms in meiosis.

DISCUSSION

Here, we report the first method for globally mapping TISs, and thus defining translated ORFs, in budding yeast. Traditional ribosome profiling has allowed detection of translated regions genome wide, but the combined signal of initiating and elongating ribosomes makes identification of alternative and overlapping ORFs challenging. Ribosome profiling following treatment with a post-initiation translation inhibitor, first applied in mammalian cells, overcomes this issue by isolating sites of translation initiation. This type of approach has not been widely used, likely because of the difficulty of identifying drug treatment conditions that are highly specific to inhibition of initiating ribosomes and the challenges of data analysis in organisms with complex transcript architectures.

Our application of this method in vegetative and meiotic budding yeast cells indicates that genome decoding in this simple eukaryote is much more complex than previously appreciated. The many newly identified ORFs from our analyses indicate the need for substantial revision to genome annotations. We identified, for example, the second case (to our knowledge) in which a yeast locus encodes three distinct proteins (Martin and Hopper, 1994). Whereas decades of study have resulted in the validation of only a handful of non-canonical translation products, our systematic experimental approach defined many cases, including 149 near-cognate-initiated N-terminally extended proteins. This is complementary to previous studies (Table S1) and adds direct experimental evidence for widespread translation initiation at near-cognate codons in budding yeast, especially during meiosis. We also found that protein levels resulting from near-cognate initiation, for N-terminal extensions, are not proportional to peak heights observed by TIS-profiling (as exemplified by Ymr31, compare Figures 4C and 4D). Rather, we detect much lower levels than expected, suggesting fundamental differences between AUG- and near-cognate-initiated translation. Both protein synthesis and degradation could contribute to the low steady-state protein levels, but blocking proteasome degradation did not appear to increase the level of the extended isoform (Figure S6C). We favor a model in which near-cognate-initiating ribosomes pause longer at TISs and are, thus, captured there more efficiently by ribosome profiling. It is also possible that ribosomes initiating at near-cognate and AUG TISs differ in their susceptibility to LTM-based inhibition, leading to preferential capture of reads at near-cognate sites by TIS-profiling.

Although previous studies have identified individual cases of extensions or predicted potential extensions computationally, it has not been possible to experimentally determine the pervasiveness of alternate protein isoforms beginning at non-AUG codons. This has become a recent area of interest, with three of the six established cases in yeast identified in just the last 3 years (Heublein et al., 2019; Kritsiligkou et al., 2017; Monteuuis et al., 2019). One of these studies predicted this class of proteins to be common, based largely on elegant computational analyses (Monteuuis et al., 2019). Our data are consistent with their general prediction, providing the first direct and comprehensive evidence for translation of a large set of N-terminally extended proteins in budding yeast. We also report these proteins to be conditionally unmasked, with their translation enriched in the context of meiosis.

The few known loci that encode extended proteins have been studied either by mutating the upstream near-cognate codon to encode an AUG, or by using a strong promoter to increase production of the extended protein, presumably by necessity due to low natural expression of extended isoforms (Kritsiligkou et al., 2017; Monteuuis et al., 2019). Conservation and mass spectrometry analyses of N-terminally extended proteins provided evidence for function and stability of only a small subset of the proteins resulting from the alternate isoforms that our TIS-profiling predicted. Because the detection efficiency of both approaches has length-dependence, however, it is not surprising that this class of short protein extensions are generally poorly detected.

Moreover, the low abundance of these isoforms, as a class, might explain their especially poor detection by mass spectrometry. The lack of PhyloCSF signal for this class of coding regions may also suggest species-specific translation or unusual constraints on the amino acid sequence. For example, the extended portion of the alanyl tRNA synthetase Ala1 did not show evidence of conserved coding potential despite its critical role in mitochondrial translation. This extension was also not detected by mass spectrometry analysis, highlighting the challenges in using existing global approaches to comprehensively identify this class of alternative protein isoforms.

The large class of non-AUG-initiated 5′ extended ORFs defined in this study reveals trends that could not be determined from the few such cases previously confirmed *in vivo*. Our study also highlights the challenges of studying near-cognate-initiated extended protein isoforms by classical approaches and the reasons that few have been confirmed to date. First, as noted above, the protein levels for extended proteins appear low relative to the canonical isoform, making it difficult to study their localization or activity compared with the canonical form, or even to detect their presence in many cases. The efficiency of initiation at near-cognate codons has been reported at between 1% and 10% that of AUG initiation (Chen et al., 2008; Kearse and Wilusz, 2017; Kolitz et al., 2009), and a model in which many fewer ribosomes initiate at the near-cognate TIS relative to the canonical AUG is consistent with our data. Second, the length of the extension relative to the rest of the protein is small (with a median of 21 amino acids in our set), making it difficult to resolve the two isoforms by western blotting. Of the extensions validated by western blot here, only Ymr31 had a large enough size difference to discriminate the two isoforms, while all others necessitated mutating the canonical start codon (*M1A* constructs) to confirm production of the extended isoform. However, we also found that isolated production of the extended isoforms from the *M1A* construct can result in low mRNA levels due to NMD, presumably caused by downstream initiation at an out-of-frame AUG (Celik et al., 2017). The degree to which such transcripts are targets of NMD varied greatly, and these differences did not seem to correlate with the distance from the newly used out-of-frame stop codon to the end of the transcript, a distance proposed to affect NMD (Hug et al., 2016). Interestingly, however, a moderate positive association was seen with the distance from the beginning of the transcript to the downstream out-of-frame AUG. It is currently unclear how or if this observation might inform the mechanism of NMD for these transcripts, but it is intriguing in light of our incomplete understanding of what defines an NMD target in budding yeast.

Are near-cognate-initiated alternate protein isoforms translated from the same transcripts as canonical isoforms or from distinct transcript isoforms? Our TIS-profiling data cannot distinguish between these possibilities, but we favor the former model for several reasons. First, as discussed above, ribosomes frequently bypassing the near-cognate TIS in favor of initiating at the canonical AUG TIS would make translation of the two isoform types in concert possible from one transcript. Second, 5′RACE analysis of two genes with near-cognate-initiated extensions showed the vast majority (33/34) of transcription start sites to be upstream of the extension's TIS (Figures S9A and S9B). Finally, the data for genes in which the canonical AUG start is mutated (*M1A*, Figures 4C, 5E, 5F, 6B, and 6E) support both isoforms being translated from the same pool of transcripts. Otherwise, we would not expect ATG mutation to result in dramatic downregulation of extended isoform production and deletion of *UPF1* (and the resultant NMD deficiency) to rescue it. Finally, in the case of previously studied extensions encoded by *ALA1* and *HFA1*, the transcription start sites identified by 5′RACE were all upstream of the near-cognate TIS (Suomi et al., 2014; Tang et al., 2004).

Although we identified 149 genes for which translation initiation from a 5′ leader-positioned near-cognate codon produces an alternate extended isoform of a characterized protein, this represents only ~3% of possible in-frame TISs upstream of annotated ORFs. It is unclear which *cis* factors contribute to this strong specificity, although a bias for the usage of some near-cognate codons over others appears to be a factor. The preferential usage of these codons, including prominently CUG and UUG, is consistent with previous studies of near-cognate translation initiation (Chen et al., 2008; Diaz de Arce et al., 2018; Kolitz et al., 2009). The basis for the additional specificity beyond near-cognate codon identity cannot be explained by optimal context cues used to define the set of AUG start codons used for translation of traditional ORFs. Our attempts to identify simple shared context motifs around the near-cognate codons used to translate alternate isoforms did not reveal signal beyond the preference for a central U in the start codon itself (data not shown). Identifying the context cues that underlie the strong specificity that we observe is an interesting future area of study that may illuminate differences in the mechanism of translation initiation at AUG and near-cognate codons. It is possible that the case of *HFA1* is informative in this respect, as it is one of only two genes encoding extended isoforms for which we do not see translation initiation at the annotated downstream AUG. This is suggestive of very efficient initiation at the upstream near-cognate codon that prevents leaky downstream scanning of initiation complexes. The sequence downstream of the near-cognate (AUU) start codon for *HFA1* has very high nucleotide-level conservation in yeast, with many positions intolerant to even synonymous mutations (Figure S9C). Such constraint typically indicates function beyond protein coding, such as RNA structure. Consistently, a conserved, stable RNA structure is predicted downstream of the AUU by RNAz analysis, (Figure S9C), which may contribute to the high initiation efficiency at this site (Kozak, 1990).

We found that eIF5A is a *trans* factor that contributes to translation of near-cognate-initiated protein isoforms in meiotic cells. eIF5A is known to associate with 60S ribosomal subunits and has been reported to affect multiple aspects of translation (Gregio et al., 2009; Melnikov et al., 2016; Schuller et al., 2017). We found low eIF5A association with ribosomal subunits in meiosis, leading us to investigate its role in meiotic cells. Inducing higher levels of eIF5A decreased translation of a reporter for near-cognate-initiated translation, and re-analysis of published data for eIF5A depletion in mitotic cells showed higher translation within 5′ leaders generally (consistent with Manjunath et al., 2019; Schuller et al., 2017). Strikingly, the subset of genes that we identified as having near-cognate-initiated translation in 5′ leaders during meiosis were *the same genes* that were responsible for the higher 5′ leader ribosome occupancy in eIF5A-depleted cells, suggesting that the specific near-cognate TISs that we report here are coordinately and selectively “unmasked” by low eIF5A levels. A possible mechanism for this enhanced near-cognate initiation is elongation stalling at specific motifs in eIF5A-deficient cells, leading to ribosome queuing and increased opportunity to initiate at upstream near-cognate sites (Gutierrez et al., 2013; Ivanov et al., 2018; Manjunath et al., 2019; Schuller et al., 2017). The recent finding that low eIF5A enhances CUG-initiated MYC translation in mammals, as well, suggests a conserved mechanism in the regulation of near-cognate-initiated protein isoforms (Manjunath et al., 2019).

An especially intriguing outstanding question raised by this study is the potential function of the many new protein extensions that were identified. Their generally low conservation suggests that they could expand the function of conserved proteins in a species-specific manner. All six known cases of near-cognate-initiated alternate protein isoforms result in mitochondrial targeting of the extended protein and dual mitochondrial and cytoplasmic targeting has been suggested as a general role for this type of alternate isoform (Pujol et al., 2007; Yogev and Pines, 2011). However, mitochondrial localization signals are not significantly enriched in the full set of such extensions that we identify (Figure S9D), leaving investigation of their function (or range of functions) an important area of future study. It remains unclear whether most extensions mediate key cellular roles, akin to the case for Ala1, or whether they might represent noisy expression that provides a selective advantage to cells only under specific new or stressful conditions. Because one-third of random DNA sequences can mediate organellar protein localization, modified protein localization is an attractive hypothesis for the function of these extended isoforms that could drive the evolution of new roles for existing protein products (Kaiser and Botstein, 1990). That these alternative protein isoforms can be induced in concert, potentially by a decrease in the stringency of start codon selection during translation initiation, points to a simple strategy for cells to modulate the features of a discrete subset of the proteome in response to a change in condition.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Yeast Strain Construction
  - ○ Yeast Growth and Sporulation
- METHOD DETAILS
  - ○ TIS-Profiling
  - ○ Polysome Gradient Analysis
  - ○ Mass-Spectrometry-Based Protein Identification of the 40S/60S Peaks by iTRAQ-Labeling
  - ○ Western Blotting
  - ○ qPCR
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Analysis of TIS-Profiling Data
  - ○ Footprint Quantification and Correlation Analysis
  - ○ Start Codon Analysis
  - ○ Context Analysis
  - ○ Conservation Analysis
  - ○ Deep Proteome Identification of Peptides and Proteins

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.06.011.

ACKNOWLEDGMENTS

We thank Elçin Ünal, George Otto, Helen Vander Wende, and Tina Sing for critical reading of this manuscript and members of the Brar and Ünal labs for helpful suggestions. G.A.B., A.R.E., and A.L.H. are supported by the National Institutes of Health (DP2-GM-119138 and 1R35GM134886), investigator awards from the Alfred P. Sloan Foundation (FG-2016-6229) and Pew Charitable Trusts (00029624), and UC-Berkeley start-up funding, including from the Bowes Foundation, to G.A.B. A.L.H. was also funded in part by NSF predoctoral fellowship DGE 1752814 and A.R.E. and A.L.H. were funded in part by NIH training grants to the MCB department (T32 GM 0007232 and T32 HG 00047). M.J. is funded by the NIH (R35-GM-128802) and Columbia start-up funding. I.J. is supported by the National Human Genome Research Institute of the National Institutes of Health (U41HG007234 and HG004037).

AUTHOR CONTRIBUTIONS

A.R.E. and G.A.B. conceived most aspects of this study. Ribosome profiling experiments and analysis were performed by A.R.E., I.H., P.D.D., and G.A.B. ORF-RATER analysis was performed by A.P.F. and A.L.H. Conservation analysis was performed by I.J. and M.K. Mass spectrometry analysis was performed by M.J. All other experiments were performed by A.R.E. Data analyses were performed by A.R.E., A.L.H., and G.A.B. The manuscript was written and edited by A.R.E., A.L.H., and G.A.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 10, 2020
Revised: May 18, 2020
Accepted: June 24, 2020
Published: July 24, 2020

REFERENCES

Aitken, C.E., and Lorsch, J.R. (2012). A mechanistic overview of translation initiation in eukaryotes. Nat. Struct. Mol. Biol. *19*, 568–576.

Almagro Armenteros, J.J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., and Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance *2*, e201900429.

Benne, R., and Hershey, J.W. (1978). The mechanism of action of protein synthesis initiation factors from rabbit reticulocytes. J. Biol. Chem. *253*, 3078–3087.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science *335*, 552–557.

Brent, M.R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. *15*, 1777–1786.

Celik, A., Baker, R., He, F., and Jacobson, A. (2017). High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA *23*, 735–748.

Chang, C.-P., Chen, S.-J., Lin, C.-H., Wang, T.-L., and Wang, C.-C. (2010). A single sequence context cannot satisfy all non-AUG initiator codons in yeast. BMC Microbiol. *10*, 188.

Chang, K.-J., and Wang, C.-C. (2004). Translation initiation from a naturally occurring non-AUG codon in Saccharomyces cerevisiae. J. Biol. Chem. *279*, 13778–13785.

Chen, S.-J., Lin, G., Chang, K.-J., Yeh, L.-S., and Wang, C.-C. (2008). Translational efficiency of a non-AUG initiation codon is significantly affected by its sequence context in yeast. J. Biol. Chem. *283*, 3173–3180.

Cheng, Z., Otto, G.M., Powers, E.N., Keskin, A., Mertins, P., Carr, S.A., Jovanovic, M., and Brar, G.A. (2018). Pervasive, coordinated protein-level changes driven by transcript isoform switching during meiosis. Cell *172*, 910–923.e16.

Clements, J.M., Laz, T.M., and Sherman, F. (1988). Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. Mol. Cell. Biol. *8*, 4533–4536.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

Davis, C.A., Grate, L., Spingola, M., and Ares, M. (2000). Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. *28*, 1700–1706.

Diaz de Arce, A.J., Noderer, W.L., and Wang, C.L. (2018). Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. Nucleic Acids Res. *46*, 985–994.

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. Mol. Cell *60*, 816–827.

Fresno, M., Jiménez, A., and Vázquez, D. (1977). Inhibition of translation in eukaryotic systems by harringtonine. Eur. J. Biochem. *72*, 323–330.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J., et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. *22*, 2208–2218.

Fukasawa, Y., Tsuji, J., Fu, S.-C., Tomii, K., Horton, P., and Imai, K. (2015). MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol. Cell. Proteomics *14*, 1113–1126.

Gregio, A.P.B., Cano, V.P.S., Avaca, J.S., Valentini, S.R., and Zanelli, C.F. (2009). eIF5A has a function in the elongation step of translation in yeast. Biochem. Biophys. Res. Commun. *380*, 785–790.

Gutierrez, E., Shin, B.-S., Woolstenhulme, C.J., Kim, J.-R., Saini, P., Buskirk, A.R., and Dever, T.E. (2013). eIF5A promotes translation of polyproline motifs. Mol. Cell *51*, 35–45.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47*, D853–D858.

Harris, R.S. (2007). Improved pairwise alignment of genomic DNA., PhD Thesis (The Pennsylvania State University).

Henderson, A., and Hershey, J.W. (2011). Eukaryotic translation initiation factor (eIF) 5A stimulates protein synthesis in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *108*, 6415–6419.

Hershey, J.W., Smit-McBride, Z., and Schnier, J. (1990). The role of mammalian initiation factor eIF-4D and its hypusine modification in translation. Biochim. Biophys. Acta *1050*, 160–162.

Heublein, M., Ndi, M., Vazquez-Calvo, C., Vögtle, F.N., and Ott, M. (2019). Alternative translation initiation at a UUG codon gives rise to two functional variants of the mitochondrial protein Kgd4. J. Mol. Biol. *431*, 1460–1467.

Hinnebusch, A.G. (2011). Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol. Mol. Biol. Rev. *75*, 434–467.

Homann, O.R., and Johnson, A.D. (2010). MochiView: versatile software for genome browsing and DNA motif analysis. BMC Biol. *8*, 49.

Hood, H.M., Neafsey, D.E., Galagan, J., and Sachs, M.S. (2009). Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. Annu. Rev. Microbiol. *63*, 385–409.

Hossain, M.A., Rodriguez, C.M., and Johnson, T.L. (2011). Key features of the two-intron Saccharomyces cerevisiae gene SUS1 contribute to its alternative splicing. Nucleic Acids Res. *39*, 8612–8627.

Hug, N., Longman, D., and Cáceres, J.F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. *44*, 1483–1495.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Ivanov, I.P., Shin, B.-S., Loughran, G., Tzani, I., Young-Baird, S.K., Cao, C., Atkins, J.F., and Dever, T.E. (2018). Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. Mol. Cell *70*, 254–264.e6.

Juneau, K., Nislow, C., and Davis, R.W. (2009). Alternative splicing of PTC7 in *Saccharomyces cerevisiae* determines protein localization. Genetics *183*, 185–194.

Kaiser, C.A., and Botstein, D. (1990). Efficiency and diversity of protein localization by random signal sequences. Mol. Cell. Biol. *10*, 3163–3173.

Kearse, M.G., and Wilusz, J.E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. *31*, 1717–1731.

Kemper, W.M., Berry, K.W., and Merrick, W.C. (1976). Purification and properties of rabbit reticulocyte protein synthesis initiation factors M2Balpha and M2Bbeta. J. Biol. Chem. *251*, 5551–5557.

Kim Guisbert, K.S., Zhang, Y., Flatow, J., Hurtado, S., Staley, J.P., Lin, S., and Sontheimer, E.J. (2012). Meiosis-induced alterations in transcript architecture and noncoding RNA expression in S. cerevisiae. RNA *18*, 1142–1153.

Kolitz, S.E., Takacs, J.E., and Lorsch, J.R. (2009). Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. RNA *15*, 138–152.

Kozak, M. (1978). How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell *15*, 1109–1123.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. *12*, 857–872.

Kozak, M. (1990). Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc. Natl. Acad. Sci. USA *87*, 8301–8305.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. Gene *234*, 187–208.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. Gene *299*, 1–34.

Kritsiligkou, P., Chatzi, A., Charalampous, G., Mironov, A., Grant, C.M., and Tokatlidis, K. (2017). Unconventional targeting of a thiol peroxidase to the mitochondrial intermembrane space facilitates oxidative protein folding. Cell Rep. *18*, 2729–2741.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. USA *109*, E2424–E2432.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282.

Longtine, M.S., McKenzie, A., Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., and Pringle, J.R. (1998). Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast *14*, 953–961.

Lopo, A.C., Lashbrook, C.C., Infante, D., Infante, A.A., and Hershey, J.W. (1986). Translational initiation factors from sea urchin eggs and embryos: functional properties are highly conserved. Arch. Biochem. Biophys. *250*, 162–170.

Machkovech, H.M., Bloom, J.D., and Subramaniam, A.R. (2019). Comprehensive profiling of translation initiation in influenza virus infected cells. PLoS Pathog. *15*, e1007518.

Manjunath, H., Zhang, H., Rehfeld, F., Han, J., Chang, T.-C., and Mendell, J.T. (2019). Suppression of ribosomal pausing by eIF5A is necessary to maintain the fidelity of start codon selection. Cell Rep. *29*, 3134–3146.e6.

Martin, N.C., and Hopper, A.K. (1994). How single genes provide tRNA processing enzymes to mitochondria, nuclei and the cytosol. Biochimie *76*, 1161–1167.

Melnikov, S., Mailliot, J., Shin, B.-S., Rigger, L., Yusupova, G., Micura, R., Dever, T.E., and Yusupov, M. (2016). Crystal structure of hypusine-containing translation factor eIF5A bound to a rotated eukaryotic ribosome. J. Mol. Biol. *428*, 3570–3576.

Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D.R., Burgess, M.W., Gillette, M.A., Jaffe, J.D., and Carr, S.A. (2013). Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat. Methods *10*, 634–637.

Monteuuis, G., Miścicka, A., Świrski, M., Zenad, L., Niemitalo, O., Wrobel, L., Alam, J., Chacinska, A., Kastaniotis, A.J., and Kufel, J. (2019). Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. Nucleic Acids Res. *47*, 5777–5791.

Morris, D.R., and Geballe, A.P. (2000). Upstream open reading frames as regulators of mRNA translation. Mol. Cell. Biol. *20*, 8635–8642.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Pujol, C., Maréchal-Drouard, L., and Duchêne, A.-M. (2007). How can organellar protein N-terminal sequences be dual targeting signals? In silico analysis and mutagenesis approach. J. Mol. Biol. *369*, 356–367.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. *2*, 1896–1906.

Saini, P., Eyler, D.E., Green, R., and Dever, T.E. (2009). Hypusine-containing protein eIF5A promotes translation elongation. Nature *459*, 118–121.

Sapkota, D., Lake, A.M., Yang, W., Yang, C., Wesseling, H., Guise, A., Uncu, C., Dalal, J.S., Kraft, A.W., Lee, J.-M., et al. (2019). Cell-type-specific profiling of alternative translation identifies regulated protein isoform variation in the mouse brain. Cell Rep. *26*, 594–607.e7.

Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., and Liu, J.O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat. Chem. Biol. *6*, 209–217.

Schreier, M.H., Erni, B., and Staehelin, T. (1977). Initiation of mammalian protein synthesis. I. Purification and characterization of seven initiation factors. J. Mol. Biol. *116*, 727–753.

Schuller, A.P., Wu, C.C.-C., Dever, T.E., Buskirk, A.R., and Green, R. (2017). eIF5A functions globally in translation elongation and termination. Mol. Cell *66*, 194–205.e5.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., Hengel, H., et al. (2012). Decoding human cytomegalovirus. Science *338*, 1088–1093.

Sugawara, K., Nishiyama, Y., Toda, S., Komiyama, N., Hatori, M., Moriyama, T., Sawada, Y., Kamei, H., Konishi, M., and Oki, T. (1992). Lactimidomycin, a new glutarimide group antibiotic. Production, isolation, structure and biological activity. J. Antibiot. *45*, 1433–1441.

Suomi, F., Menger, K.E., Monteuuis, G., Naumann, U., Kursu, V.A., Shvetsova, A., and Kastaniotis, A.J. (2014). Expression and evolution of the non-canonically translated yeast mitochondrial acetyl-CoA carboxylase Hfa1p. PLoS One *9*, e114738.

Suzuki, Y., St Onge, R.P., Mani, R., King, O.D., Heilbut, A., Labunskyy, V.M., Chen, W., Pham, L., Zhang, L.V., Tong, A.H.Y., et al. (2011). Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. Nat. Methods *8*, 159–164.

Tang, H.-L., Yeh, L.-S., Chen, N.-K., Ripmaster, T., Schimmel, P., and Wang, C.-C. (2004). Translation of a yeast mitochondrial tRNA synthetase initiated at redundant non-AUG codons. J. Biol. Chem. *279*, 49656–49663.

Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A.C., and Vagner, S. (2003). Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol. Cell *95*, 169–178.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., et al. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. USA *106*, 3264–3269.

Yogev, O., and Pines, O. (2011). Dual targeting of mitochondrial proteins: mechanism, regulation and function. Biochim. Biophys. Acta *1808*, 1012–1020.

Zalatan, J.G., Coyle, S.M., Rajan, S., Sidhu, S.S., and Lim, W.A. (2012). Conformational control of the Ste5 scaffold protein insulates against MAP kinase misactivation. Science *337*, 1218–1222.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-GFP	Clontech (Mountain View, CA, United States of America)	Cat#632381, RRID:AB_2313808
Rabbit anti-hexokinase antibody	Rockland	Cat#100-4159, RRID:AB_21991
Anti-mouse 800 secondary	LI-COR	Cat#926-32210, RRID:AB_621842
Anti-rabbit 680 secondary	LI-COR	Cat#926-68071, RRID:AB_10956166
Rat anti-tubulin	Serotec (Oxford)	Cat#MCA78G, RRID:AB_325005
Anti-rat 680 secondary	LI-COR	Cat#926-68076, RRID:AB_10956590
Chemicals, Peptides, and Recombinant Proteins		
Lactimidomycin (LTM)	Millipore	Cat#506291
RNaseI	Ambion	Cat#AM2294
Sequencing grade modified trypsin	Promega (Southampton, UK)	Cat#V5111
PMSF	Sigma	Cat#78830
cOmplete mini EDTA-free protease inhibitor cocktail	Roche	Cat#29384100
Acid-washed glass beads	Sigma	Cat#G8772
PBS Odyssey Blocking Buffer	LI-COR	Cat#927-40100
Superscript III	Thermo (Waltham, MA, United States of America)	Cat#18080044
SYBR green mix	Thermo	Cat#4309155
Critical Commercial Assays		
Q5 Site Directed Mutagenesis kit	NEB	Cat#E0554S
TURBO DNA-free kit	Ambion	Cat#AM1907
Deposited Data		
Raw and analyzed sequencing data	This study	GEO: GSE150375
Experimental Models: Organisms/Strains		
All yeast strains are described in Table S5.	Brar-Ünal Lab and this study	N/A
Oligonucleotides		
oGAB-2736 GFP-qPCR_f ctccggtgaaggtgaaggtg	IDT	N/A
oGAB-2737 GFP-qPCR_r aggttggccatggaactgg	IDT	N/A
oGAB-3301 PFY1-qPCR_f acggtagacatgatgctgagg	IDT	N/A
oGAB-3302 PFY1-qPCR_r acggttggtggataatgagc	IDT	N/A
oGAB-7864 HYP2-qPCR_f TGTCAAGGCTCCAGAAGGTGA	IDT	N/A
oGAB-7865 HYP2-qPCR_r CCCATAGCGGAGATGATGGT	IDT	N/A
Recombinant DNA		
pÜB1/pFA6A-KanMX	Addgene	Cat#39296
pÜB189/pFA6A-KanMX-pCUP1 (derived from pFA6A)	Brar-Ünal Lab	N/A
pÜB731 (derived from pNH604)	Brar-Ünal Lab	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
MochiView	Homann and Johnson, 2010	http://www.johnsonlab.ucsf.edu/mochi/
ImageStudio Lite Software	LI-COR	https://www.licor.com/bio/products/software/image_studio_lite/index.html

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gloria Brar (gabrar@berkeley.edu).

Materials Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gloria Brar (gabrar@berkeley.edu).

Data and Code Availability

The accession number for the sequencing data reported in this paper is GEO: GSE150375.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Yeast Strain Construction

All yeast strains used were *Saccharomyces cerevisiae* of the SK1 background. Strains used in this study are listed in Table S5. GFP-tagged strains were created using single-integration plasmids constructed by Gibson assembly of PCR-amplified genomic regions including 5' leader regions and PCR-amplified single-integration vector pÜB731/pNH604 (which contains a TRP1 selection marker, yEGFP tag and *ADH1* terminator; described in Zalatan et al., 2012). Plasmids were mutated using the Q5 Site Directed Mutagenesis kit. *M1A* constructs were generated by mutating the annotated ATG to a GCT, and for genes where the next downstream ATG was in-frame, this ATG was also mutated to a GCT. *Ustop* constructs were generated by mutating the codon prior to the annotated ATG to a stop codon. Deletion strains were created using pÜB1/pFA6A-KanMX (described in Longtine et al., 1998), and overexpression strains were created using pÜB189/pFA6A-*KanMX-pCUP1*.

Yeast Growth and Sporulation

Vegetative cells were grown in YEPD at 30°C, with exponentially growing cells grown from an OD600 of 0.2 to an OD600 of 1, and saturated cells to an OD600 >10. For meiotic time courses, strains were inoculated in YEPD for 24 hours, then diluted to an OD600 of 0.2 in buffered YTA and grown for 16 hours at 30°C. Cells were washed once in water, resuspended in sporulation media (SPO), and grown at 30°C. Time points were taken at times indicated in figures.

METHOD DETAILS

TIS-Profiling

Cells were treated with 3 μM LTM (Millipore) for 20 minutes, then harvested by filtration and flash freezing in liquid nitrogen. Samples were lysed by mixermilling at 15 Hz for 6 rounds of 3 minutes each. Samples were thawed at 30°C and spun down at 3000 rcf for 5 minutes at 4°C. The supernatant was removed and cleared at 20,000 rcf for 10 minutes at 4°C, and 200 μL aliquots of cleared supernatant were flash frozen. Ribosome profiling library preparation was as in (Brar et al., 2012). In brief, samples were treated with RNaseI (Ambion), then monosome peaks were collected from sucrose gradients. RNA was extracted, size selected, dephosphorylated, polyA-tailed, subjected to rRNA subtraction, RT-PCR, circularization, and PCR amplification. Samples were sequenced on an Illumina HiSeq 2500, 50SRR, with multiplexing, at the UC-Berkeley Vincent Coates QB3 Sequencing facility.

Polysome Gradient Analysis

Extract from mixermilling flash-frozen cells was subjected to polysome gradient analysis as described in (Ingolia et al., 2009). In short, 200 μl extract was loaded on 10-50% sucrose gradients with or without RNAseI treatment, depending on if sample would be used for ribosome profiling or 40S/60S isolation, respectively. Samples were centrifuged in a Beckman XL-70 Ultracentrifuge, using a Sw-Ti41 rotor for 3 hours at 35,000 rpm at 4°C. Tube was loaded on a Bio-Comp Gradient Station and analyzed for absorbance at 260 nm. For

Please cite this article in press as: Eisenberg et al., Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.011

mass spectrometry of 40S/60S fraction, sucrose fraction was collected and flash frozen prior to precipitation and mass spectrometry.

Mass-Spectrometry-Based Protein Identification of the 40S/60S Peaks by iTRAQ-Labeling

Proteins from the collected 40S/60S fractions were precipitated by adding -20°C cold acetone to the lysate (acetone to eluate ratio 10:1) and overnight incubation at -20°C. The proteins were pelleted by centrifugation at 20,000 g for 15 min at 4°C. The supernatant was discarded and the pellet was left to dry by evaporation. The protein pellet was reconstituted in 100 µl urea buffer (8 M Urea, 75 mM NaCl, 50 mM Tris/HCl pH 8.0, 1 mM EDTA) and protein concentrations were determined by BCA assay (Pierce). 10 µg of total protein per sample (with the exception of the ''Master spike-in Total Extract'' where we used 20 µg – see below) were processed further. Disulfide bonds were reduced with 5 mM dithiothreitol and cysteines were subsequently alkylated with 10 mM iodoacetamide. Samples were diluted 1:4 with 50 mM Tris/HCl (pH 8.0) and sequencing grade modified trypsin (Promega) was added in an enzyme-to-substrate ratio of 1:50. After 16 hours of digestion, samples were acidified with 1% formic acid (final concentration). Tryptic peptides were desalted on C18 StageTips according to (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator. Desalted peptides were labeled with the iTRAQ reagent according to the manufacturer's instructions (AB Sciex) and as previously described (Mertins et al., 2013). Briefly, replicate 1 and replicate 2 were each measured in their own iTRAQ mix. In addition, each mix had the same two ''Master spike-in'' samples added. The ''Master spike-in Total Lysate'' contained an equal mix of total protein extract from vegetative, meiotic cells and spores. The ''Master spike-in Polysomes'' contained an equal mix of proteins from all polysome fractions from vegetative, meiotic cells and spores. Briefly, 0.33 units of iTRAQ reagent were used per IP. Peptides were dissolved in 10 µl of 0.5 M TEAB pH 8.5 solution and the iTRAQ reagent was added in 23 µl of ethanol. After 1 h incubation the reaction was stopped with 50 mM Tris/HCl (pH 8.0). Differentially labeled peptides were mixed and subsequently desalted on C18 StageTips (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator. Peptides were reconstituted in 50 µl 3% MeCN/0.1% formic acid. LC-MS/MS analysis was performed as previously described (Mertins et al., 2013).

All mass spectra were analyzed with the Spectrum Mill software package v4.0 beta (Agilent Technologies, Santa Clara, CA, United States of America) according to (Mertins et al., 2013) using the yeast UniProt database (UniProt.Yeast.completeIsoforms.UP000002311; strain ATCC 204508 / S288c). For identification, we applied a maximum FDR of 1% separately on the protein and peptide level and proteins were grouped in subgroup specific manner. We calculated intensity ratios relative to iTRAQ channel 117 (''Master spike-in Polysomes'') and subsequently median normalized these ratios for each sample.

Mix 1		
Sample	*iTRAQ label*	*Peptides labeled (µg)*
Master spike-in Total Lysate	114	20
40S/60S Meiosis Repl. 01	115	10
40S/60S Vegetative Repl. 01	116	10
Master spike-in Polysomes	117	10
Mix 2		
Sample	*iTRAQ label*	*Peptides labeled (µg)*
Master spike-in Total Lysate	114	20
40S/60S Vegetative Repl. 02	115	10
40S/60S Meiosis Repl. 02	116	10
Master spike-in Polysomes	117	10

Western Blotting

Strains were grown in YEPD or SPO, with 3.5 ODs of cells harvested at indicated time points. Cells were fixed in 5% TCA for at least 10 minutes, then washed once with acetone and dried overnight. Samples were resuspended in 50 mM Tris-HCl, 1 mM EDTA, 3 mM DTT, 1.1 mM PMSF (Sigma) and 1x cOmplete mini EDTA-free protease inhibitor cocktail (Roche), then lysed with glass-bead-based agitation for 5 minutes, then boiled in SDS loading buffer for 5 minutes at 95°C. Samples were spun down for 5 min at 20,000 rcf prior to running on a 4-12% Bis-Tris gel at 175 V for 30 minutes. Transfer to nitrocellulose membrane was performed using a Turbo Transfer semi-dry standard 30 minute transfer. Membrane was blocked with 5% milk in PBST for 1 hour, and incubated in primary antibody overnight at 4°C. Primary antibodies were diluted 1:2,000 for mouse anti-GFP (Clontech), 1:10,000 for rabbit anti-hexokinase (Rockland), and 1:10,000 for rat anti-tubulin (Serotec) in PBS blocking buffer (LI-COR). Membrane was washed with PBST 5 times for 5 minutes each time, then incubated in secondary antibody (1:15,000 anti-mouse 800, anti-rabbit 680, or anti-rat 680 (LI-COR) in PBS blocking buffer) for 2 hours, then washed with PBST 5 times for 5 minutes each time. Images were acquired using the LI-COR Odyssey imager, and analysis and quantification was performed in ImageStudio Lite Software (LI-COR).

qPCR

Samples were flash frozen in liquid nitrogen, then resuspended in TES buffer (10 mM Tris 7.5, 10 mM EDTA, 0.5% SDS), with acid-washed glass beads (Sigma) and acid phenol:chloroform:isoamyl alcohol (125:24:1; pH 4.7). Samples were centrifuged for 10 minutes at 21000 rcf at 4°C, then the aqueous phase was removed and added to chloroform. Samples were centrifuged again for 5 minutes at 21,000 rcf, then the aqueous phase was removed and added to isopropanol and 0.33 M NaOAc. Samples were precipitated at 4°C overnight, then centrifuged for 20 minutes at 21,000 rcf at 4°C. Pellets were washed with 80% ethanol, air-dried, and resuspended in water. The TURBO DNA-free kit (Thermo) was used to treat 2.5 μg RNA with DNAse, then samples were incubated with random hexamers for 5 min at 65°C. Superscript III buffer (Thermo), DTT, and dNTPs were added, then Superscript III was added and samples were incubated at 25°C for 10 minutes, 42°C for 50 min, and 70°C for 10 minutes. cDNA was quantified by 7500 FAST Real-Time PCR machine with SYBR green mix (Thermo) and the following qPCR primers listed in the Key Resources Table: *GFP* (oGAB-2736/oGAB-2737), PFY1 (oGAB-3301/oGAB-3302), and *HYP2* (oGAB-7864/oGAB-7865).

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of TIS-Profiling Data

Sequencing data were aligned using bowtie2 (Langmead and Salzberg, 2012), and ORF-RATER was applied to TIS-profiling data and standard profiling data. Genome browser analysis and visualization was done using MochiView (Homann and Johnson, 2010). The distribution of read lengths by this approach was approximately 2 nucleotides longer than seen for standard ribosome profiling (peaking at 30 nt, rather than 28 nt), and we found that the a-site offset typically used for standard ribosome profiling data visualization required shifting of 2 nt upstream, as well. To calculate expression values, footprint values from standard ribosome profiling for annotated genes were averaged, and an expression cutoff greater than or equal to 5 RPKM was used for analysis shown in Figures S2A and S2B; Table S2.

Footprint Quantification and Correlation Analysis

Standard RPKM calculations were used for cycloheximide profiling. For TIS-profiling, we counted reads mapping to the region spanning 3 nt up- and downstream of the start codon and normalized by total reads at all initiation sites. The spearman correlation between TIS-profiling and standard profiling was calculated for each gene. The distribution of correlation scores was compared to a null distribution generated by shuffling gene names and performing the same correlation analysis. Statistical significance was determined using a K.S. test. For 5′ UTR quantification, read counts were determined for the region from the canonical start to 99 nt upstream. Counts were normalized by total reads at initiation sites to account for library size differences.

Start Codon Analysis

The region 30-99 nt upstream of canonical starts was used as a proxy for 5'UTRs. The upper cutoff was based on average transcript lengths in yeast and the lower cutoff was matched to the minimum length cutoff used for extensions. Within this region, we counted the number of AUG and near-cognate in-frame start codons that did not also have an in-frame stop codon before the canonical TIS. These counts gave the "expected" distribution of codon usage given no start codon selection bias. The expected counts were compared to the counts that were observed among called extensions. Statistical significance was determined using Fisher's Exact Test for each individual codon. As a control, we also analyzed the regions within 30 ntbp upstream of canonical start codons, which would encode short (<10 amino acid) extensions. This class does not show the same start codon bias as is seen for the longer set (Figures S4B and S4C).

Context Analysis

Maximum motif score analysis was performed using Mochiview (Homann and Johnson, 2010) for the regions 10 nt up- and downstream of all annotated genes, recapitulating the known Kozak sequence. The enrichment for this motif in regions 10 nt up- and downstream of other start codon classes and control regions were plotted using the maximum motif score enrichment tool in Mochiview.

Conservation Analysis

PhyloCSF scores for the extensions were computed using the 7yeast parameter set and the default mle and AsIs options, applied to the extension, starting at the upstream start codon and continuing up to but not including the annotated start codon. Alignments used as input to PhyloCSF and shown in CodAlignView were extracted from the MULTIZ whole genome alignment of seven *Saccharomyces* species based on the sacCer3 *S. cerevisiae* S288C reference assembly, obtained from the UCSC Genome Browser (Haeussler et al., 2019). Extensions were first mapped from the SK1 strain assembly to the the S288C strain sacCer3 assembly using an ad hoc alignment created with LASTZ (Harris, 2007). We did not compute PhyloCSF scores for the two extensions of *YBR012C* because of difficulty mapping to the S288C strain. In some cases, we also computed PhyloCSF scores of 10-codon windows 5' of the detected TIS to determine if the ancestral extension was longer than the one detected.

Deep Proteome Identification of Peptides and Proteins

First, we generated a concatenated search database including all canonical proteins in the yeast UniProt database (release 2014_09, strain ATCC 204508 / S288c), and the newly predicted alternative proteoforms (e.g. N-terminal extensions) and proteins identified by ORF-RATER (a set including scores >0.1). Raw data generated previously to investigate proteome changes during yeast meiosis at deep coverage (Cheng et al., 2018) were analyzed with the MaxQuant software version 1.6.0.16 (Cox and Mann, 2008) against the above mentioned concatenated search database, and MS/MS searches were performed with the following parameters: TMT-11plex labeling on the MS2 level, oxidation of methionine and protein N-terminal acetylation as variable modifications; carbamidomethylation as fixed modification; Trypsin/P as the digestion enzyme; precursor ion mass tolerances of 20 p.p.m. for the first search (used for nonlinear mass re-calibration) and 4.5 p.p.m. for the main search, and a fragment ion mass tolerance of 20 p.p.m. For identification, we applied a maximum FDR of 1% separately on protein and peptide level.

Cell Systems, Volume *11*

Supplemental Information

Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast

Amy R. Eisenberg, Andrea L. Higdon, Ina Hollerer, Alexander P. Fields, Irwin Jungreis, Paige D. Diamond, Manolis Kellis, Marko Jovanovic, and Gloria A. Brar

Figure S1

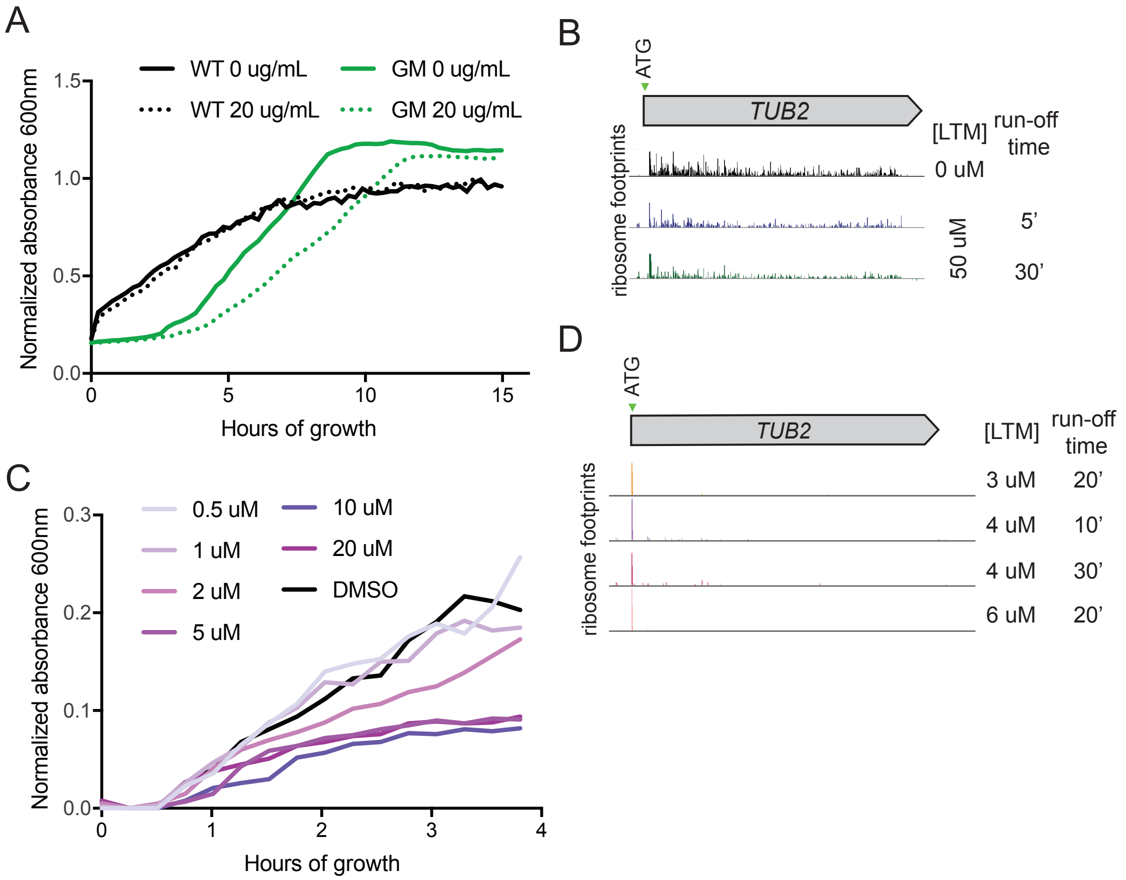


Figure S1: Optimization of TIS-profiling conditions for yeast, Related to Figure 1

(A) Growth curve of WT cells or Green Monster (GM) mutant cells treated with harringtonine. The GM strain lacks 16 ABC transporter drug efflux genes. Solid lines indicate no treatment and dotted lines indicate 20 ug/mL of harringtonine. Absorbance at 600 uM was used to measure growth over 16 hours. Estimated doubling time for WT cells is 3.7 and 3.3 hours for 0 and 20 ug/mL harringtonine respectively, and 1.9 and 2.8 hours for GM cells for 0 and 20 ug/mL harringtonine respectively.

(B) Ribosome profiling reads from cells treated with 0 or 50 μM LTM and either 5 or 30 minutes run-off time for a representative gene, *TUB2*.

(C) Growth curve of WT yeast treated with LTM at concentrations between 0-20 μM. Absorbance at 600 uM was used to measure growth over four hours. Estimtated doubling time for 0 μM LTM was 1.1 hours, and increased to 1.8 hours for 20 μM LTM.

(D) Ribosome profiling reads from cells treated with varying LTM concentration and run off times for a representative gene, *TUB2*.

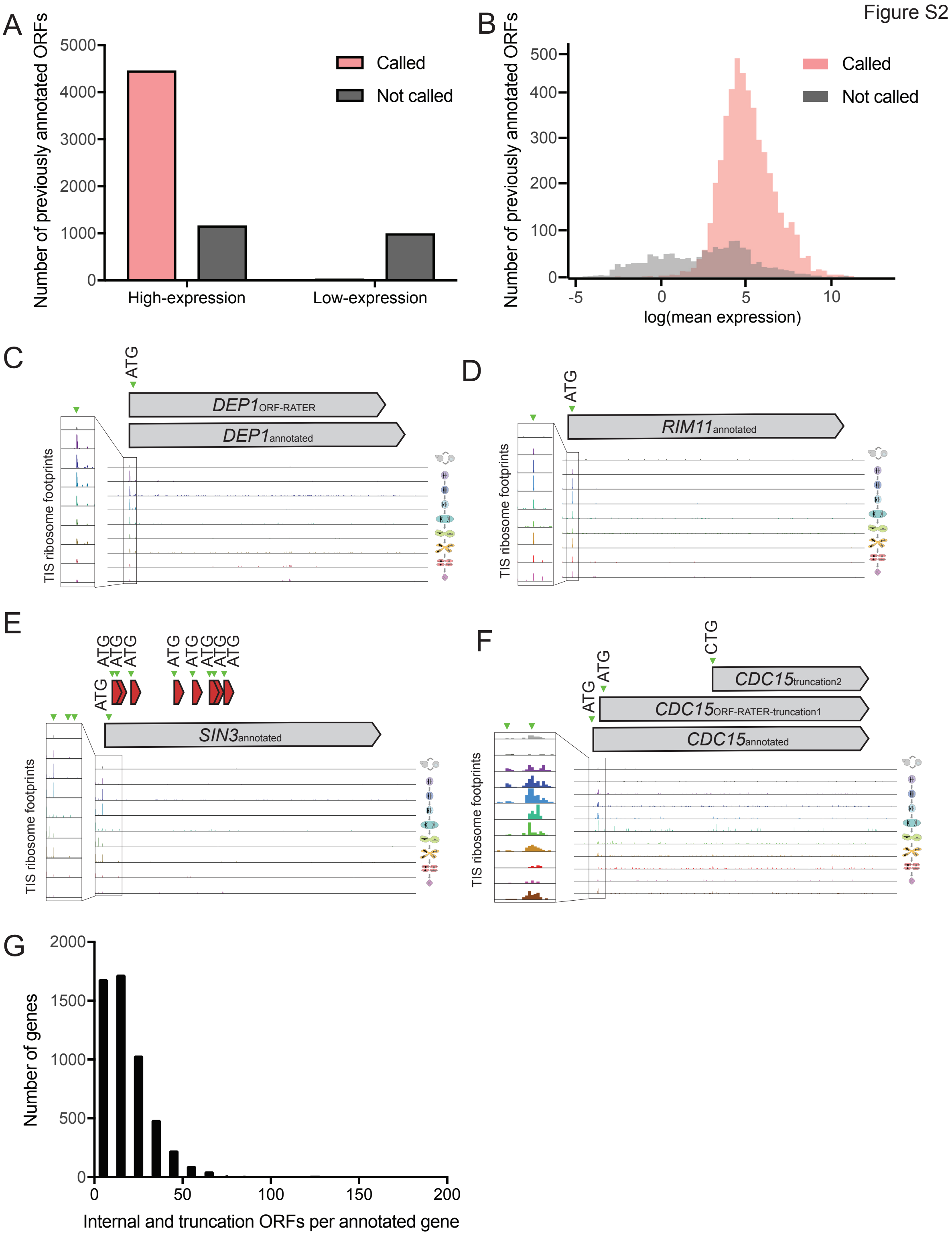


Figure S2: Categories of false positive and false negative ORF-RATER calls, Related to Figure 2

(A) Previously annotated ORFs that are called (pink) or not called (gray), at expression values greater (high-expression) or less than (low-expression) 5 mean RPKM. Approximately half of annotated ORFs that were not called have low expression.

(B) Distribution of expression (mean RPKM of all time points) for annotated ORFs that are called (pink) versus not called (gray).

(C) TIS-profiling for *DEP1,* a gene showing a change in stop codon annotation leading to it not being called as an annotated ORF by ORF-RATER.

(D) TIS-profiling for *RIM11*, a gene that is an example of a false negative, where an apparent peak is present at the annotated ATG but was not identified as a TIS by ORF-RATER.

(E) TIS-profiling for *SIN3*, a gene with many internal ORFs called, most of which are likely false positives.

(F) TIS-profiling for *CDC15*, a gene with two truncated ORFs called, the first of which represents a likely misannotation and the second of which is a likely false positive.

(G) Number of internally initiated ORFs called per annotated gene.

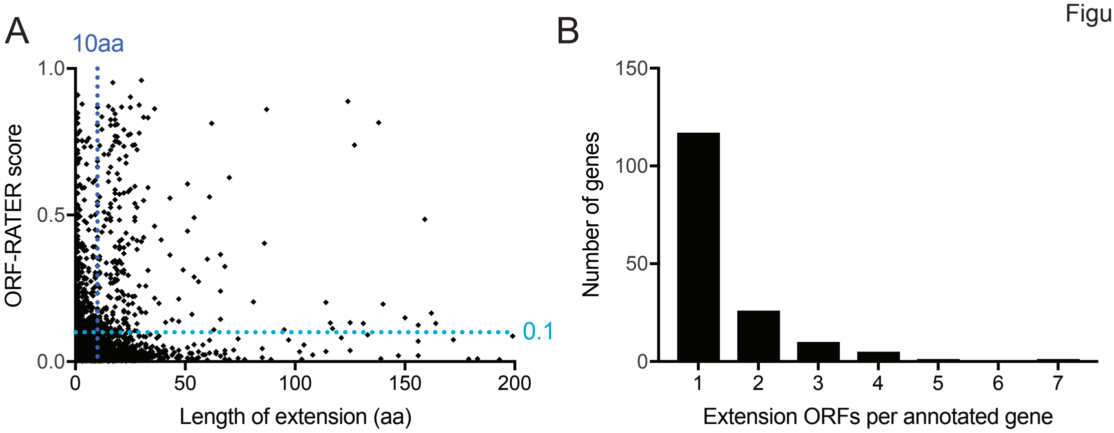


Figure S3: Properties of extension ORFs used for setting cutoffs, Related to Figure 2

(A) Length versus score for all extension ORFs, with a line showing the length cutoff at 10 amino acids and the score cutoff of 0.1.

(B) Number of extension ORFs called per annotated gene.

Figure S4

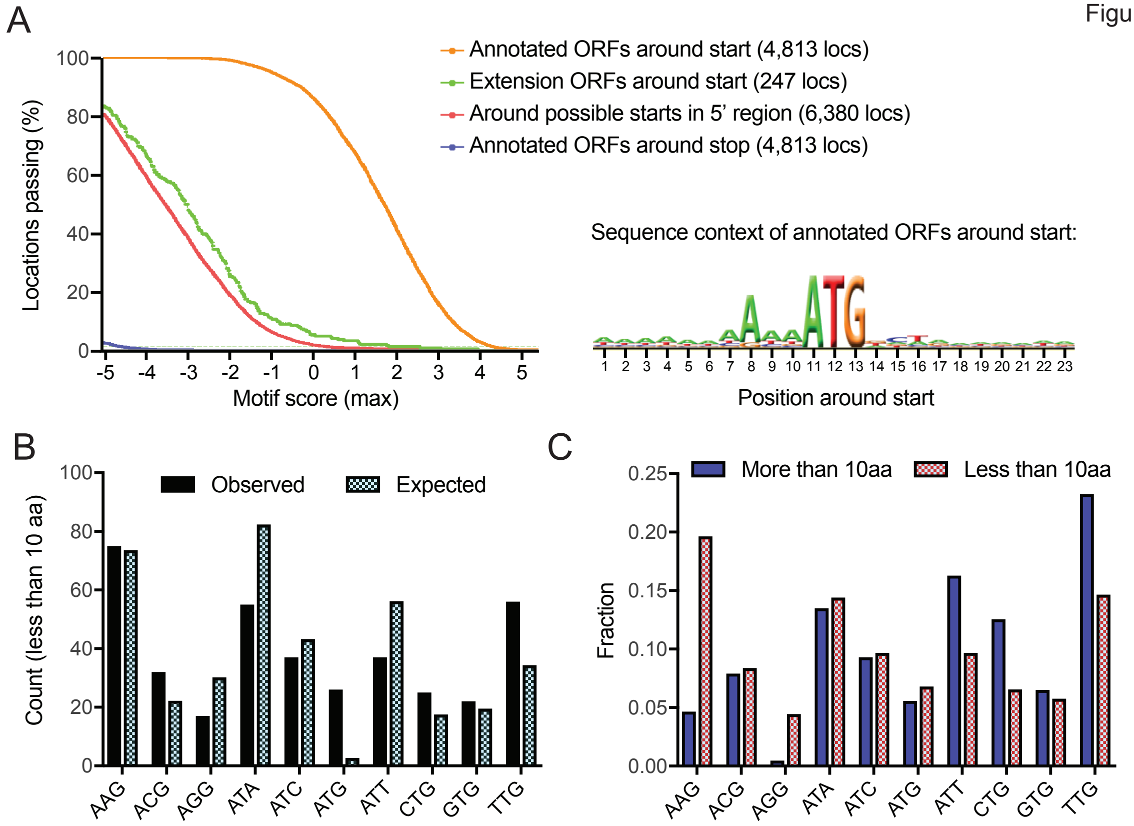


Figure S4: Translated near-cognate-initiated ORFs do not show Kozak sequence context enrichment, Related to Figure 3, Methods

(A) Enrichment plot (left) for yeast Kozak motif in the 10 bp region up and downstream of ORF-RATER called annotated genes (orange), near-cognate extensions (green), all possible in-frame near-cognate start codons (red), and stop codons for annotated genes (blue). Sequence context logo (right) was derived from annotated ORFs.

(B) Comparison of start codon usage for called extensions less than 10aa from canonical start codon (observed) to prevalence within UTR (expected), showing a lack of codon bias relative to what was observed for longer, more likely functional extensions (as seen in Figure 3F).

(C) Comparison of start codon usage between extensions that initiate more than and less than 10 amino acids upstream of the canonical start codon. Longer extensions show a stronger bias toward better start codons and against weaker start codons.

Figure S5

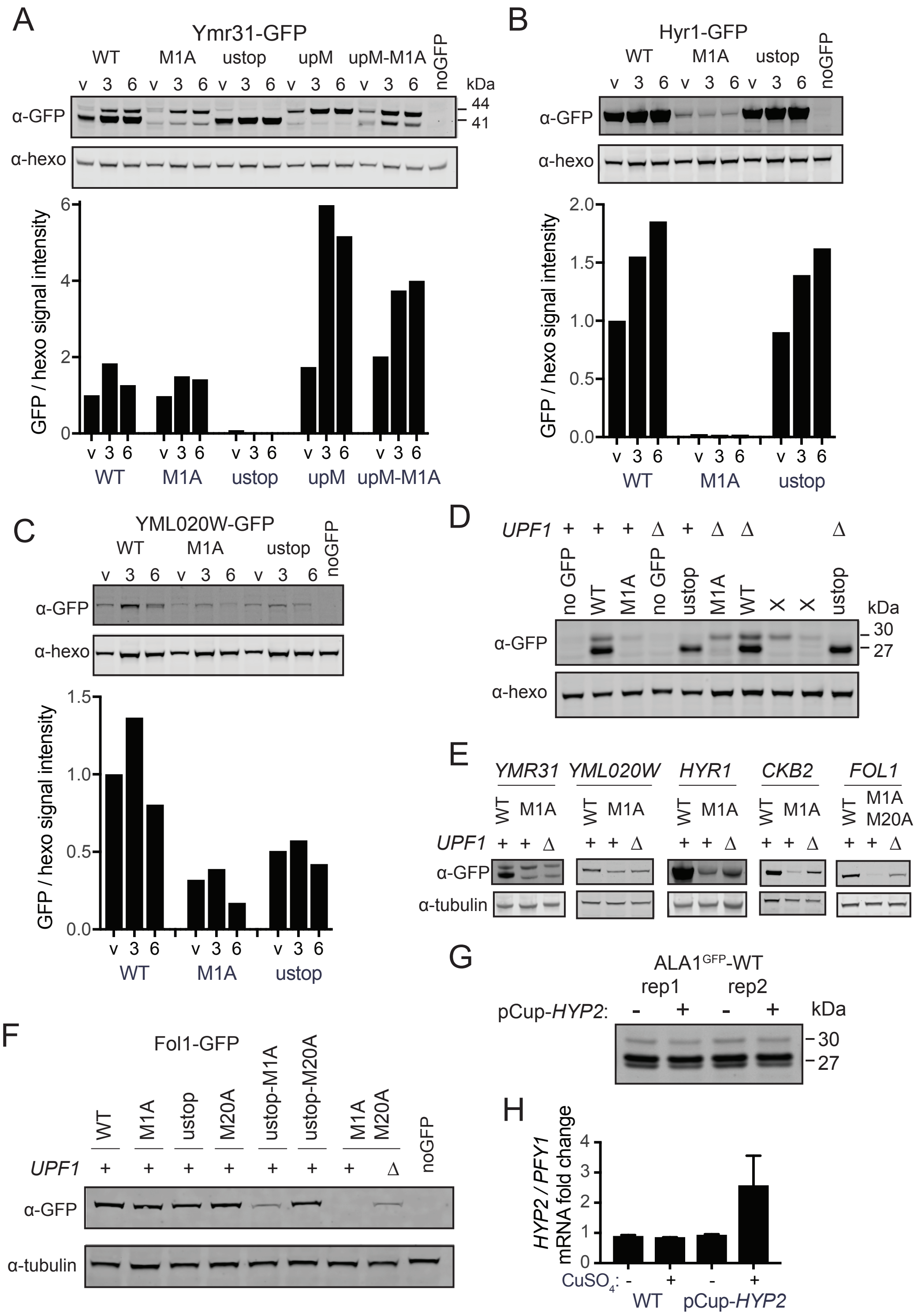


Figure S5: Western blot replicates and quantification for alternate isoforms, Related to Figures 4-7

(A) Replicate western blot of *YMR31-GFP* constructs, as in Figure 4C (top) and quantification of upper GFP band relative to hexokinase loading control for three replicates (bottom).

(B) Replicate western blot of *HYR1-GFP* replicates, as in Figure 5E (top) and quantification of GFP relative to hexokinase loading control for three replicates (bottom).

(C) Replicate western blot of *YML020W-GFP* replicates, as in Figure 5F (top) and quantification of GFP relative to hexokinase loading control for three replicates (bottom).

(D) Replicate western blot of *ALA1*^{GFP} reporter constructs, as in Figure 6A. Xs indicate samples that were not discussed in this study.

(E) Replicate western blots of *YMR31-GFP, YML020W-GFP, HYR1-GFP, CKB2-GFP and FOL1-GFP* with and without *upf1*Δ, as in Figure 6E.

(F) Replicate western blot of *FOL1-GFP* constructs, as in Figure 6I.

(G) Western blot of *ALA1*^{GFP}*-WT* reporter for cells with and without the *pCup-HYP2* construct with copper ($CuSO_4$) addition leading to overexpression of eIF5A for two replicates, which is quantified in Figure 7C.

(H) qPCR fold change of *HYP2* transcript relative to *PFY1* for cells with and without the *pCup-HYP2* construct with and without copper ($CuSO_4$) addition for three replicates. Related to Figure 7C.

Figure S6

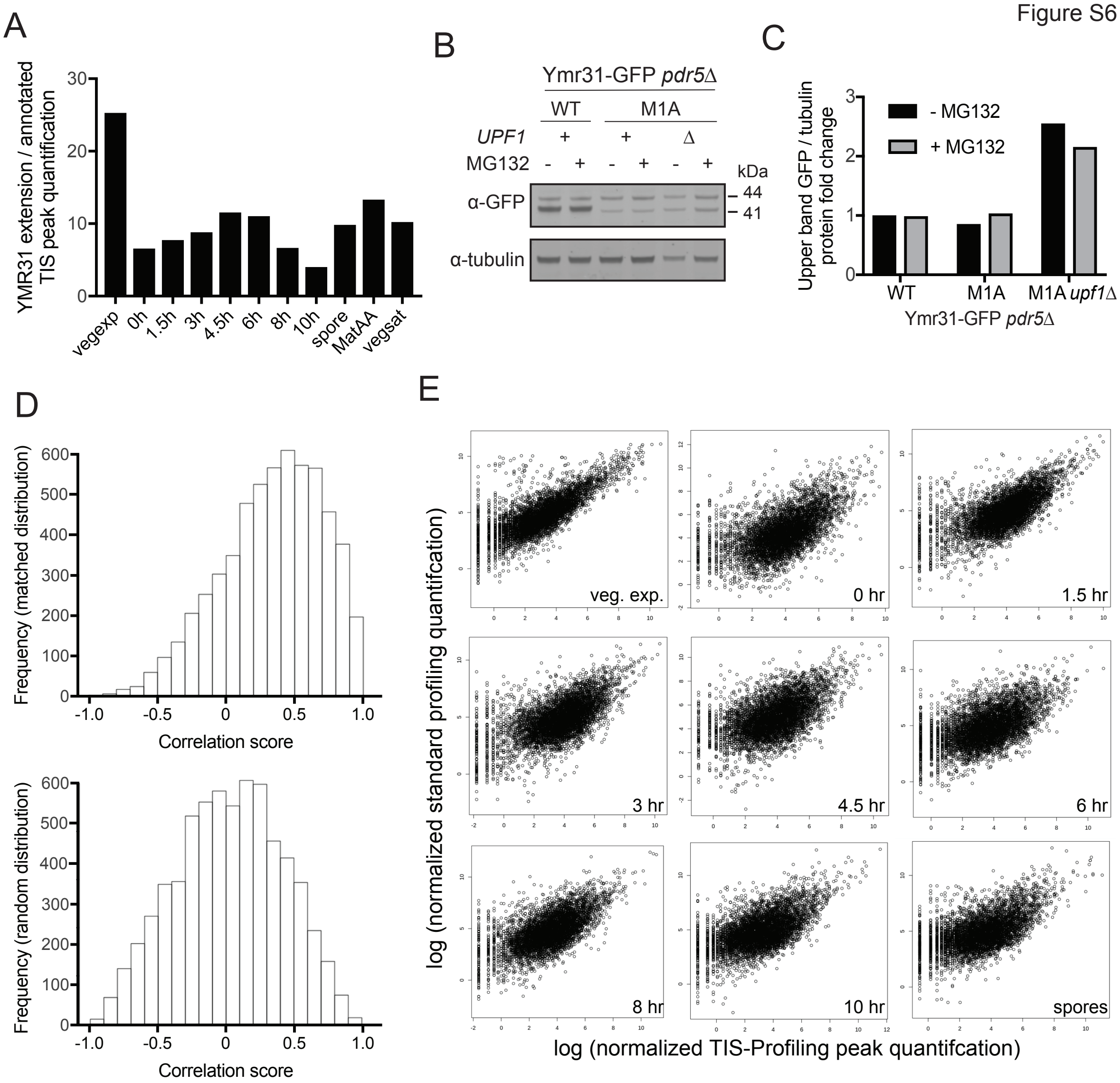


Figure S6: Positive correlation of TIS peaks with gene expression for annotated AUG sites but not near-cognate sites, Related to Figure 4

(A) Quantification of *YMR31* TIS-profiling peaks for the extension peak relative to the annotated peak. For all timepoints, the non-AUG extension peak is higher than the annotated AUG peak.

(B) Western blot of Ymr31-GFP with the proteasome inhibitor MG132. WT, M1A and M1A *upf1Δ* strains were treated with 100 uM MG132 for one hour. All strains are *pdr5Δ* to allow MG132 to enter cells, and samples were taken at 4h in meiosis.

(C) Quantification of the upper GFP band relative to tubulin for Figure S6B.

(D) Distribution of spearman correlation scores for peak height quantification comparing standard and TIS-profiling across all meiotic time points for all annotated genes (top) compared to a matched random distribution set (bottom). The set of annotated genes is significantly enriched for positive correlation scores, as seen by a K.S. test with a p-value of $<2.2\text{x}10^{-16}$.

(E) Scatter plots comparing peak quantification of TIS versus standard profiling for each timepoint.

Figure S7

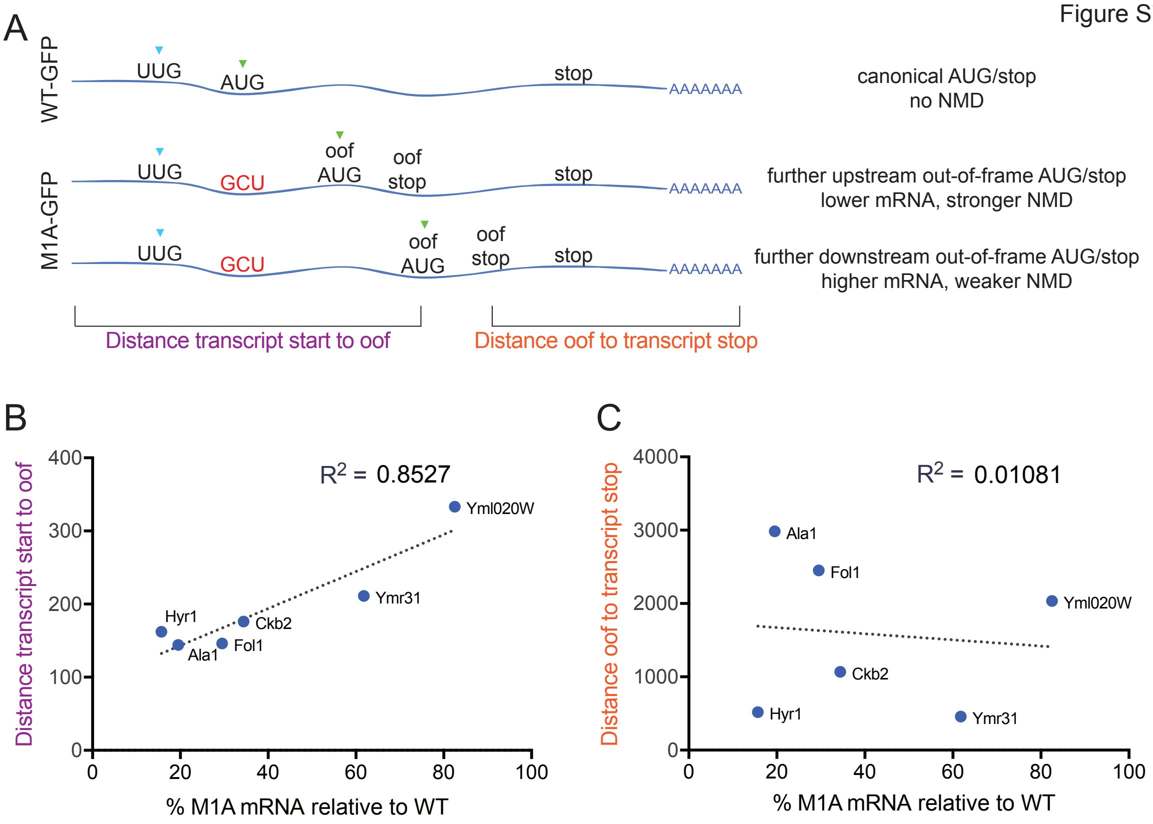


Figure S7: Effect of NMD for M1A transcripts does not correlate with distance from premature stop to transcript end, Related to Figure 6

(A) Diagram of a canonical ORF (*WT-GFP*) compared to two possible *M1A-GFP* constructs where the annotated AUG is mutated, leading to initiation at a later, out-of-frame (oof) AUG. Two different positions of the oof AUG/stop are shown, leading to different outcomes of NMD effect. For the mutated *M1A* construct, two distances are indicated, the distance between the transcript start to the oof AUG/stop (purple), and the distance from the oof AUG/stop to the transcript stop (orange).

(B) Correlation between the distance from the transcript start to the newly created oof ORF relative to the percent of *M1A / WT* mRNA level from Figure 6G, where a lower percentage indicates a stronger NMD effect and a higher percentage indicates a weaker NMD effect. A correlation with an R^2 value of 0.8527 is seen, indicating that a shorter distance from the transcript start to the oof ORF correlates positively with less *M1A* mRNA relative to *WT* and therefore stronger NMD.

(C) Correlation between the distance from the end of the newly created oof ORF to the end of the transcript relative to the percent of *M1A / WT* mRNA level from Figure 6G. A correlation with an R^2 value of 0.01081 is seen, indicating esentially no association between the distance from the oof ORF to transcript stop and the strength of NMD.

Figure S8

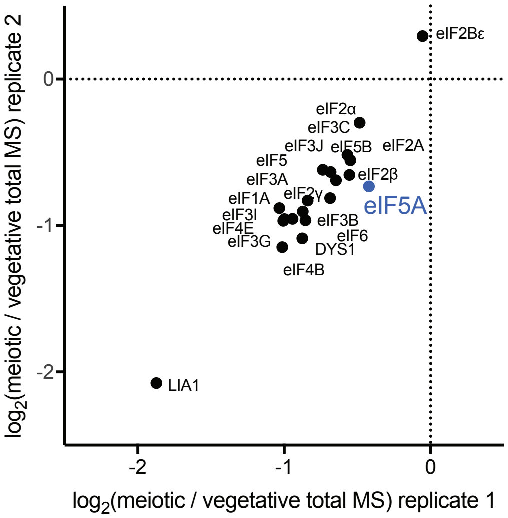


Figure S8: Total protein abundance of initiation and hypusination factors, Related to Figure 7

Enrichment of translation factors (as in Figure 7B) and hypusination factors Lia1 and Dys1 comparing meiotic and vegetative samples for two replicates, determined by quantitative (TMT10) mass spectrometry of whole cell extract from meiotic and vegetative cells.

Figure S9

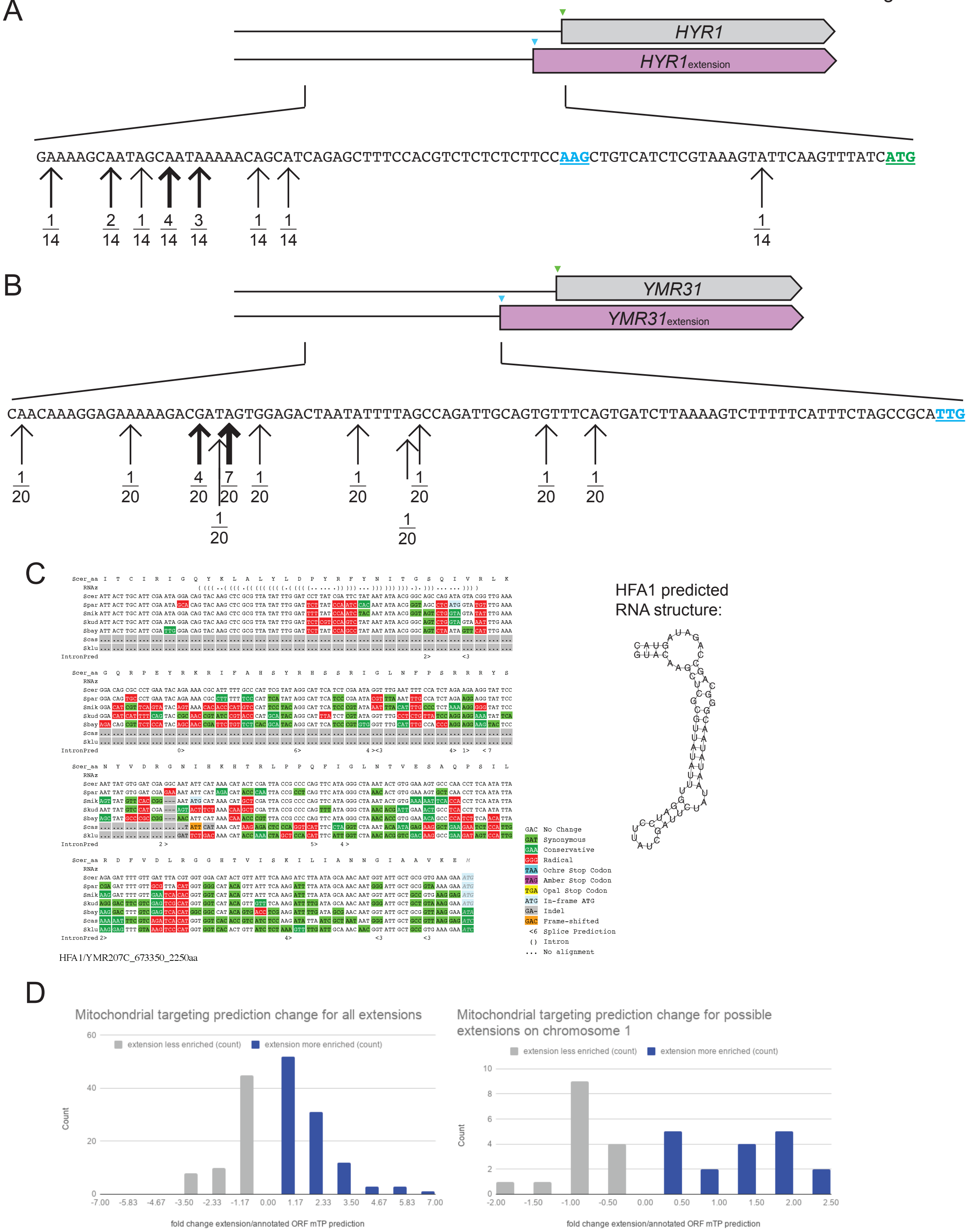

A
HYR1
HYR1extension
GAAAAGCAATAGCAATAAAAACAGCATCAGAGCTTTCCACGTCTCTCTCTTCCAAGCTGTCATCTCGTAAAGTATTCAAGTTTATCATG
1/14 2/14 1/14 4/14 3/14 1/14 1/14 1/14
B
YMR31
YMR31extension
CAACAAAGGAGAAAAAGACGATAGTGGAGACTAATATTTTAGCCAGATTGCAGTGTTTCAGTGATCTTAAAAGTCTTTTTCATTTCTAGCCGCATTG
1/20 1/20 4/20 7/20 1/20 1/20 1/20 1/20 1/20 1/20 1/20
C
Scer_aa
RNAz
Scer
Spar
Smik
Skud
Sbay
Scas
Sklu
IntronPred
HFA1/YMR207C_673350_2250aa
GAC No Change
GAT Synonymous
GAA Conservative
GGG Radical
TAA Ochre Stop Codon
TAG Amber Stop Codon
TGA Opal Stop Codon
ATG In-frame ATG
GA- Indel
GAC Frame-shifted
<6 Splice Prediction
() Intron
... No alignment
HFA1 predicted RNA structure:
D
Mitochondrial targeting prediction change for all extensions
extension less enriched (count)
extension more enriched (count)
Count
60
40
20
0
-7.00 -5.83 -4.67 -3.50 -2.33 -1.17 0.00 1.17 2.33 3.50 4.67 5.83 7.00
fold change extension/annotated ORF mTP prediction
Mitochondrial targeting prediction change for possible extensions on chromosome 1
extension less enriched (count)
extension more enriched (count)
Count
10
8
6
4
2
0
-2.00 -1.50 -1.00 -0.50 0.00 0.50 1.00 1.50 2.00 2.50
fold change extension/annotated ORF mTP prediction

Figure S9: HFA1 RNA structure and mitochondrial targeting sequence prediction, Related to Discussion

(A) 5’RACE analysis of *HYR1.* Locations of transcription start sites are indicated with arrows, with the number of sequencing reads at that site indicated. A total of 14 transcription start sites were sequenced.
(B) 5’RACE analysis of *YMR31*. Locations of transcription start sites are indicated with arrows, with the number of sequencing reads at that site indicated. A total of 20 transcription start sites were sequenced.
(C) Structure prediction for *HFA1*, shown by RNAz depiction in alignment (left), and in predicted structure form (right).
(D) Mitochondrial targeting prediction score changes for extension ORFs relative to the annotated ORF’s score (left) and for possible extensions of annotated ORFs on chromosome 1 relative to the annotated ORF’s score (right).