Title

Non-Canonical Protein Isoforms Produce Diversity of Protein Function and Localization in Budding Yeast

Author

Higdon, Andrea Lynn

Publication Date

2023

Peer reviewed|Thesis/dissertation

Non-Canonical Protein Isoforms Produce Diversity of Protein Function and Localization in Budding Yeast

By

Andrea Lynn Higdon

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gloria Brar, Chair
Professor Nicholas Ingolia
Professor Britt Glaunsinger
Professor Nir Yosef

Summer 2023

Abstract

Non-Canonical Protein Isoforms Produce Diversity of Protein Function and Localization in Budding Yeast

by

Andrea Lynn Higdon

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Gloria Brar, Chair

Global methods for assaying translation have greatly improved our understanding of the protein-coding capacity of the genome. In particular, it is now possible to perform genome-wide and condition-specific identification of translation initiation sites through modified ribosome profiling methods that selectively capture initiating ribosomes (TIS-profiling). Here we apply this approach to meiotic and mitotic timepoints in budding yeast and show the surprising diversity of protein products that can be revealed by such methods.

Chapter 2 describes our use of TIS-profiling to globally annotate translation initiation sites in yeast, with a particular focus on alternative N-terminally extended protein isoforms, which initiate from near-cognate (non-AUG) start codons upstream of annotated AUG start codons. We identified 149 genes with an extended isoform and show that these isoforms are produced in concert with canonical isoforms and are 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.

Despite our success in identifying extended protein isoforms, high-confidence identification of new coding regions that entirely overlap annotated coding regions – including those that encode truncated protein isoforms – remained challenging. As described in Chapter 3, we developed a sensitive and robust algorithm focused on identifying N-terminally truncated proteins genome-wide, identifying 388 truncated protein isoforms. We performed extensive experimental validation of these truncated proteins and defined two general classes. The first set lack large portions of the annotated protein sequence and tend to be produced from a truncated transcript. We show two such cases, Yap5^{truncation} and Pus1^{truncation}, to have condition-specific regulation and functions that appear distinct from their respective annotated isoforms. The second set of N-terminally truncated proteins lack only a small region of the annotated protein and are less likely to be regulated by an alternative transcript isoform. Many localize to different subcellular compartments than their annotated counterpart,

representing a common strategy for achieving dual localization of otherwise functionally identical proteins. Together these findings support the adoption of less static views of gene identity and a broader framework for considering the translational capacity of the genome.

Table of Contents

List of Figures

List of Tables

Acknowledgements

A PhD journey is long and requires all kind of help and support, both large and small. I want to thank everyone who helped make this work possible.

I have to start by thanking my advisor Gloria Brar, who has been a constant source of support throughout my PhD. Gloria is an amazing cheerleader and a patient mentor, and I always came out of meetings with her feeling more positive about my science and more ready to tackle the next challenges. She also gave me an incredible amount of freedom to explore my interests and trusted me to take my project in directions that were out of her wheelhouse, while always providing insight, support, and encouragement. I could not have done this work without her.

The BrÜn Lab has been a wonderful place to work over that last several years. It's amazing to come to work and be surrounded by such a supportive, smart, and fun group of people. Everyone is quick to lend a hand, a reagent, or a word of advice.

Amy Eisenberg was my rotation mentor and was so welcoming and patient with me as I got started in the lab. Once I joined the lab, she continued to be an incredibly generous mentor, always there to help me out or talk things through. She also welcomed me onto her project as a collaborator, and it was such a pleasure to work with her over the years.

I would like to thank all of my mentees over the course of my PhD. They brought energy, enthusiasm, and new perspectives to the work and I learned so much from working with them all. Special thanks to Nathan Won for all his hard work on the truncations project.

The lab would not be what it is without Elçin Ünal. Her sharp intellect and passion for science makes everyone around her better. Christiane Brune is the best bay mate and best lab manager, keeping everything organized in the lab while always having time to ask how my day is going. A big thank you to Kate Morse and Amanda Su for lending their RNA expertise and for being so fun to commiserate with about late-stage grad school. Tina Sing is so generous with her expertise and is just a wonderful person to grab a tea with and talk life and science. Emily Powers is such an incredible friend and coworker and the person I go to with all my stupidest questions and most half-baked ideas. I miss our Asha dates so much.

I would like to thank my thesis committee – Nick Ingolia, Britt Glaunsinger, and Nir Yosef – for sharing their advice and expertise over the years. I feel very lucky to have been able to rotate with both Nick and Britt. I learned a lot in my short time in their labs that I have taken with me through my PhD. Over the years, Nick has always been generous with his time and expertise in translation. Britt's MCB290 seminar on how to give a good talk was so helpful for me in learning how to communicate science, and I think about it every time I'm making a presentation.

I'm so grateful for the friends I made through MCB. I miss having JJJAMM all together in Berkeley but I'm so glad we've stayed close despite scattering across the country. Thanks to Maiko for tea dates that were always a highlight of my week, Mark for fun photo walks and his dark sense of humor, Jack for always bringing so much energy and enthusiasm to everything, and Jana for being such a calm and steady presence but always ready to laugh. Natalie and Snigdha were always so fun to grab a beer with and chat about grad school life.

I would also like to thank the communities at Berkeley Ballet Theater and Berkeley Ironworks. Without ballet and climbing I don't know how I would have made it through grad school.

Thank you to Josh for being a wonderful partner and supporting me in so many ways. He helped my science in direct ways – being a sounding board for ideas, listening to practice talks, reading drafts of writing, helping me when I got stuck on programming – and helped me in countless indirect ways as well. He picks up the slack when I'm going through a busy time and is always there to listen and help me relax when I get too stressed. No one is better at helping me stay calm and focus on what is really important.

Lastly, I have to thank my family. My sister Lauren is incredibly smart and hardworking and has always been a role model for me. She has also always been there to support me and talk things through. My parents have always supported my education in every way possible and encouraged my interest in science and the natural world. None of this would have been possible without them.

Chapter 1: General Introduction

Portions of this chapter were adapted from the following publication:
Higdon, A.L., Brar, G.A., 2020. Rules are made to be broken: a “simple” model organism reveals the complexity of gene regulation. Current Genetics 1–8. https://doi.org/10.1007/s00294-020-01121-8

1.1 Our understanding of the protein coding capacity of the genome is incomplete

A key outcome of most gene expression events is the production of proteins, the tiny workhorses that execute a complex array of tasks within the cell. In principle, the information necessary to determine the identity, function, and regulation of proteins is encoded in the genome. However, despite knowing the full genome sequence of budding yeast for over two decades, our reading and interpretation of its small and compact genome remains incomplete and largely ignores conditional differences in genome decoding (Goffeau et al., 1996; Wood et al., 2019).

Decades of fruitful research on the functions and regulation of proteins has benefited from gene annotations, which integrate DNA sequence information with predictions about the regions of these sequences that are eligible for decoding into proteins. Initial gene annotations provided a valuable starting point for identifying protein-coding regions, but they were intrinsically limited by the methods available at the time. They relied on our understanding that proteins are made from open reading frames (ORFs) that begin with an AUG start codon and end at an in-frame stop codon (reviewed in Aitken and Lorsch, 2012). And because there would be an overwhelming number of short ORFs throughout the genome, even with these rules, length restrictions were imposed in order to prioritize more likely protein-coding genes. Length limits (usually greater than 100 codons) were based on sizes of well-characterized proteins and assumptions about the length of polypeptide chain needed to fold stably (reviewed in Dinger et al., 2008). While such rules were necessary to avoid an unwieldy number of erroneous predictions, they also excluded many gene products that we now know to be functional.

The advent of RNA-seq was critical for deepening our understanding of eukaryotic genome decoding by revealing transcribed regions of the genome without the biases intrinsic to single-gene and microarray studies, which provided important insights but depended on existing gene annotations. RNA-seq, by comparison, enabled comprehensive identification of the regions of the genome that produce RNA and are therefore candidate protein-coding regions (Nagalakshmi et al., 2008; reviewed in Wang et al., 2009). This method proved especially valuable in organisms with prevalent alternative splicing, which had previously made gene predictions particularly difficult. It also revealed an abundance of transcription in regions without predicted ORFs, which was provisionally assumed to correspond to production of non-coding RNAs, a subset of which have since been shown to serve important RNA-based cellular functions (reviewed in Schmitt and Chang, 2017).

The wealth of information contained in the transcriptome can easily lead to the assumption that once the identity and abundance of a transcript is known, it is straightforward to predict the identity and abundance of the resulting proteins. However, the invention of ribosome profiling, which allows global empirical measurements of what proteins are made and when, has revealed unexpected complexity to translation (Ingolia et al., 2009). In applying this method to budding yeast meiosis, for example, we observed thousands of cases where transcript abundance over time does not predict protein abundance for annotated genes. In fact, in several hundred cases, we found an *inverse* relationship between mRNA and translation or protein levels, upending longstanding paradigms for how gene expression generally works in this simple eukaryote (Cheng et al., 2018).

In this work, we revisit the question of which regions of the yeast genome are decoded into protein using a global method for translation initiation site (TIS) mapping (Eisenberg et al., 2020). This revealed many cases that defy three of the simplest assumptions about genome coding in yeast: that a given locus produces one mature transcript which produces one protein product, that coding regions always initiate at AUG start codons, and that gene identity is statically encoded by genome sequence. Importantly, we are able to identify many protein isoforms that would easily fly under the radar with interrogation by standard molecular biology approaches and are challenging to identify even by standard ribosome profiling, which captures elongating ribosomes (Ingolia et al., 2009). Our findings suggest not only a need for revision to the broadly accepted rules of gene regulation, but also that there is more information encoded in the genome than can be readily inferred from sequence analysis alone, even in the well-studied and simple budding yeast.

1.2 Translation initiation site profiling detects non-canonical protein isoforms

Our straightforward definition of what constitutes a coding region remained more or less intact for decades in the absence of tools to empirically and systematically put it to the test. Ribosome profiling, a method for capturing and sequencing ribosome-protected fragments of mRNA, started to change our understanding by providing a global picture of the positions and levels of translation (Brar and Weissman, 2015; reviewed in Ingolia, 2014; Ingolia et al., 2009). A modified version of ribosome profiling—in which cells are pretreated with a specific type of translation inhibitor, such as harringtonine or lactimidomycin, which block the first elongation cycle of the ribosome—strongly favors the capture of ribosomes that have just completed translation initiation (Ingolia et al., 2011; Lee et al., 2012; Schneider-Poetsch et al., 2010). Relative to traditional ribosome profiling, this allows cleaner detection of translation initiation sites, unobscured by signal from elongating ribosomes within ORFs (Figure 2.1A). Its application to mammalian cells revealed complexity in TIS usage, but these data are challenging to interpret (Fields et al., 2015; Ingolia et al., 2011). Identifying the coding region based on the start codon depends on transcript isoform definitions, which are incomplete and likely to be highly conditionally regulated in mammals, based on the small subset that have been studied in great depth (reviewed in Baralle and Giudice, 2017). In addition, when

multiple transcript isoforms are present, it is difficult to unambiguously assign TIS peaks to a specific transcript isoform (Figure 1.1).

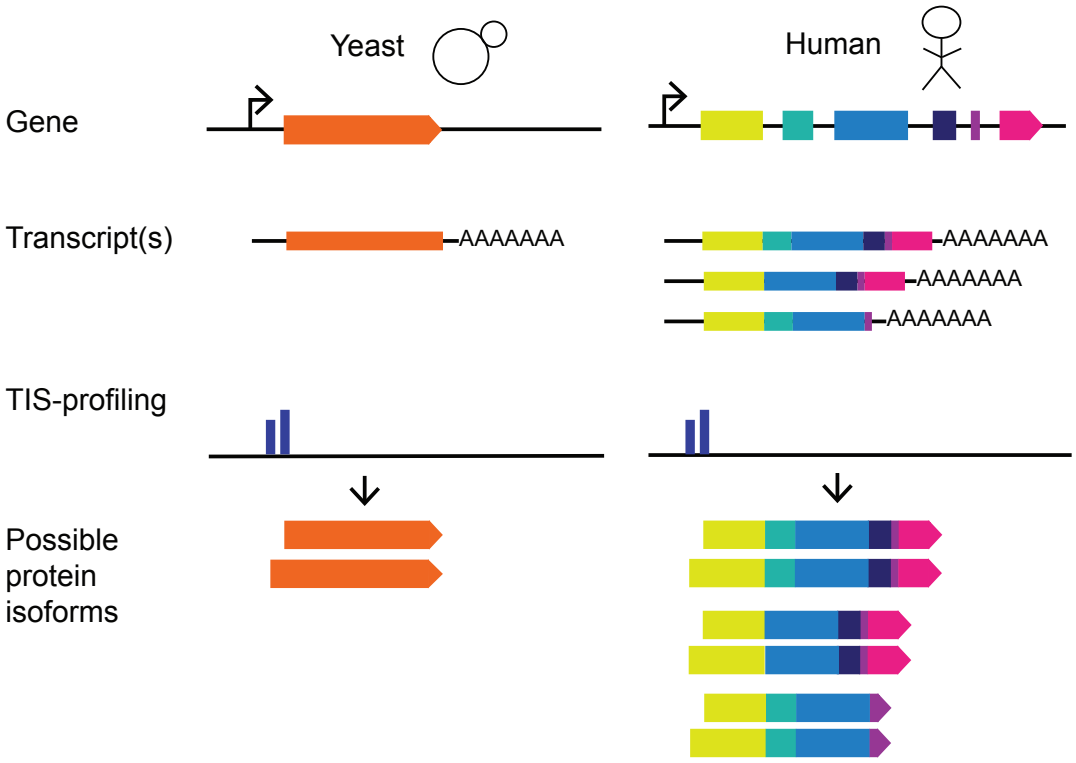


Figure 1.1 The compact nature of the yeast genome facilitates unambiguous protein isoform identification
Comparison of protein isoform prediction from TIS-profiling data in yeast (left panel) and humans (right panel). Alternative transcript isoforms present in humans contribute to ambiguity in protein isoform identity.

1.3 Yeast meiosis as a context for studying non-canonical translation

Meiosis is a highly regulated developmental program that, in yeast, converts a diploid cell into four haploid spores. It requires dynamic and precisely regulated waves of gene expression changes to achieve dramatic morphological changes to cellular components (reviewed in Marston and Amon, 2004; van Werven and Amon, 2011). Standard ribosome profiling revealed large temporally regulated changes in the quantities of proteins made from nearly all annotated genes, and also hinted at qualitative changes in the identity of the proteins being produced, including evidence for translation in 5' leaders (traditionally defined as UTRs or "untranslated regions") of approximately half of mRNAs expressed in meiosis (Brar et al., 2012). The translation in many of these 5' leaders could be attributed to upstream open reading frames (uORFs), and transcripts showing such translation often appeared to have translation of several overlapping uORFs. Based on this, and the ensemble nature of ribosome profiling data, which reveals all translated positions for the pool of a given transcript in the samples collected, it was difficult to unambiguously assign reads to specific ORFs for these transcripts. TIS-profiling, in contrast, enables comprehensive identification of ORF start codons by

enriching for ribosome footprint signal representing post-initiation ribosomes. By applying this method to samples spanning the stages of meiosis, we observed widespread translation initiation for non-canonical regions, including uORFs and N-terminally extended and truncated proteins (Eisenberg et al., 2020).

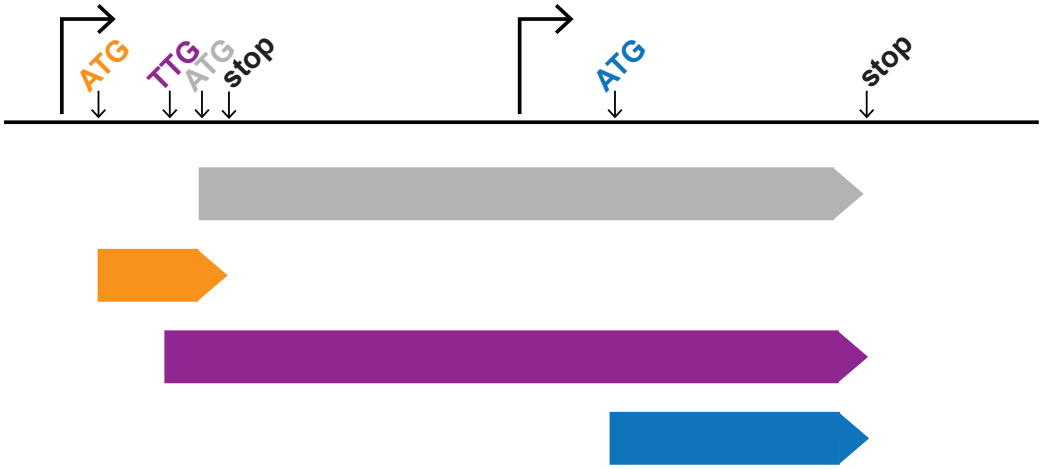


Figure 1.2 Types of non-canonical protein isoforms
An annotated protein isoform (gray) is illustrated along with three types of non-canonical gene products discussed in this work: an upstream open reading frame (uORF; orange), an extended protein isoform (purple) and a truncated protein isoform (blue). Transcript architecture is illustrated above, with arrows representing transcripts start sites. Start codons (ATG or near-cognate TTG) for protein isoforms are indicated in corresponding colors.

The work in Chapter 2 focuses on N-terminally extended proteins, which are produced via translation initiation from in-frame near-cognate start codons upstream of annotated start codons. In our data, we identified 149 genes with such extended isoforms, representing a small but notable fraction of the ~6000 annotated yeast genes. These isoforms are, as a class, more abundantly produced during meiosis relative to vegetative growth, fitting with a general trend of increased translation from canonical and non-canonical start codons in upstream regions during meiosis. Our approach in this chapter was robust for identifying extended protein isoforms but appeared to have a large false-negative and false-positive rate for truncated protein isoforms. Chapter 3 describes our efforts to address this issue through the development of an algorithm specifically designed for identifying truncated isoforms using the TIS-profiling data. With this method we identified 388 truncated protein isoforms. Like extended isoforms, truncations were also more prevalent in meiosis than mitotic growth, suggesting that non-canonical translation products are generally more common in meiosis.

Chapter 2: Translation initiation site profiling reveals widespread synthesis of non-AUG initiated isoforms in yeast

This chapter was adapted from the following manuscript:
Eisenberg, A.R., Higdon, A.L., Hollerer, I., Fields, A.P., Jungreis, I., Diamond, P.D., Kellis, M., Jovanovic, M., Brar, G.A., 2020. Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast. Cell Systems 11, 145-160.e5. https://doi.org/10.1016/j.cels.2020.06.011

2.1 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 (~30nt) 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 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 pre-treated 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 have relatively simple transcript architectures with fewer known cases of complexity, such as

from alternative splicing, despite extensive analyses of their transcriptome (Davis et al., 2000; Guisbert et al., 2012; Hossain et al., 2011; Juneau et al., 2009; 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. 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.

2.2 Results

2.2.1 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 2.1A; 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 S2.1A; 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 S2.1B; 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 S2.1C) and resulted in strong TIS enrichment of ribosome footprints (Figure S2.1D). 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 2.1B). 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 2.1C). 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* (Figure 2.1C, 2.1D).

We confirmed that our data accurately reported the expected positions and condition-specificity of both canonical and non-canonical start sites 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 2.1E). TIS-profiling also revealed peaks at known non-canonical TISs, including the four AUG-initiated uORFs known to regulate *GCN4* (Figure 2.1F). 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 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 2.1G) and concluded that our TIS-profiling protocol could capture both known canonical and non-canonical TISs.

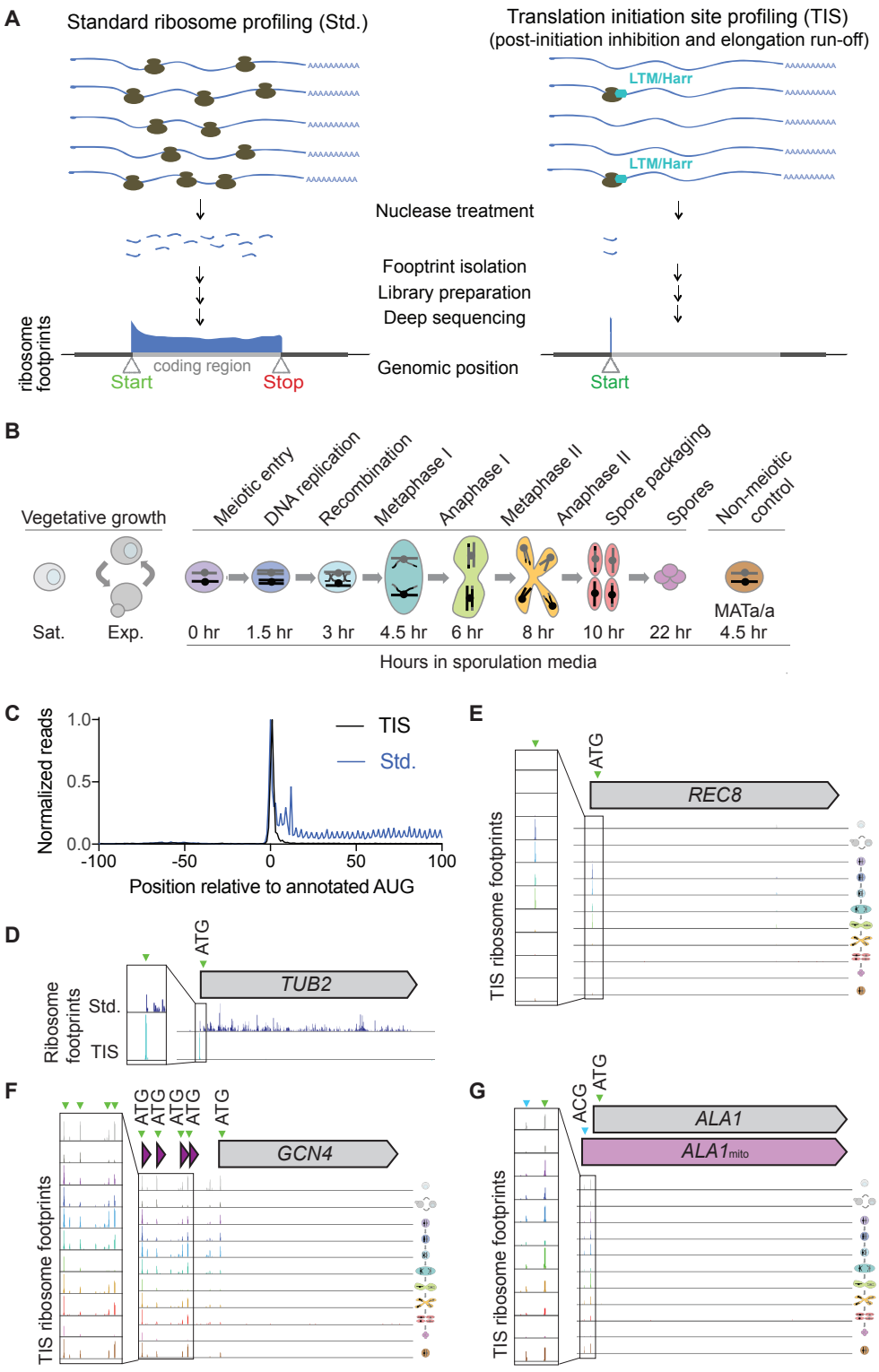


Figure 2.1 Translation initiation site ribosome profiling in mitotic and meiotic yeast cells
(A) Cartoon comparing standard (Std., left) and translation initiation site (TIS, right) ribosome profiling, with representative ribosome footprint profiles for a typical ORF.
(B) Schematic of yeast cell stages and samples collected for TIS-profiling, including vegetative saturated (sat.), vegetative exponential (exp.), 0 hr, 1.5 hr, 3 hr, 4.5 hr, 6 hr, 8 hr, 10 hr, and 22 hr after addition to sporulation media, and a MATa/a non-meiotic control taken at 4.5 hr 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. Reads are normalized to position zero.
(D) Comparison of standard and TIS-profiling for TUB2, a representative gene, from all timepoints combined. Green arrowheads indicate ATG initiation sites and inset shows close-up view of region around initiation site.
(E) TIS-profiling of REC8,
(F) GCN4
(G) and ALA1, showing ribosome footprints at the time points indicated in Figure 2.1B. Green arrowheads indicate ATG initiation sites and blue arrowheads indicate non-ATG initiation sites.

2.2.2 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 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 timepoint-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 2.2A). 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; Figure S2.2A, S2.2B). An interesting category of uncalled annotated ORFs are cases of apparent misannotation, such as *PEX32* and *RSB1*, where the likely predominant initiation site based on TIS-profiling and ORF-RATER analysis is upstream or downstream of the annotated TIS. 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 (Figure 2.2B, 2.2C). 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 S2.2C). 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 site in the TIS-profiling data (Figure S2.2D).

It is not surprising that ORF-RATER was generally successful at calling annotated canonical ORFs because the approach trains on this set. To assess the success of 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. 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 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 2.2A). This number was much higher than we expected to represent true translated regions, and we thus 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 2.2A) are likely to be false positives, based on visual analysis of the LTM data (such as for *SIN3* and *CDC15*; Figure S2.2E, S2.2F), and the fact that there were a median of 16 internal ORFs called per annotated gene (score >0.1; Figure S2.2G). 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 internal 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 2.1G), 30,642 uORFs, and 450 downstream ORFs in which translation initiates within predicted 3'UTR regions (Figure 2.2A). 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 2.2D), N-terminal extensions (for example, *CKB2*; Figure 2.2E), and downstream ORFs. We further refined the N-terminal extension class based on length, with a cutoff of greater than 10 amino acids based on

the minimum length predicted for function such as targeting signal or binding domains (Figure S2.3A; Almagro Armenteros et al., 2017; Fukasawa et al., 2015). Excluding AUG-initiated extensions, many of which are likely to represent misannotations (as for *RSB1*, Figure 2.2C), left 231 extensions, representing 160 unique genes, as some genes contained multiple predicted extensions (Figure S2.3B; this number was ultimately adjusted to 149 based on misannotations discovered through conservation analysis).

A

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

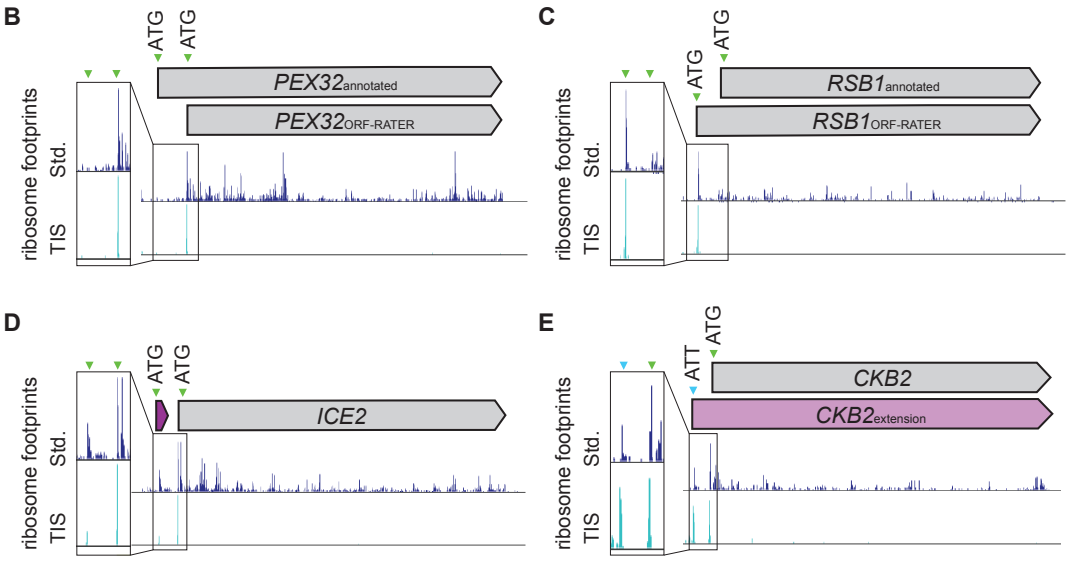


Figure 2.2 ORF-RATER annotations of TIS-profiling
(A) Numbers of different types of ORFs called by ORF-RATER at two different score cutoffs - a high score cutoff (> 0.5) and a low score cutoff (> 0.1). Truncation and internal out-of-frame numbers are likely overestimates due to high rates of false positives, indicated with a *.
(B) Comparison of standard and TIS-profiling for: (B) PEX32, which has a likely incorrect start site annotation. The likely correct (later) 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 extension ORF with a non-AUG TIS identified by ORF-RATER.

2.2.3 Translation of uORFs and N-terminal extension 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 2.3A) but no difference between the vegetative and meiotic LTM-based ribosome footprints in regions surrounding annotated stop codons (Figure 2.3B). The increased read density in 5' leaders during meiosis could reflect an increase in translation of either uORFs or N-terminal extension 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 extensions are increased in meiosis, while the number of annotated and downstream ORFs are similar between the two conditions (Figure 2.3C). 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 2.3D). The translation of both uORFs and N-terminal extensions 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 of 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 *ALA1*, *YMR31/KGD4*, *HYR1/GPX3, TRZ1* and *HFA1* loci, as well as 155 other genes, which we proceeded to evaluate in more detail (Heublein et al., 2019; Kritsiligkou et al., 2017; Monteuuis et al., 2019; Suomi et al., 2014; Tang et al., 2004).

2.2.4 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 2.3D) 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. Four of the remaining seven are misannotations, similar to *RIM11* (Figure S2.2D), and one (*YPL034W*) includes a likely frameshifting event. 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 start-site 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 (David et al., 2006; Nagalakshmi et al., 2008). We found 6446 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 (Figure 2.3E, 2.3F). 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., 2008). 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 (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.

2.2.5 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 S2.3A). 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 2.4A), 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.

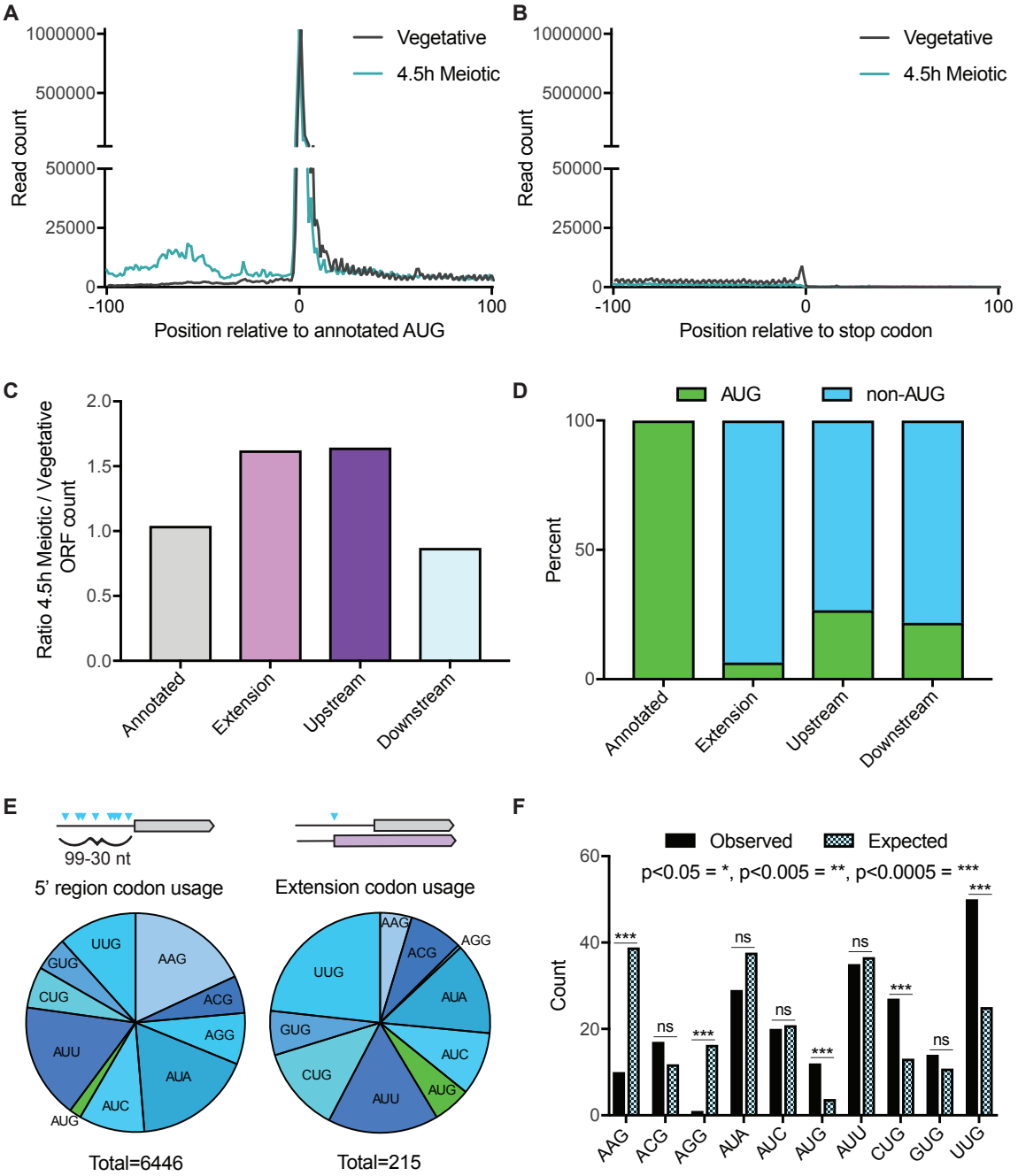


Figure 2.3 Specificity of uORF and N-terminal extension translation is partly dependent on condition and start codon identity

(A) Metagene plot of read counts from vegetative exponential and 4.5 hr time points, 100 nucleotides upstream and downstream of annotated AUG start codons. Reads are normalized to aligned reads for that timepoint. Increased read density is observed for the meiotic timepoint upstream of annotated start codons, but not after.
(B) Metagene plot of read counts from vegetative exponential and 4.5 hr time points, 100 nucleotides upstream and downstream of annotated stop codons. Reads are normalized to aligned reads for that timepoint.
(C) Relative numbers of ORFs from different ORF categories, comparing the 4.5 hr meiotic time point to vegetative exponential. More extension 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 sites, while 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 sites for all possible TISs (left) and called extension ORFs (right). Of the 6446 sites possible in 5' regions, 215 are observed to initiate translation of extension ORFs called by ORF-RATER.
(F) Near-cognate codon usage for called extensions (observed) compared to relative abundance of all possible near-cognate codons within UTRs (expected). Expected distribution is derived from counts of all possible TISs in the 99-30 nt upstream of annotated AUG start sites. P-values calculated by Fisher's exact test, with $p<0.05$ = *, $p<0.005$ = **, $p<0.0005$ = ***, and ns = not significant.

2.2.6 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* isoforms, a C-terminally GFP tagged version of this protein was expressed with either the wild-type start codon (*WT*), the annotated ATG start codon 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 2.4B). Samples were collected in vegetative cells, and at 3h and 6h 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 (Figure 2.4C, S2.5A). The extended form of Ymr31 was ten times lower in abundance than the canonical form in *WT* cells by western blot analysis (Figure S2.5A), 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 (Figure 2.4D, S2.6A).

Mutation of the near-cognate 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 2.4C). This suggested that the native near-cognate TIS is used inefficiently for translation initiation relative to AUG, consistent with *in vitro* experiments comparing AUG and near-cognate initiation (Chen et al., 2008; Kolitz et al., 2008). 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 (Figure S2.6B, S2.6C).

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 start site peaks appears to reflect known regulation patterns during meiosis for characterized genes. Across annotated ORFs, there is 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 (Figure S2.6D, S2.6E). This is seen by comparisons of individual time points (Figure S2.6E), as well as by calculating correlation scores for each gene across all time points (Figure S2.6D). Individual examples, such as Rec8 (Figure 2.1E), show a strong correlation between TIS-profiling peaks and standard profiling reads (Pearson correlation coefficient = 0.833), and the correlation scores are significantly enriched for positive scores compared to a random distribution of genes (Figure S2.6D). This is consistent with a study using a similar approach in mammalian cells that suggest 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.

A

Extension ORF 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	

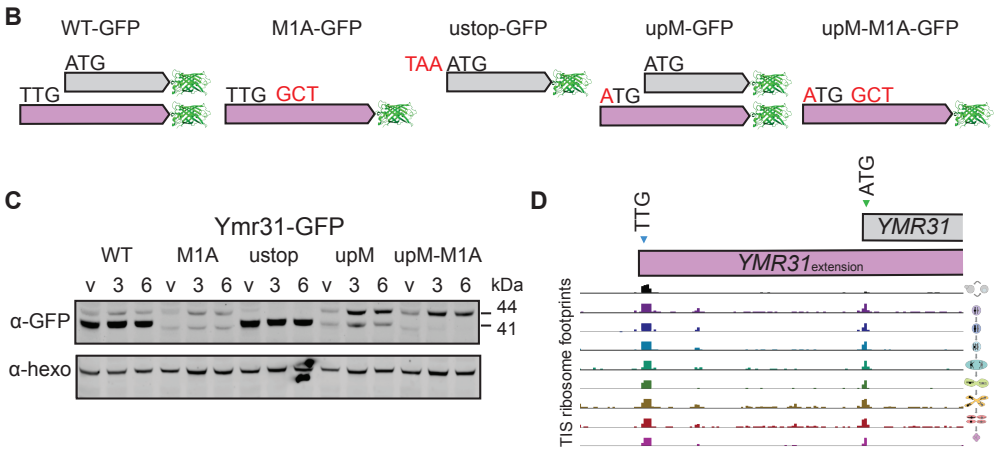


Figure 2.4 The abundance of near-cognate-initiated isoforms is not reflective of TIS-profiling peak height

(A) Extension ORFs 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 extension ORFs. 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 upstream non-AUG start codon to a methionine.

(C) Western blot of Ymr31-GFP showing the WT construct with two bands corresponding to the extension ORF (44 kDa) and annotated ORF (41 kDa). M1A and ustop constructs show the extension ORF 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 3h and 6h 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 2.1B, with the extension (TTG) and annotated (ATG) start sites indicated.

2.2.7 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 scores are more likely in conserved coding regions (Figure 2.5A). 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. Alignments of individual true extensions illustrate the degree of conservation, which for Ymr31 is high, reflected in its high PhyloCSF score (Figure 2.5B). We further evaluated two true extensions with high PhyloCSF scores, Hyr1 and Yml020w (Figure 2.5C, 2.5D). In these cases, as well as for nearly every other extension- containing 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 (Figure 2.5E, S2.5B). Similarly, we detect an N-terminally extended isoform of Yml020w in cells carrying the *YML020W-M1A* construct (Figure 2.5F, S2.5C).

The majority of extensions analyzed had scores below zero, suggesting a lack of conserved functionality (Figure 2.5A). 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 site and the reading frame are conserved in five species but the extension itself had a negative PhyloCSF score (-3.587; Figure 2.5A, 2.5G). 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.

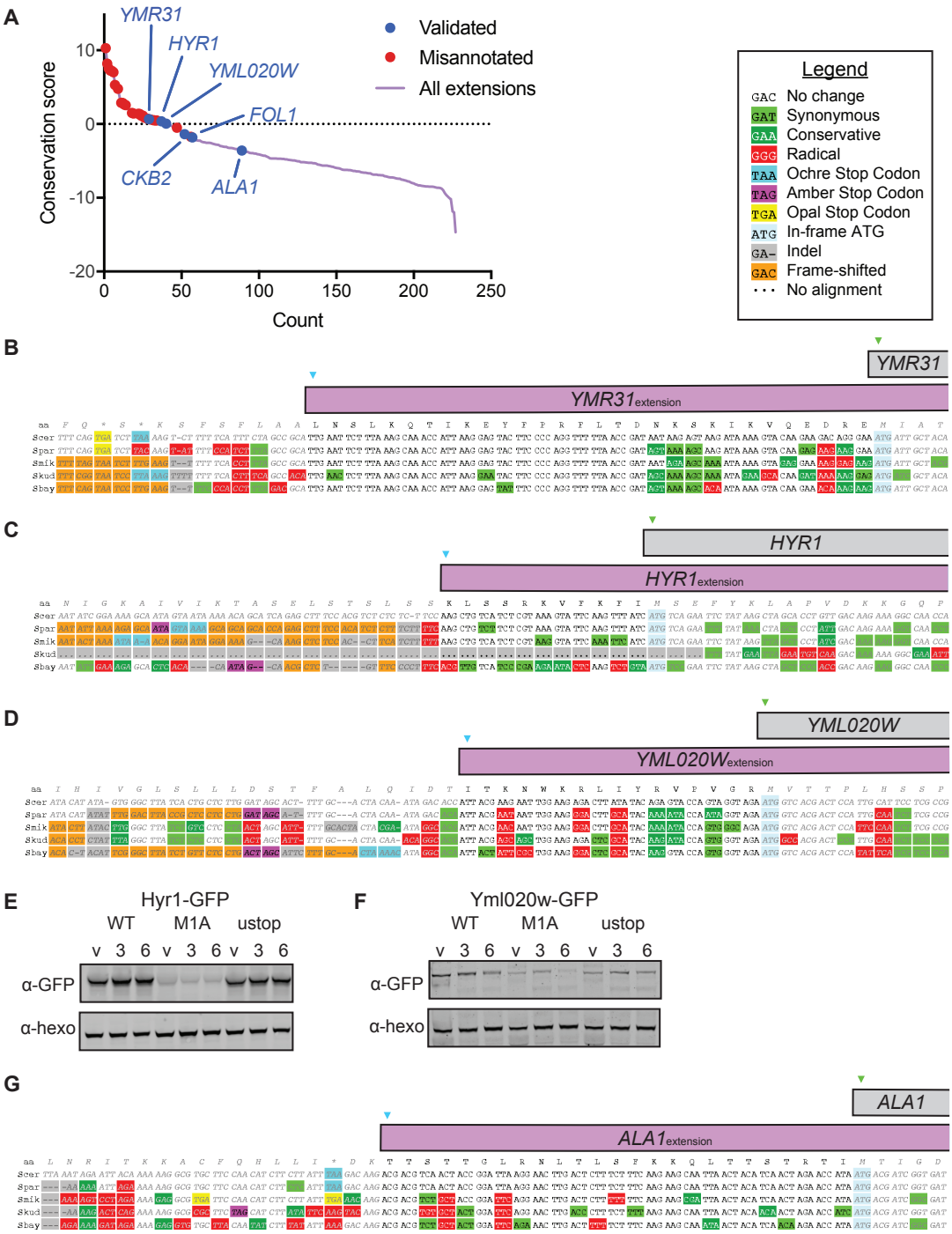


Figure 2.5 Most ORF extensions are poorly conserved
(A) Plot of PhyloCSF conservation scores for extension ORFs. Misannotated extensions are shown with red dots, and validated extensions are shown with blue dots, including three previously validated extensions (YMR31, HYR1 and ALA1). The additional validated extensions (YML020W, CKB2 and FOL1) were validated in this study.
(B) Alignments showing level of conservation for YMR31,
(C) HYR1,
(D) and YML020W, 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 3h and 6h 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 3h and 6h after addition to sporulation media.
(G) Alignment showing level of conservation for ALA1, which has a negative conservation score.

2.2.8 Transcripts with canonical start site mutations are NMD 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 2.6A). When the canonical start codon was present in the reporter (*ALA1^{GFP}-WT*), both Ala1 reporter isoforms were observed (Figure 2.6B, 2.6C, S2.5D). The canonical Ala1 reporter isoform could be detected alone in extract from cells carrying the *ALA1^{GFP}-ustop* construct (Figure 2.6B, 2.6C). Surprisingly, in cells carrying the *ALA1^{GFP}-M1A* construct, however, we could not detect production of either protein isoform (Figure 2.6B, 2.6C). The dramatic difference in production of the extended reporter with and without the canonical start site 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 alone (Chen et al., 2008; Clements et al., 1988; Kolitz et al., 2008). 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 2.6D). 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 ATG that results in early translation termination (Figure S2.7A). 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 (*upf1*Δ*)*, although not to the level of the extended isoform in the *ALA1^{GFP}-WT* reporter construct (Figure 2.6B-D).

In addition to the *ALA1* reporters, several other *M1A* constructs showed little to no tagged protein in otherwise WT cells. This was consistent with our findings for the extended isoform of Hyr1*,* which was detected in our mass spec dataset (Figure 2.4A) but was detected at extremely low levels in cells carrying the *HYR1-M1A* construct (Kritsiligkou et al., 2017). Analysis of the *HYR1-M1A* construct in *upf1*Δ cells revealed increased levels of the N-terminally extended protein and *HYR1* mRNA (Figure 2.6E, 2.6G, S2.5E), consistent with NMD targeting of the mutant transcript. Analyses in the *upf1*Δ 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 that removed the AUG start codon(s) (M1A for *CKB2;* M1A M20A for *FOL1*, see below) was not detected with *UPF1* present, but was in *upf1*Δ cells (Figure 2.6E, 2.6G).

For the two examples that were robustly detected in a *WT* background, Ymr31 and Yml020w, little increase in protein levels from *M1A* constructs in *upf1*Δ cells was seen for the extended versions (Figure 2.4C, 2.5F). Consistently, *YMR31-M1A* and *YML020W-M1A* mRNA levels were not dramatically decreased in WT cells relative to unmutated constructs (Figure 2.6G). 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 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 (Figure S2.7A-C), although this distance is thought to be a key factor in specifying yeast NMD substrates (Hug et al., 2016). We did, however, observe a moderately positive association between the distance of the transcript start site to the location of the first downstream ATG (which is out-of-frame) in the *M1A* constructs and the degree of NMD (Figure S2.7B).

2.2.9 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 reveals 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 2.6H). The relative usage of these start sites, 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 2.6I). 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 (Figure 2.6H, 2.6I, S2.5F). Such coding complexity is surprising to find in a eukaryote as simple as budding yeast and would not have been readily identifiable without TIS-profiling data.

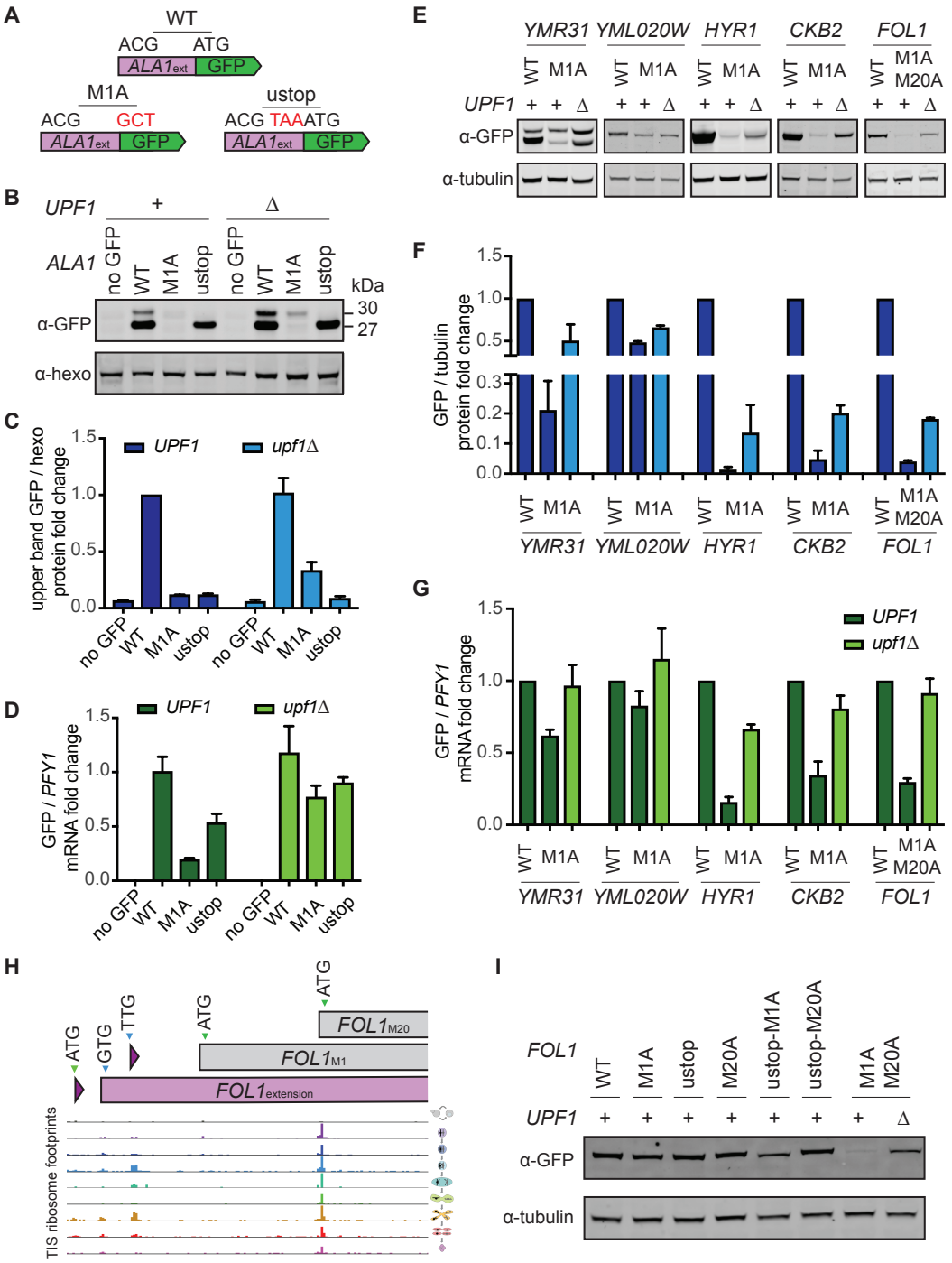


Figure 2.6 Extended ORF transcripts with no in-frame ATG are degraded by NMD
(A) Schematic for ALA1 tagging strategy, using a reporter including the region upstream of the ATG, and either including (WT) or not including (M1A) the in-frame ATG in front of the GFP, and a mutant with a stop codon upstream of the in-frame ATG (ustop).
(B) Western blot for Ala1GFP reporters in WT and upf1Δ 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 WT cells. In a upf1Δ background, the M1A construct shows the extension due to blocking nonsense mediated decay (NMD) of this transcript with no in-frame ATG.

(C) Western blot quantification of Ala1-GFP upper band intensity from Figure 2.6B 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 upf1Δ 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 upf1Δ cells at 4.5 hours in meiosis.
(F) Western blot quantification of GFP tagged proteins from Figure 2.6E normalized to tubulin for 3 replicates.
(G) qPCR fold change of GFP transcripts relative to PFY1 for 3 replicates from strains from Figure 2.6E.
(H) TIS-profiling of FOL1, showing ribosome footprints at the time points indicated in Figure 2.1B, with the positions of the extension (GTG), M1 (ATG) and M20 (ATG) start 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 possible in-frame start codons.

2.2.10 eIF5A levels alter non-AUG TIS usage in yeast meiosis

The preferential translation of non-AUG-initiated ORFs in meiotic cells (Figure 2.3C), and the increase in TIS-profiling reads in 5' leader regions in meiotic time points relative to vegetative cells suggests condition-specific modulation of translation initiation (Figure 2.7A). 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 (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 2.7B). 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 S2.8). 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 extension, 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 start site 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* (Figure 2.7C, S2.5G, S2.5H), 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 lead to lower Hyp2 activity (Figure S2.8). Moreover, our data suggests that meiotic ribosomal subunits show changes in association of multiple translation initiation factors relative to vegetative cells, some of which are known to be involved in TIS selection (Figure 2.7B; 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.

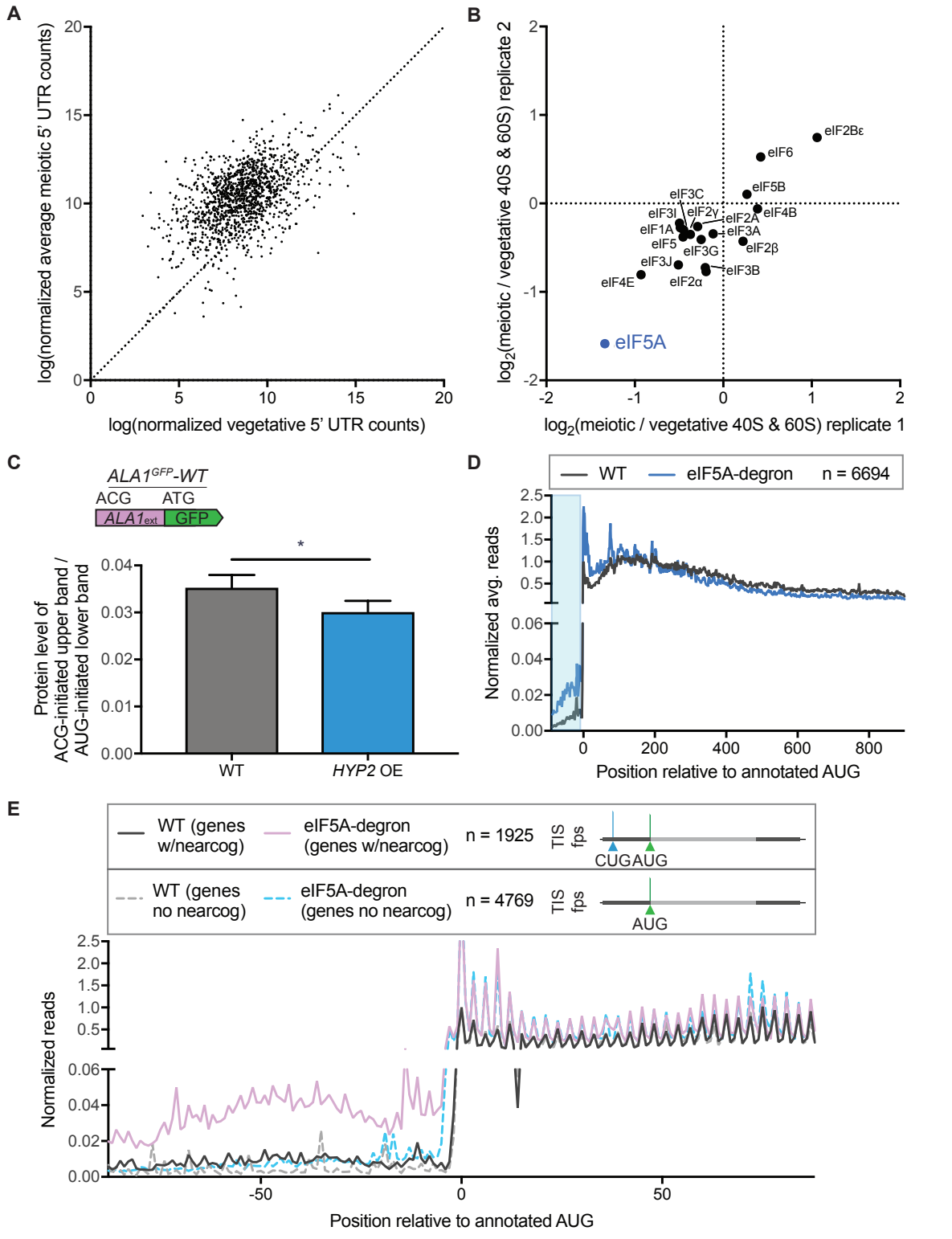


Figure 2.7 eIF5A levels regulate pervasive non-AUG-initiated translation
(A) Comparison of vegetative and average meiotic 5' read density measurements.
(B) Enrichment of translation factors comparing meiotic and vegetative samples for two replicates, determined by quantitative mass spectrometry of 40S and 60S ribosomal subunits isolated by sucrose gradient centrifugation of cell extract from meiotic and vegetative cells.

(C) Western blot quantification of ALA1GFP-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 is normalized to AUG-initiated GFP lower band, which runs as a doublet. Both bands were used for quantification. The decrease seen in 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 = 6694). Reads are normalized to WT at position zero and averaged across three nucleotides. Ribosome profiling data was 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 near-cognate-initiated ORFs annotated by ORF-RATER (n=1925, WT (genes w/nearcog): solid black and eIF5A-degron (genes w/nearcog): solid purple) and genes that do not contain near-cognate ORFs (n=4769, 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 with near-cognate start codons.

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 2.7D; 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=1925) 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=4769), relative to WT cells (Figure 2.7E). This shows that low eIF5A levels alone 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.

2.3 Discussion

Here we report the first method for globally mapping translation initiation sites, 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 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 Figure 2.4C 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 S2.6C). 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 three 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 genetically, by mutating the upstream near-cognate codon to an ATG, or by using a strong promoter to increase production of the extended protein, by necessity (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 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, 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 to 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 1-10% that of AUG initiation based on *in vitro* experiments (Chen et al., 2008; Kearse and Wilusz, 2017; Kolitz et al., 2008), and a model in which many fewer 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 site (*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 (Figure S2.9A, S2.9B). Finally, the data for genes in which the canonical AUG start is mutated (M1A, Figure 2.4C, 2.5E, 2.5F, 2.6B and 2.6E) supports both isoforms being translated from the same pool of transcripts. Otherwise, we would not expect AUG 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 *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 *in vitro* studies of near-cognate translation initiation (Chen et al., 2008; Diaz de Arce et al., 2018; Kolitz et al., 2008). 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 sites 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 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 S2.9C). 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 S2.9C), 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 of its role in meiotic cells. Inducing higher levels of eIF5A decreased translation of a reporter for near-cognate-initiated translation, and reanalysis 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., 2011;

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/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 S2.9D), leaving investigation of their function (or range of functions) an important area of future study. It remains unclear whether most 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 general role for 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 site selection during translation initiation, points to a simple strategy for cells to modulate the features of a subset of the proteome in response to a change in condition.

2.4 Materials and Methods

2.4.1 Yeast strain construction

All yeast strains used were Saccharomyces cerevisiae of the SK1 background. Strains used in this study are listed below.

BrÜn Strain No.	Genotype
13	MATa wild-type
14	MATalpha wild-type
15	MATa/alpha wild-type
1362	MATa/alpha wild-type
5805	MATa/a wild-type
12507	MATa/alpha ymr31::KanMX; trp1::YMR31-WT-yEGFP::TRP1
12508	MATa/alpha ymr31::KanMX; trp1::YMR31-M1A-yEGFP::TRP1
12509	MATa/alpha ymr31::KanMX; trp1::YMR31-ustop-yEGFP::TRP1
12510	MATa/alpha hyr1::KanMX; trp1::HYR1-WT-yEGFP::TRP1
12511	MATa/alpha hyr1::KanMX; trp1::HYR1-ustop-yEGFP::TRP1
12880	MATa/alpha hyr1::KanMX; trp1::HYR1-M1A-yEGFP::TRP1
16920	MATalpha trp1::ALA1-yEGFP_WT::TRP1
16922	MATalpha trp1::ALA1-yEGFP_M1A::TRP1
18006	MATa/alpha ymr31::KanMX; trp1::YMR31-upM-M1A-yEGFP::TRP1
18039	MATa/alpha ymr31::KanMX; trp1::YMR31-upM-yEGFP::TRP1
18547	MATalpha trp1::ALA1-yEGFP_ustop::TRP1

18766	MATa upf1::NatMX
19023	MATa trp1::ALA1-yEGFP_WT::TRP1; upf1::NatMX
19025	MATa trp1::ALA1-yEGFP_M1A::TRP1; upf1::NatMX
19033	MATa trp1::ALA1-yEGFP_ustop::TRP1; upf1::NatMX
19302	MATa/alpha ymr31::KanMX; trp1::YMR31-WT-yEGFP::TRP1; pdr5::HygMX
19303	MATa/alpha ymr31::KanMX; trp1::YMR31-M1A-yEGFP::TRP1
19430	MATa/alpha hyr1::KanMX; trp1::HYR1-M1A-yEGFP::TRP1; upf1::NatMX
20203	MATa/alpha ymr31::KanMX; trp1::YMR31-M1A-yEGFP::TRP1; upf1::NatMX
20858	MATa/alpha ymr31::KanMX; trp1::YMR31-M1A-yEGFP::TRP1; upf1::NatMX pdr5::HygMX
21423	MATa/alpha trp1::FOL1-WT-yEGFP::TRP1
21426	MATa/alpha trp1::CKB2-WT-yEGFP::TRP1
21640	MATa/alpha trp1::YML020W-WT-yEGFP::TRP1
21716	MATa/alpha trp1::FOL1-M20A-yEGFP::TRP1
21719	MATa/alpha trp1::CKB2-M1A-yEGFP::TRP1
21723	MATa/alpha trp1::YML020W-ustop-yEGFP::TRP1
21816	MATa/alpha trp1::CKB2-M1A-yEGFP::TRP1; upf1::NatMX
22159	MATa/alpha trp1::FOL1-M1A-yEGFP::TRP1
22353	MATa/alpha trp1::FOL1-M1A-M20A-yEGFP::TRP1
22434	MATa/alpha trp1::FOL1-M1A-M20A-yEGFP::TRP1; upf1::NatMX
22526	MATa/alpha trp1::YML020W-M1A-M31A-yEGFP::TRP1
22529	MATa/alpha trp1::YML020W-M1A-M31A-yEGFP::TRP1; upf1::NatMX
23156	MATa/alpha trp1::FOL1-ustop-yEGFP::TRP1
23157	MATa/alpha trp1::FOL1-ustop-M20A-yEGFP::TRP1
23955	MATa/alpha trp1::FOL1-ustop-M1A-yEGFP::TRP1
23983	MATa/alpha trp1::ALA1-yEGFP_WT::TRP1; hyp2::pCup1-HYP2::KanMX

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.

2.4.2 Yeast growth and sporulation

Vegetative cells were grown in YEPD, 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. Cells were washed once in water and resuspended in sporulation media (SPO). Time points were taken at times indicated in figures.

2.4.3 TIS-profiling

Cells were treated with 3 μM LTM (Millipore) for 20 min, 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 30C and spun down at 3000 rcf for 5 minutes at 4C. The supernatant was removed and cleared at 20,000 rcf for 10 minutes at 4C, and 200 uL 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.

2.4.4 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 ul 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 4oC. Tube was loaded on a Bio-Comp Gradient Station and analyzed for absorbance at 260 nm. For mass spectrometry of 40S/60S fraction, sucrose fraction was collected and flash frozen prior to precipitation and mass spectrometry.

2.4.5 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 h 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).

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

All mass spectra were analyzed with the Spectrum Mill software package v4.0 beta (Agilent Technologies) 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.

2.4.6 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 min, then boiled in SDS loading buffer for 5 min at 95C. Samples were spun down for 5 min at 20,000g prior to running on a 4-12% Bis-Tris gel for 175V 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 4C. 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 at RT, 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).

2.4.7 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 4C, then the aqueous phase was removed and added to chloroform. Samples were centrifuged again for 5 minutes at 21000 rcf at RT, then the aqueous phase was removed and added to isopropanol and 0.33 M NaOAc. Samples were precipitated at 4C overnight, then centrifuged for 20 min at 21000 rcf at 4C. Pellets were washed with 80% ethanol, air-dried, and resuspended in water. The TURBO DNA-free kit (Thermo) was used to treat 2.5 ug RNA with DNAse, then samples were incubated with random hexamers for 5 min at 65C. Superscript III (Thermo) buffer, DTT, dNTPs added, then superscript 25C 10min, 42C 50 min, 70C 10 min. 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).

2.4.8 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 Figure S2.2A-B.

2.4.9 Footprint quantification and correlation analysis

Standard RPKM calculations were used for cycloheximide profiling. For TIS-profiling, we counted reads mapping to the region spanning 3bp up- and downstream of the start codon and normalized by total reads at start 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 UTR quantification, read counts were determined for UTRs within the region from the canonical start to 99bp upstream. Counts were normalized by total reads at start sites to account for library size differences.

2.4.10 Start/stop codon analysis

The region 30-99bp 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 site 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 30bp 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 (Figure S2.4B, S2.4C).

2.4.11 Context analysis

Maximum motif score analysis was performed using Mochiview for the regions 10 basepairs up- and down-stream of all annotated genes, recapitulating the known Kozak sequence. The enrichment for this motif in regions 10 basepairs up- and down- stream of other start codon classes and control regions were plotted using the maximum motif score enrichment tool in Mochiview.

2.4.12 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.

2.4.13 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 extension) and proteins identified by ORF- RATER (an expanded set including scores 0.1 and above). 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.

2.4.14 Data and Code Availability

The datasets generated during this study are available at NCBI GEO, with accession number GSE150375.

2.5 Supplemental Figures

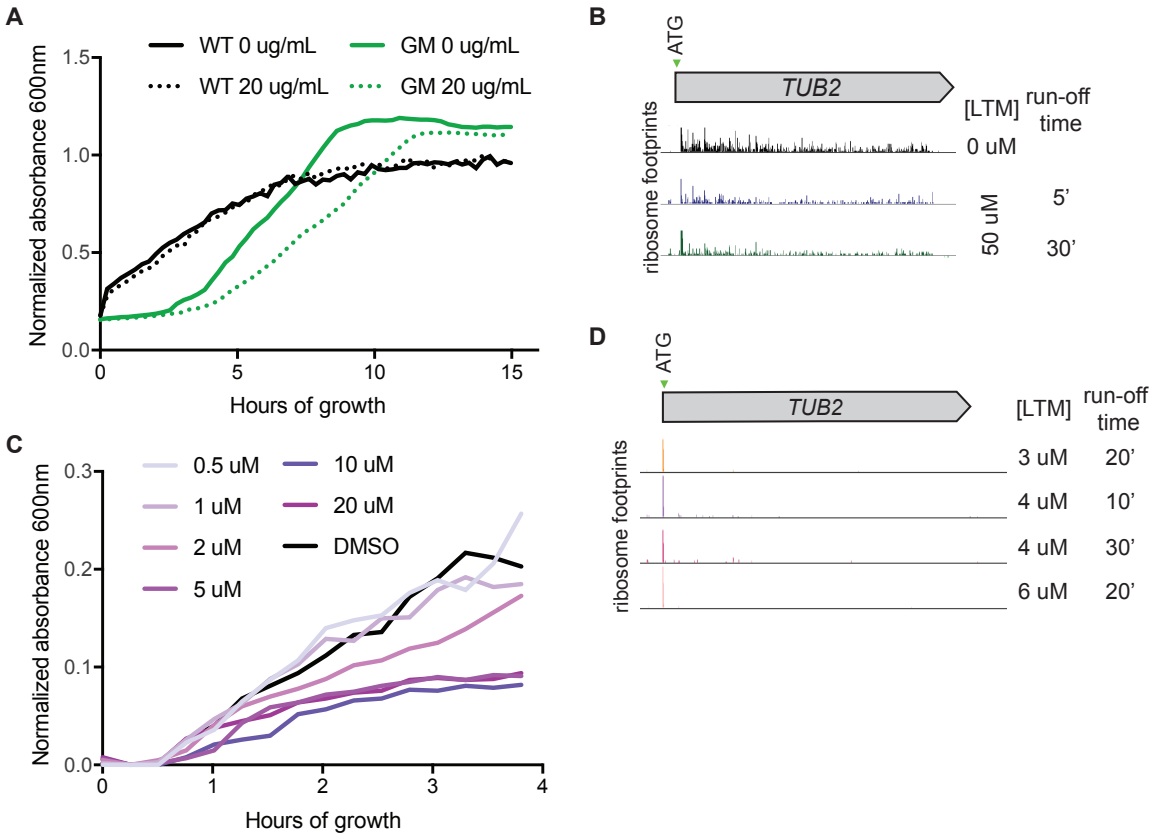


Figure S2.1 Optimization of TIS-profiling conditions for yeast

(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.

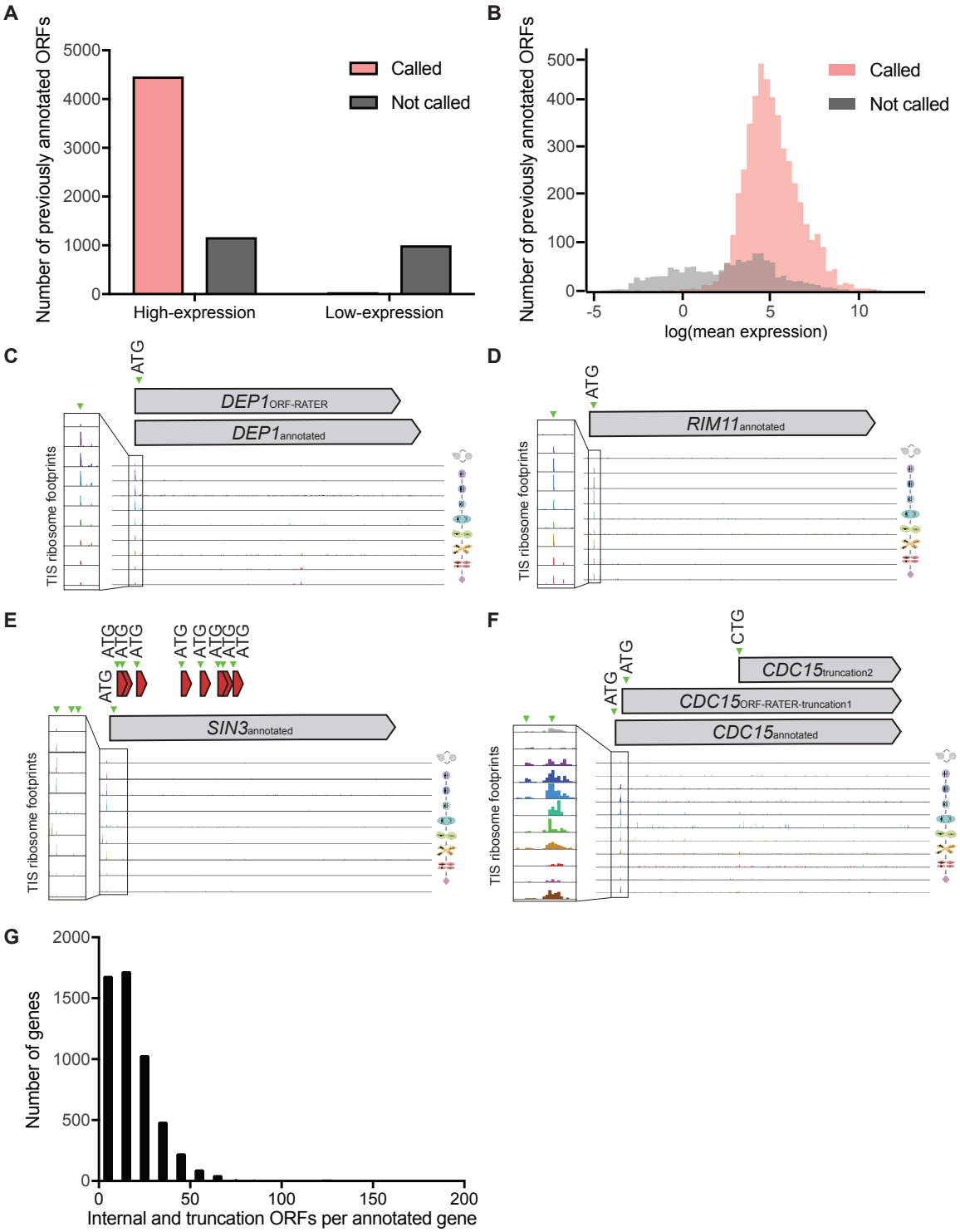


Figure S2.2 Categories of false positive and false negative ORF-RATER calls
(A) Previously annotated ORFs that are called (pink) or not called (gray), at expression values greater (high-expression) or less than (low-expression) 5 mean RPKM. Approximately half of annotated ORFs that were not called have 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.
(E) Number of internally initiated ORFs called per annotated gene.

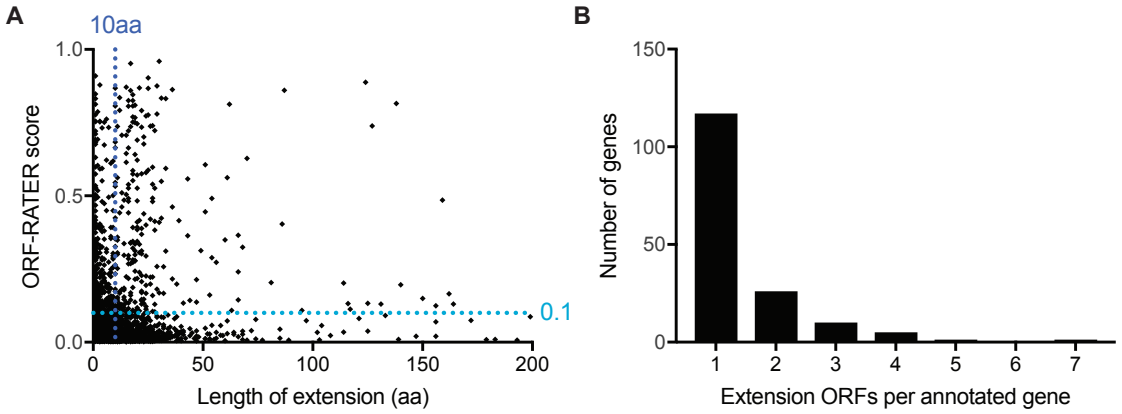


Figure S2.3 Properties of extension ORFs used for setting cutoffs
(A) Length versus score for all extension ORFs, with a line showing the length cutoff at 10 amino acids and the score cutoff of 0.1.
(B) Number of extension ORFs called per annotated gene.

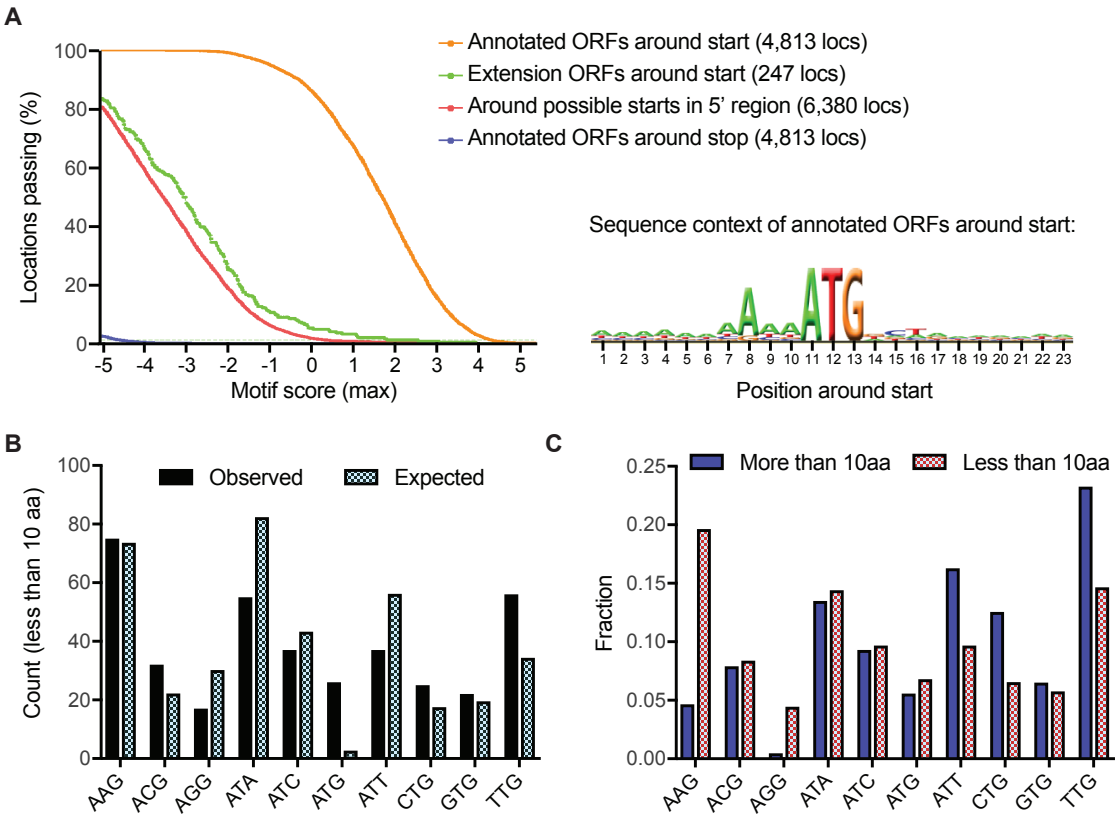


Figure S2.4 Translated near-cognate-initiated ORFs do not show Kozak sequence context enrichment

(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 2.3F).

(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.

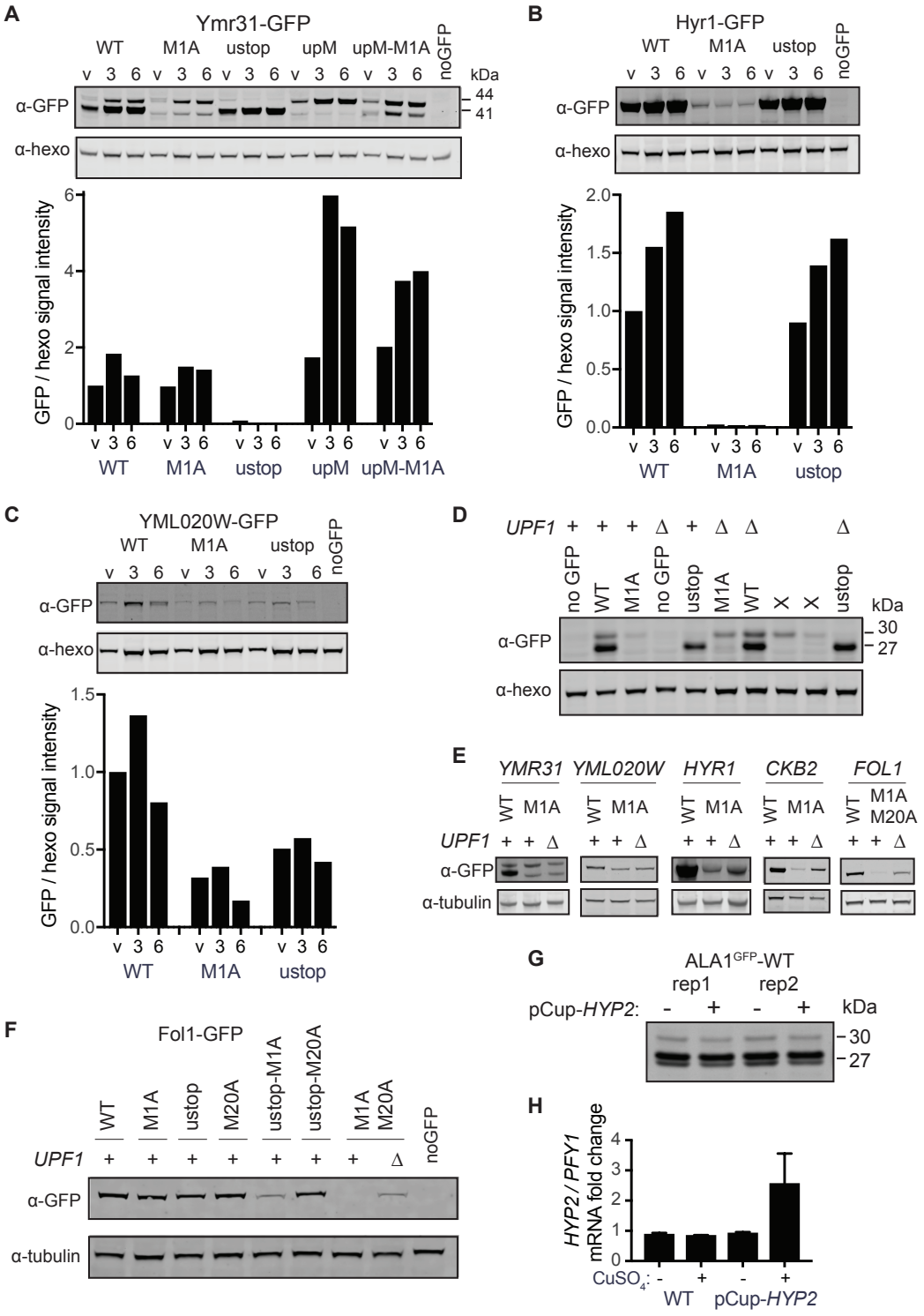


Figure S2.5 Western blot replicates and quantification for alternate isoforms
(A) Replicate western blot of YMR31-GFP constructs, as in Figure 2.4C (top) and quantification of upper GFP band relative to hexokinase loading control for three replicates (bottom).
(B) Replicate western blot of HYR1-GFP replicates, as in Figure 2.5E (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 2.5F (top) and quantification of GFP relative to hexokinase loading control for three replicates (bottom).

(D) Replicate western blot of ALA1GFP reporter constructs, as in Figure 2.6A. 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 upf1Δ, as in Figure 2.6E.
(F) Replicate western blot of FOL1-GFP constructs, as in Figure 2.6I.
(G) Western blot of ALA1GFP-WT reporter for cells with and without the pCup-HYP2 construct with copper (CuSO4) addition leading to overexpression of eIF5A for two replicates, which is quantified in Figure 2.7C.
(H) qPCR fold change of HYP2 transcript relative to PFY1 for cells with and without the pCup-HYP2 construct with and without copper (CuSO4) addition for three replicates.

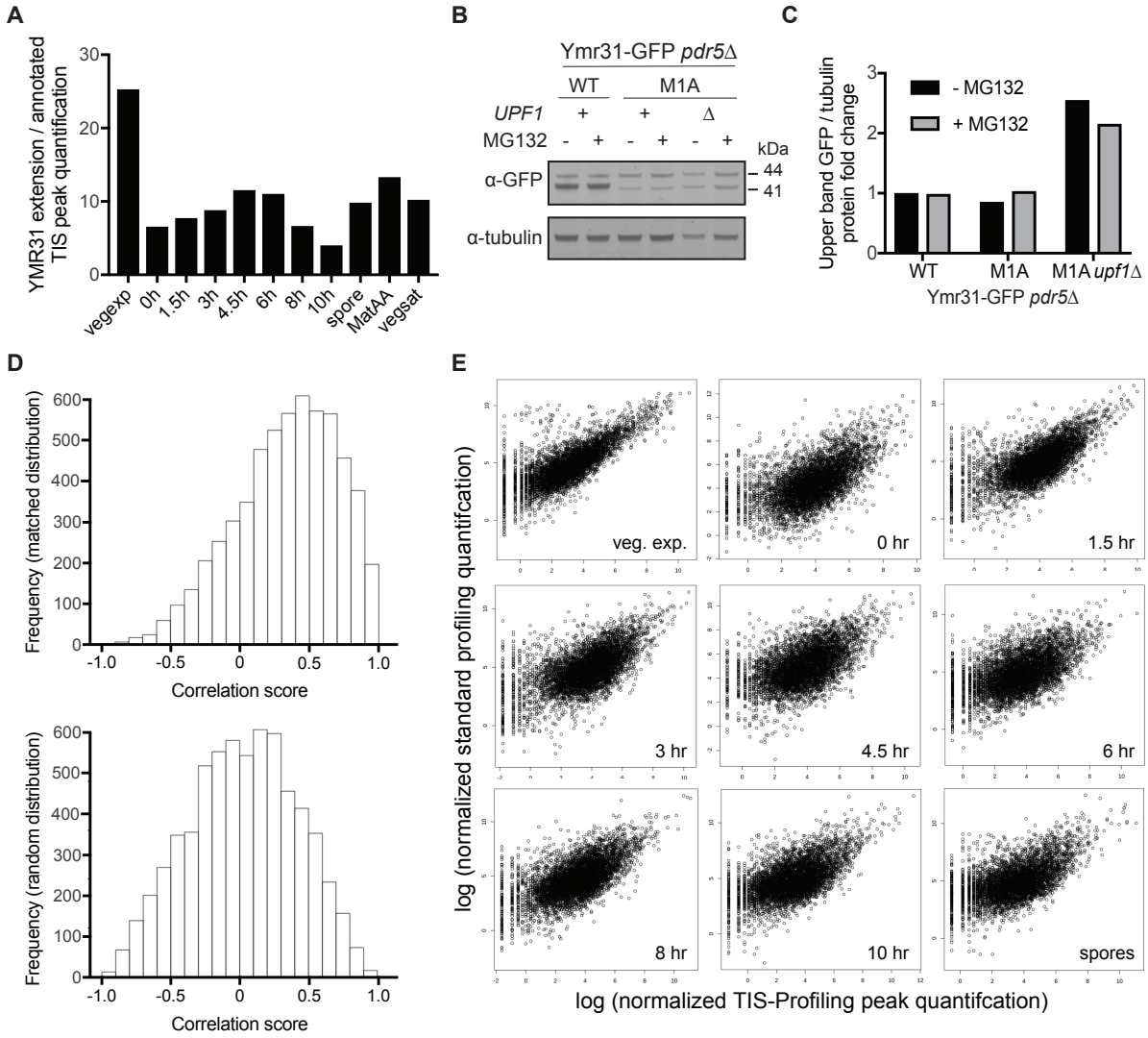


Figure S2.6 Positive correlation of TIS peaks with gene expression for annotated AUG sites but not near-cognate sites
(A) Quantification of YMR31 TIS-profiling peaks for the extension peak relative to the annotated peak. For all timepoints, the non-AUG extension peak is higher than the annotated AUG 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 2.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 <2.2x10-16.
(E) Scatter plots comparing peak quantification of TIS versus standard profiling for each timepoint.

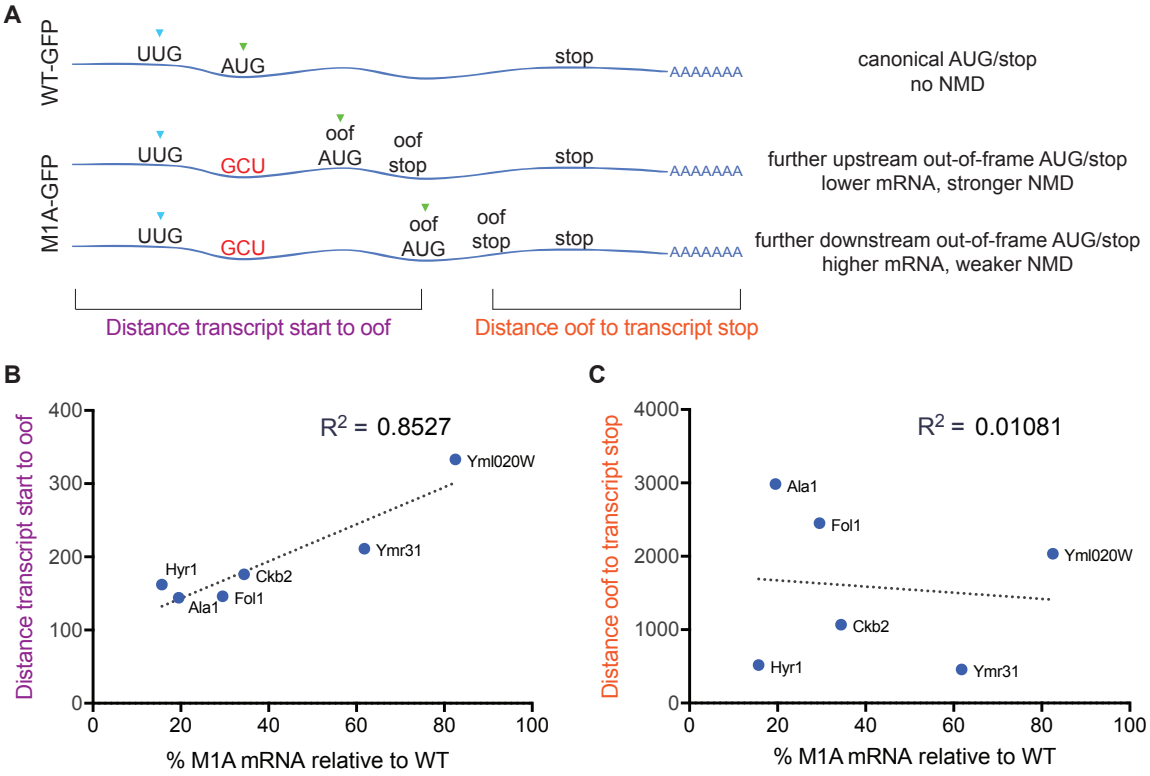


Figure S2.7 Effect of NMD for M1A transcripts does not correlate with distance from premature stop to transcript end
(A) Diagram of a canonical ORF (WT-GFP) compared to two possible M1A-GFP constructs where the annotated AUG is mutated, leading to initiation at a later, out-of-frame (oof) AUG. Two different 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 2.6G, where a lower percentage indicates a stronger NMD effect and a higher percentage indicates a weaker NMD effect. A correlation with an R2 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 2.6G. A correlation with an R2 value of 0.01081 is seen, indicating essentially no association between the distance from the oof ORF to transcript stop and the strength of NMD.

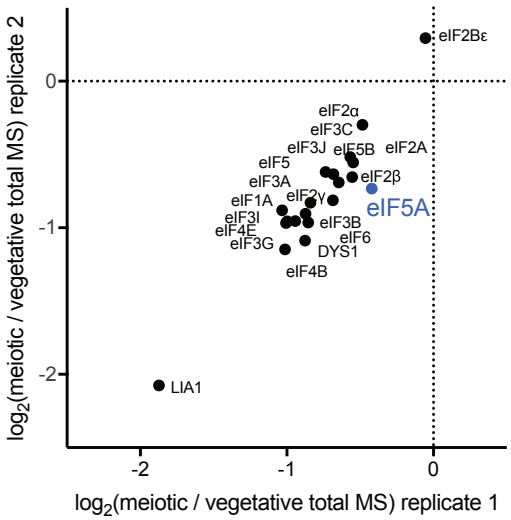


Figure S2.8 Total protein abundance of initiation and hypusination factors
Enrichment of translation factors (as in Figure 2.7B) and hypusination factors Lia1 and Dys1 comparing meiotic and vegetative samples for two replicates, determined by quantitative (TMT10) mass spectrometry of whole cell extract from meiotic and vegetative cells.

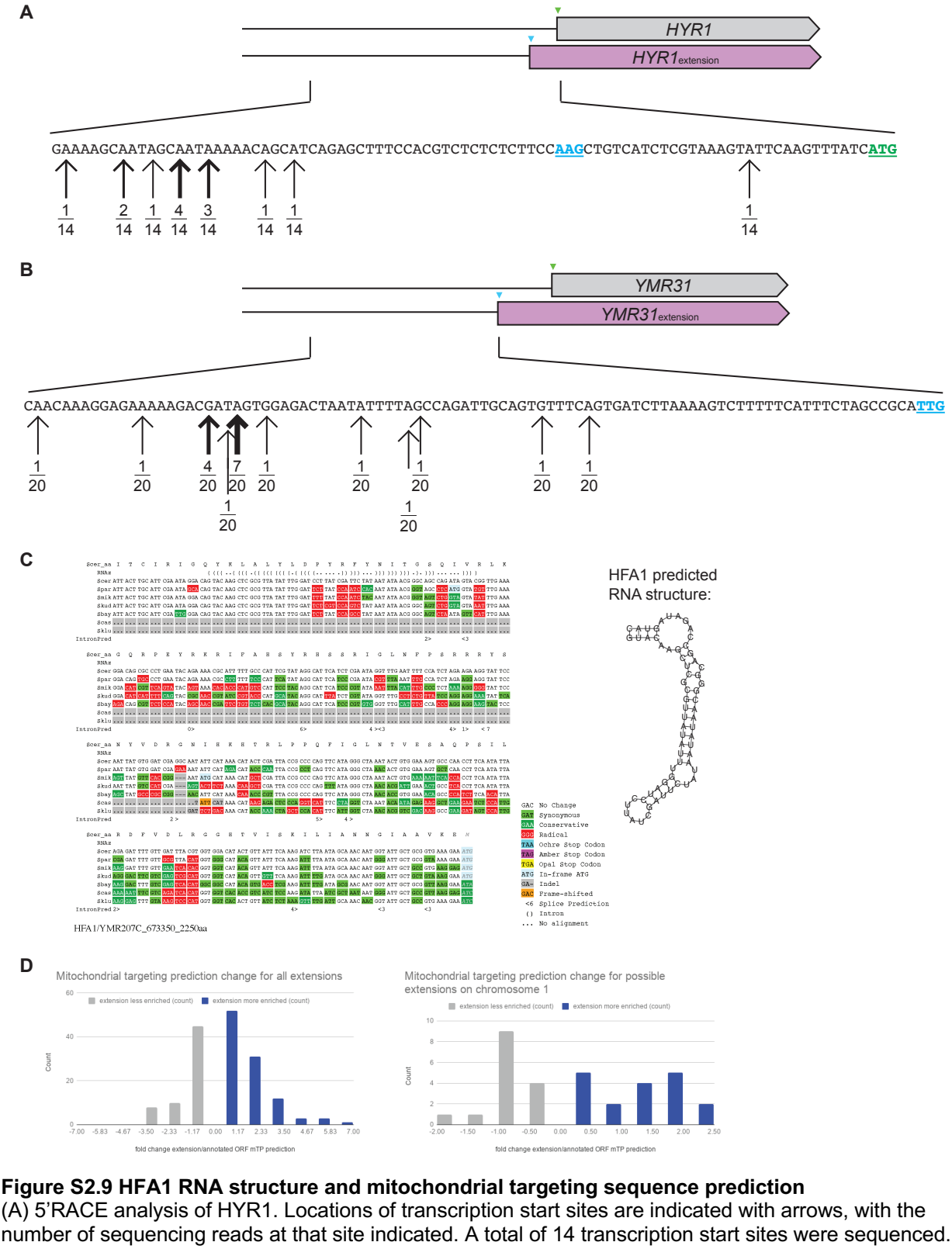


Figure S2.9 HFA1 RNA structure and mitochondrial targeting sequence prediction
(A) 5'RACE analysis of HYR1. Locations of transcription start sites are indicated with arrows, with the number of sequencing reads at that site indicated. A total of 14 transcription start sites were sequenced. (B) 5'RACE 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).

Chapter 3: Truncated protein isoforms generate diversity of protein localization and function in yeast

Portions of this chapter were adapted from the following manuscript:
Higdon, A.L., Won, N.H., Brar, G.A., 2023. Truncated protein isoforms generate diversity of protein localization and function in yeast. bioRxiv. https://doi.org/10.1101/2023.07.13.548938

3.1 Introduction

Defining the set of proteins encoded by an organism allows understanding of cellular function. This fundamental idea was the motivation for systematic coding sequence prediction immediately following the generation of whole genome sequences. Initial annotation strategies had known limitations, including difficulty in predicting splice isoforms and a reliance on strict rules to pare down the number of predicted open reading frames (ORFs), such as a minimum length requirement of 100 codons (reviewed in Dinger et al., 2008). Alternative splicing is generally considered to be the major force driving diversity of protein products from a single locus, but this process is relatively rare in budding yeast (although less rare than previously appreciated) (Ares et al., 1999; Douglass et al., 2019). Therefore, a simple model in which one gene encodes one transcript which is then decoded into one protein product has been adopted as a general rule in this organism, with a few isolated exceptions. Genomic techniques, such as transcript isoform sequencing and ribosome profiling, have greatly expanded our understanding of the diversity of transcript and protein products encoded by even very compact genomes like that of budding yeast (Brar et al., 2012; Chia et al., 2021; Eisenberg et al., 2020; Ingolia et al., 2011; Pelechano et al., 2013). Ribosome profiling, in particular, has facilitated the identification of a diverse array of translated open reading frames (ORFs), including upstream open reading frames (uORFs), intergenic short ORFs (sORFs), and N-terminally extended protein isoforms.

Meiosis in budding yeast is an excellent system for identifying fundamental principles of genome decoding and regulated gene expression. The process of differentiation from a diploid progenitor into haploid gametes requires an intricate and precisely timed series of cellular remodeling events (reviewed in Marston & Amon, 2004; van Werven & Amon, 2011). Underlying these dramatic changes is a gene expression program that requires dynamic regulation of most of the yeast proteome (Brar et al., 2012). Ribosome profiling and transcription start site sequencing show that production of non-canonical protein products and alternative transcripts is particularly prevalent during meiosis relative to mitotic growth (Brar et al., 2012; Chia et al., 2021; Sing et al., 2022). The specific functional relevance of all but a few non-canonical gene products, however, remains unclear. N-terminally truncated proteins, which initiate from downstream in-frame start codons within annotated genes, are a particularly interesting class. Since they are variants of existing proteins, they would seem likely to have molecular function, but their identification has remained difficult, both by sequencing-based and proteomics methods, due to specific technical challenges arising from their complete overlap with annotated coding regions.

Here, we develop a novel algorithm and identify hundreds of N-terminally truncated isoforms using data from a modified version of ribosome profiling for identifying translation initiation sites. In addition, we describe two distinct types of regulation responsible for their production, give experimental evidence for their existence, and provide insights into their functions. A handful of N-terminally truncated proteins have been validated in several organisms, including humans, and ribosome-profiling-based analyses suggest they might be common in mammalian cells (Ingolia et al., 2011). Although several individual examples of N-terminally truncated proteins were identified in previous single-gene studies in yeast (Table 3.1, 3.2), our genome-wide approach, and subsequent experimental validation, sheds new light on their prevalence in this heavily studied model organism. Most of the 388 truncations we identified are dynamically regulated during the meiotic program or under other non-standard laboratory growth conditions. The compactness of the yeast genome and minimal use of splicing have led to the presumption that yeast do not widely use alternative protein isoforms. In fact, these very features facilitated our robust global analyses and investigation of their production.

We classified truncations into two broad classes that generally differentiate regulatory and functional characteristics of the isoforms. The first class of truncations lack a large portion of the annotated protein's N-terminus ("distal" truncations) and tend to be encoded by a truncated transcript isoform. We identify and characterize such two examples, produced from the *PUS1* and *YAP5* loci, whose conserved and well-characterized annotated isoforms encode a pseudouridine synthetase and an AP-1 transcription factor, respectively. The truncated isoforms are expressed in a condition-specific manner and lack key domains, resulting in functions that seem distinct from the full-length proteins. The second class of truncations begin closer ("proximal") to the annotated start codon. Proximal truncations are more often encoded by the same transcript as the annotated isoform, likely requiring bypass of the annotated start codon for their translation and allowing simultaneous production of the annotated and truncated isoform from a single transcript. Based on our extensive computational and experimental investigation, we posit that a common role for these truncations is in diversifying the subcellular localization of the encoded protein. We demonstrate that our predictions in this respect are remarkably robust, revealing a case in which two truncations at one locus allow three distinct and simultaneous subcellular localizations. Thus, our study elucidates the potential of truncated protein isoforms to provide proteome diversity and cellular function beyond what was previously recognized.

3.2 Results

3.2.1 Truncated protein isoforms are prevalent in budding yeast

N-terminally truncated protein isoforms initiate at in-frame start codons within canonical genes and therefore share common C-terminal sequence with their annotated isoform. These truncated isoforms are more challenging to identify than many other types of translated regions. Standard ribosome profiling data, which have enabled global maps

of translated regions, report the positions of elongating ribosomes and therefore yield reads mapping across the entire gene, including the entirety of the potential truncated isoform, masking signal for truncated isoforms (Figure 3.1A). We previously published a translation-initiation site (TIS-) profiling dataset with timepoints spanning stages of meiotic progression in budding yeast cells with the goal of identifying all types of non-canonical start sites, with a particular focus on N-terminally extended protein isoforms (Eisenberg et al., 2020). In contrast to standard ribosome profiling, which employs the translation elongation inhibitor cycloheximide (CHX), TIS-profiling uses the drug lactimidomycin (LTM) to capture ribosomes immediately after initiation, while allowing elongating ribosomes to complete translation (or "run off"; Figure 3.1A) (Ingolia et al., 2011; Lee et al., 2012). The resulting ribosome footprints are therefore highly enriched at sites of translation initiation, with minimal background reads across the ORF. TIS-profiling therefore yields clear start site peaks that are much easier to detect than with standard ribosome profiling. For example, at the *MOD5* locus, we detect both a previously identified truncated isoform ($Mod5^{truncation1}$) and an additional previously unidentified truncated isoform ($Mod5^{truncation2}$) using TIS-profiling (Figure 3.1B). Notably, neither truncated isoform is evident from standard ribosome profiling data, demonstrating the power of this method for detecting internal translation start sites.

We previously used the program ORF-RATER to call all types of open reading frames, including those that are non-canonical, using standard ribosome profiling and TIS-profiling data collected in vegetative (mitotic exponential and saturated) conditions and 8 timepoints spanning the major developmental stages of meiosis (Eisenberg et al., 2020; Fields et al., 2015). Although this algorithm performed very well for many types of ORFs, including those that were 5' extended, it was unable to identify many truncated isoforms that were extremely clear in the start-site profiling data, such as at the *YAP5* locus (ATG^2, Figure 3.1C; Fields et al., 2015). We hypothesized that this could be due to overweighting of the standard ribosome profiling and underweighting of the TIS-profiling data by the algorithm. It has also been observed that many ribosome profiling algorithms that are trained on annotated ORFs are unable to perform as well with shorter ORFs (Spealman et al., 2021). To avoid systematically biasing against sensitive detection of translation initiation sites within annotated ORFs, we developed a novel algorithm specifically designed for identifying truncated isoforms that relies solely on the TIS-profiling data.

Briefly, the goal of this algorithm is to robustly interpret TIS-profiling data by separating true start-site signal from the background noise present across translated genes, likely due to low levels of elongation inhibition by LTM. To model the background signal for each gene at each timepoint, we randomly sampled three single nucleotide positions from within the annotated open reading frame and summed their mapped reads to simulate the three nucleotides of a start codon, over which a prominent peak should be present for real translation initiation sites. Resampling 10,000 times provided an empirical null distribution representing the distribution of peak heights that would be expected from the background noise within each gene at each time point (as shown for the *YAP5* locus at 0h in SPO; Figure 3.1D). We next determined the observed peak height for all in-frame (potentially truncation-generating) start codons within each

annotated gene and used the corresponding empirical distribution to assign a p-value to each putative truncation start site. In the case of *YAP5*, there were three in-frame start codons within the annotated ORF, two of which fall within the range of peak heights expect from background and are not called (ATG^3 and ATG^4), and one which was called as significantly above background (ATG^2; Figure 3.1C-D). To increase the stringency of our calls, we required that a translation initiation site be called at 2 or more timepoints. We also required that a truncated isoform begin 5 amino acids (aa) or more from the annotated start codon and be no less than 10aa in total length. Truncated isoforms for which we could not detect the annotated isoform at any time point were also excluded (n = 25), as these often represent cases of misannotation, where the "truncated" isoform is in fact the main isoform. Additional filtering criteria were applied, as described in the methods.

Using this approach, we identified 388 truncated protein isoforms. While the existence of truncated protein isoforms has been known for decades due to observations from single-gene studies, this represents a substantial increase in the number of known cases. To our knowledge, in *S. cerevisiae*, 12 truncated isoforms (two of which are present at the same locus, *CCA1*) have been characterized that we would expect to be present in our dataset as well (Table 3.1, 3.2). Of these previously known truncations, 10 (83.33%) were called by our algorithm. Of the two that were not called, one was called by the algorithm but filtered out because its annotated isoform was not called (at the *VAS1* locus). The other (the shortest isoform of *CCA1)* was visible by eye in the TIS-profiling data but was not called by our algorithm. We concluded that our approach was sensitive to identifying translation of truncated isoforms and that they are much more prevalent than previously known, with our set of newly identified truncations representing a more than 30-fold increase over the set of characterized truncations.

Gene	**Function**	**Called in TIS-profiling?**	**Citation(s)**
CCA1 (YER168C)	tRNA CCA addition	Yes, No*	(Wolfe et al., 1994)
CRS1 (YNL247W)	Cysteinyl-tRNA synthetase	Yes	(Nishimura et al., 2019)
FUM1 (YPL262W)	Fumarase	Yes	(Wu and Tzagoloff, 1987)
GAT1 (YFL021W)	Transcription factor	Yes	(Rai et al., 2014)
GLR1 (YPL091W)	Glutathione oxidoreductase	Yes	(Outten and Culotta, 2004)
GRX2 (YDR513W)	Glutaredoxin	Yes	(Pedrajas et al., 2002; Porras et al., 2006)
KAR4 (YCL055W)	Mating TF	Yes	(Gammie et al., 1999)
LEU4 (YNL104C)	Leucine biosynthesis	Yes	(Beltzer et al., 1988, 1986)
MOD5 (YOR274W)	tRNA modification	Yes	(Boguta et al., 1994)
SUC2 (YIL162W)	Invertase	Yes	(Carlson and Botstein, 1982; Taussig and Carlson, 1983)
VAS1	Valyl-tRNA synthetase	No**	(Chatton et al., 1988)

(YGR094W)			

*Two truncated isoforms, one called, and one not called but visible by-eye in genome browser
**Truncation would have been called but was filtered out because main ORF not called
Table 3.1 Previously characterized genes with N-terminally truncated protein isoforms.

Gene	Function	Reason for exclusion	Citation(s)
CCC1 (YLR220W)	Vacuolar transporter	Not expressed in WT cells	(Amaral et al., 2021)
HTS1 (YPR033C)	Histidine-tRNA synthetase	Second isoform is an AUG extension relative to genome annotation	(Natsoulis et al., 1986)
MRK1 (YDL079C)	Glycogen synthase kinase	Start is in intron; intronic sequences not considered in our algorithm	(Zhou et al., 2017)
TRM1 (YDR120C)	tRNA modification	Second isoform is an AUG extension relative to genome annotation	(Rose et al., 1992)

Table 3.2 Other relevant genes excluded from validation set

Since our algorithm used only TIS-profiling data to identify truncated isoforms, we were able to evaluate the quality of our calls using matched standard ribosome profiling data. In these data, we would expect each truncated isoform to have a peak in reads at the start site, followed by slightly elevated read density downstream relative to upstream, since downstream read density should include the contributions of elongating ribosomes associated with both the annotated and truncated isoforms (Figure 3.1A). We performed metagene analysis for the regions surrounding the predicted start codon for all 388 truncations and indeed saw the expected read density patterns across the gene set (Figure 3.1E, upper). Read densities downstream of the start codon (+20 to +70bp; excluding the non-quantitative region immediately following the start codon) were higher those upstream (-50bp to -1bp), consistent with elevated ribosome footprint density corresponding to the translation of truncated isoforms ($p<0.01$, Mann-Whitney U Test). Importantly, this trend was not seen for in-frame start codons that were not called by our algorithm (n.s., Mann-Whitney U Test; Figure 3.1E, lower). Metagene profiles of the TIS-profiling data also showed the expect trends, with a sharp peak present at the start codon for called truncations, and virtually no signal at the start codon for uncalled truncations (Figure S3.1A-B). Together this indicates that our truncation calling approach detects true translation events.

3.2.3 Truncated isoforms are dynamically expressed and enriched in meiosis

Among our set of called truncated isoforms, we observed that many appear to be dynamically expressed during meiosis. At the *MOD5* locus, for example, the smaller truncation ($Mod5^{trunc.2}$) is not present in exponentially growing mitotic cells but is upregulated specifically during early meiosis (Figure 3.1B). Over 50 percent of truncated isoforms in our dataset were called at only two or three timepoints, indicating that dynamic expression is very common (Figure 3.1F). Truncated isoforms were also much more common in meiosis and slightly more common in vegetative saturated growth than in vegetative exponential growth, the most common laboratory growth condition (Figure 3.1G). Since we analyzed multiple samples collected during meiotic progression but

only one each during vegetative exponential growth and vegetative saturated growth, this enrichment could have been due to increased power to detect ORFs during meiosis. However, this pattern was significantly less strong among annotated protein isoforms called by our algorithm, indicating that dynamic meiotic expression of truncated protein isoforms is a true biological phenomenon ($p<2.2e^{-16}$, Fisher's Exact Test).

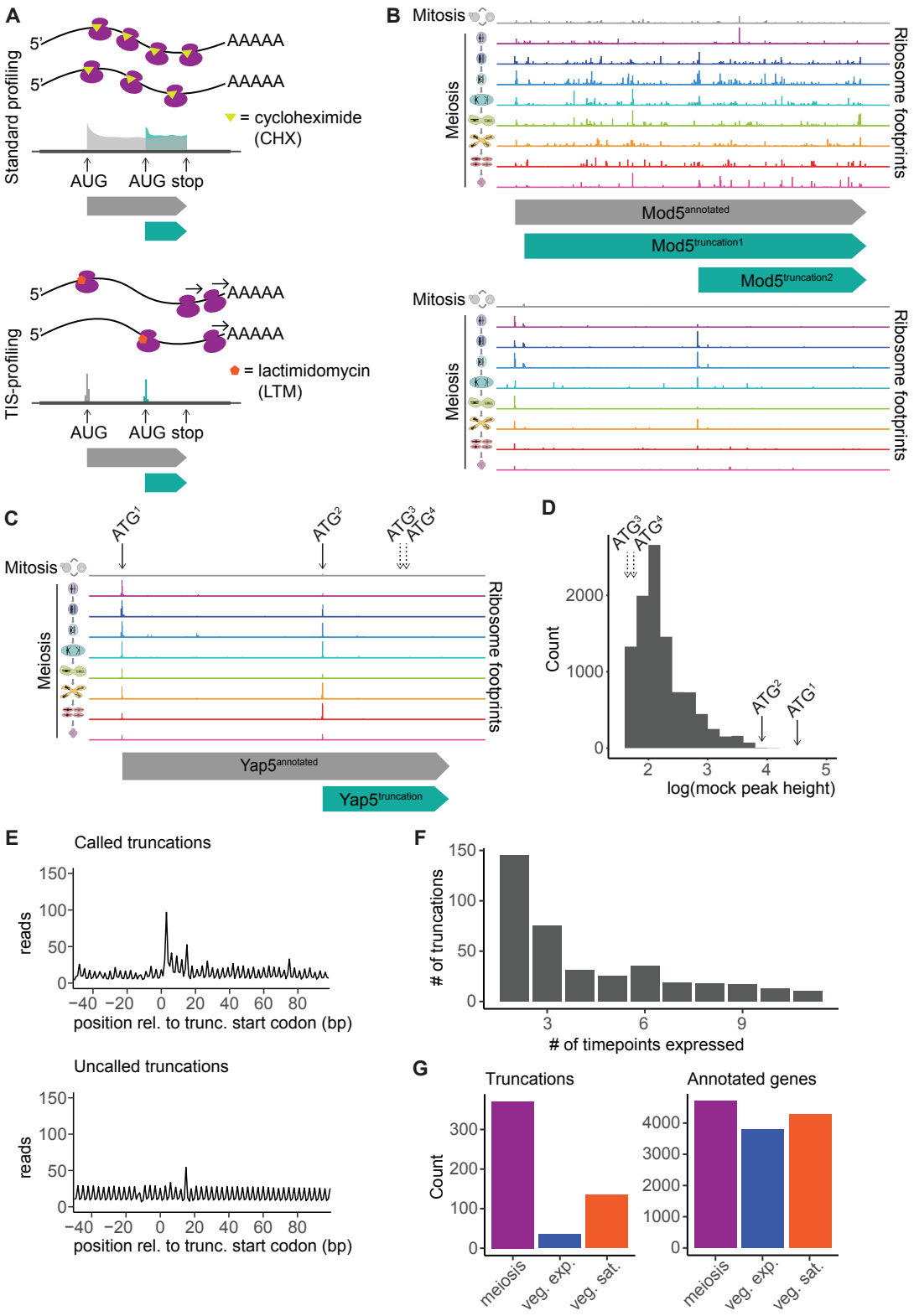


Figure 3.1 Genome-wide identification of truncated protein isoforms using TIS-profiling data
(A) Schematic comparing standard ribosome profiling (top) and TIS-profiling (bottom). In each case, a cartoon of ribosomes translating two mRNAs is shown on top and sample read density is shown below for a locus containing an annotated and truncated protein isoform.

(B) Standard ribosome profiling (top) and TIS-profiling (bottom) data for the *MOD5* locus, with the annotated open reading frame in gray and the two truncations identified by our algorithm in turquoise (middle). Cartoons to the left indicate whether the track represents mitotic cells (gray) or meiotic timepoints (rainbow). Note that the two truncations are not apparent from the standard ribosome profiling data but robust peaks at multiple timepoints are clear in the TIS-profiling data.
(C) TIS-profiling data at the *YAP5* locus. Arrows indicate all in-frame ATGs within the ORF. Cartoons to the left indicate whether the track represents mitotic cells (gray) or meiotic timepoints (rainbow).
(D) Empirical distribution representing background read density within the *YAP5* annotated ORF at a single TIS-profiling time point (0h). Arrows indicate approximate read density for called start sites (solid arrows, ATG^1 and ATG^2) and uncalled in-frame start codons (dashed arrows, ATG^3 and ATG^4).
(E) Metagene plot of standard ribosome profiling data for all called truncated isoforms (top) and uncalled controls (bottom) for the region between -50 and +100nt relative to the truncation start codon. Reads are summed across all timepoints. For called truncations, downstream (+20 to +70bp) read density is significantly higher than upstream (-50 to -1bp) read density ($p<0.01$, Mann-Whitney U Test). For uncalled truncations there was not a significant difference in upstream and downstream read densities (n.s., Mann-Whitney U Test).
(F) Histogram of number of timepoints at which truncated isoforms are expressed in TIS-profiling data.
(G) Number of truncated isoforms (left) and number of annotated isoforms (right) called in samples from meiotic timepoints, vegetative exponential growth, and vegetative saturated growth. The variation in the number of called isoforms between conditions is significantly more pronounced for truncated isoforms than for annotated isoforms ($p<2.2e^{-16}$, Fisher's Exact Test).

3.2.4 Newly predicted truncated protein isoforms can be detected *in vivo*

To evaluate the quality of our TIS-profiling data and truncation-calling algorithm, we experimentally validated the production of several newly identified truncated proteins. Due to the limitations in resolving similarly sized proteins, we focused on candidates that differed enough in size to be distinguishable from their annotated isoform by western blotting. We integrated a C-terminal epitope-tag at the endogenous locus of each protein, such that both the annotated and truncated isoforms should be tagged. We then collected samples at timepoints throughout meiosis and performed western blotting. Ten predicted truncated isoforms that we examined were clearly detectable by this method, indicating that the truncated proteins are expressed and stable. Some of these truncated protein isoforms display dynamic expression patterns distinct from their annotated isoform, including $Yap5^{trunc.}$, $Mod5^{trunc.}$, $Pex32^{trunc.}$, $Pus1^{trunc.}$, $Sas4^{trunc.}$, $Ssp1^{trunc.2}$, $Tpo1^{trunc.}$, and $Prp4^{trunc.}$ (Figure 3.2A-F, S3.2A-B). Others, like $Ssp1^{trunc.1}$ and $Yck1^{trunc.1}$, display expression patterns that mirror the annotated isoform (Figure 3.2F-G).

Although the detected truncated isoforms migrate according to their expected size, it remained possible that the bands could be degradation products of the annotated isoform rather than the product of translation at the predicted alternative in-frame start codon. To test this possibility, we generated strains for two examples, Pus1 and Yap5, for which the ATG start codons for the annotated or predicted truncated isoform were mutated to ATT (isoleucine) to abrogate expression ($M^{ann.}L$ and $M^{trunc.}L$, respectively; Figure 3.2H). Western blot analysis of these strains revealed translation of only the annotated isoform in cells carrying the $M^{trunc.}L$ mutation, and only the truncated isoform in cells carrying the $M^{ann.}L$ mutation, indicating that the truncated isoforms for each gene are indeed the product of translation initiating at the newly predicted start codon internal to the annotated ORF (Figure 3.2I-J).

Our ability to detect most of the predicted truncated isoforms that we tested suggest our approach has a low rate of false positives. However, 5 out of 15 predicted truncations that we tested were not visible by western blotting (Yck1$^{trunc.2}$, Glk1$^{trunc.}$, Ari1$^{trunc.}$, Rtt105$^{trunc.}$, and Siw14$^{trunc.}$; Figure 3.2G, S3.2C-F). There are several possible explanations for this: (1) the truncated protein may not be compatible with the specific epitope tag, (2) they may be challenging to detect for technical reasons, for example due to low expression or small size, (3) they could be false positives, or (4) they may be produced but then degraded under normal conditions. To test whether protein degradation was preventing our detection of some truncated protein isoforms, we treated cells carrying C-terminal epitope tags of predicted truncations with the proteasome inhibitor MG132. We timed MG132 treatment and sample collection for western blotting according to the expected timing of expression for each truncation based on the TIS-profiling data. We included examples in which the truncation was not detected (Yck1$^{trunc.2}$, Ari1$^{trunc.}$, Rtt105$^{trunc.}$, and Siw14$^{trunc.}$) as well as ones in which the truncation was visible but present at low abundance (Sas4$^{trunc.}$, Tpo1$^{trunc.}$, Mod5$^{trunc.}$, and Prp4$^{trunc.}$) to look for evidence of stabilization. We indeed saw increased abundance, indicating increased stabilization, for previously detectable truncations, including Sas4$^{trunc.}$ and Tpo1$^{trunc.}$ (Figure 3.2K, S3.2G). Interestingly, the previously undetectable second truncation of Yck1$^{trunc.2}$ became visible upon proteasome inhibition, suggesting that this truncation is normally translated but degraded by the proteasome to levels below detection (Figure 3.2L). Abundance of other truncated proteins, including Mod5$^{trunc.}$ and Prp4$^{trunc.}$ (Figure S3.2H-L), was minimally affected by proteasome inhibition, suggesting that proteasome-mediated degradation was not the reason for their weak detection by western blotting. Three of the previously undetectable truncations (Ari1$^{trunc.}$, Rtt105$^{trunc.}$, and Siw14$^{trunc.}$) remained undetectable upon proteasome inhibition (Figure S3.2J-L).

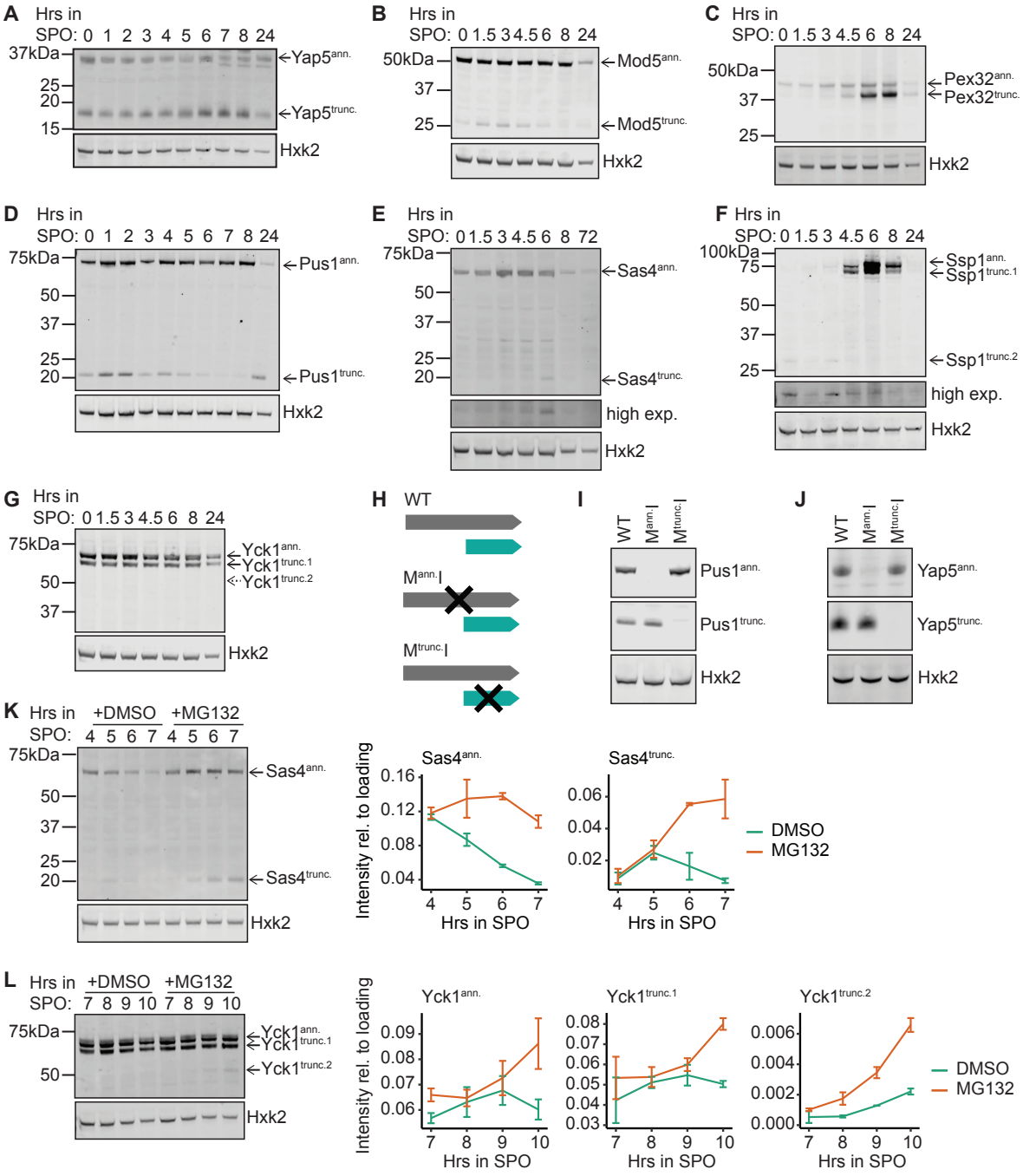


Figure 3.2 Many newly identified truncated protein isoforms can be confirmed by western blotting, some are stabilized by proteasome inhibition

(A) Western blot of samples collected at various timepoints after transfer of cells to sporulation media (SPO). Hexokinase (Hxk2) is shown as a loading control. For lowly expressed truncated isoforms, a high exposure panel is shown (high exp.). A different strain is used in each case, expressing the indicated C-terminal epitope-tagged protein and enabling detection of annotated (ann.) and truncated (trunc.) isoforms for (A) Yap5-FLAG,
(B) Mod5-3V5,
(C) Pex32-3V5,
(D) Pus1-3V5,
(E) Sas4-3V5,

(F) Ssp1-3V5, for with two truncated isoforms were predicted and detected,
(G) and Yck1-3V5, for which two truncated isoforms were predicted but only one was successfully detected. Solid arrows indicate isoforms that were predicted and detected. Dashed arrows indicate predicted but undetected isoforms.
(H) Schematic of mutagenesis approach used to validate production of truncated isoforms from predicted start codons in (I) and (J). Top image indicates the wild-type context (WT), in which both isoforms should be seen. Either the annotated (middle, $M^{ann.}I$) or predicted truncation (bottom, $M^{trunc.}I$) start codon were mutated from encoding methionine (ATG) to isoleucine (ATT) to prevent their translation.
(I) Western blot of start codon mutant strains described in (H). Hexokinase (Hxk2) is shown as a loading control. Images of bands for the annotated and truncation isoforms are from the same blot. Blotting is for Pus1-3V5. Cells were collected at 24h post-dilution in YEPD.
(J) Western blot of start codon mutant strains described in (H) for Yap5. Hexokinase (Hxk2) is shown as a loading control. Images of bands for the annotated and truncation isoforms are from the same blot. Blotting is for Yap5-FLAG. Cells were collected at 4.5h in SPO.
(K) Representative western blot (left) and quantification (right) for cells treated with proteasome inhibitor MG132 or vehicle control DMSO, showing stabilization of both the truncated and annotated isoforms of Sas4-3V5 for cells in meiosis. Note that quantification is based on 2 replicates and error bars represent standard error. Hexokinase (Hxk2) is shown as a loading control.
(L) Representative western blot (left) and quantification (right) for cells treated with proteasome inhibitor MG132 or vehicle control DMSO, showing stabilization of the truncated and annotated isoforms of Yck1-3V5 for cells in meiosis. Note that quantification is based on 2 replicates and error bars represent standard error. Hexokinase (Hxk2) is shown as a loading control.

3.2.5 "Distal" truncations are typically produced from a truncated transcript while "proximal" truncations are likely regulated by translational control

Many truncated isoforms are dynamically regulated, some in concert with their annotated isoform and others independently. There are two straightforward models to explain such regulation: either (1) translational control, in which the ribosome bypasses one or more in-frame start codons in favor of a downstream one or (2) transcriptional control, in which a truncated transcript encodes the truncated protein and translation initiation occurs at the first in-frame start codon (Figure 3.3A). To determine the extent of these two possible types of regulation for the hundreds of new truncated proteins identified by our algorithm, we leveraged a published transcript leader sequencing (TL-seq) dataset collected across meiosis in the same strain background as our TIS-profiling data (SK1) (Chia et al., 2021). Using these data, we were able to determine whether there is evidence of a 5' transcript end upstream of our called truncation TISs but downstream of their respective annotated TISs. For *YCK1*, for example, two truncated isoforms ($Yck1^{truncation1\ and\ 2}$) are produced but there is no evidence of a separate transcript isoform in the TL-seq data, consistent with translational control (Figure 3.3B). For *YAP5*, in contrast, a clear TSS is present just upstream of the TIS for the truncated isoform, supporting the transcriptional control model (Figure 3.3C).

To assess the prevalence of truncated transcripts corresponding to truncated proteins genome-wide, we constructed TL-seq metagene profiles for the regions upstream of translation initiation sites (Figure 3.3D). For annotated genes, this displays the expected profile of high read density peaking around 50bp upstream of the TIS, consistent with the average length of a yeast 5' UTR (Nagalakshmi et al., 2008). The metagene for truncated isoforms displays a strikingly similar profile, indicating that it is common for truncated proteins to be produced from truncated transcripts (Figure 3.3D). Importantly,

for this analysis we excluded truncated isoforms whose annotated isoform would start within the window included in the metagene profile (-200bp) such that the metagene profile should not include any reads derived from annotated transcript isoforms.

We hypothesized that regulation via truncated transcripts might be more common for truncations that initiate far (distal) from the annotated isoform since control at the translational level would in most cases require the ribosome to bypass several in-frame start codons, making transcriptional regulation a more parsimonious explanation. Truncated isoforms starting close (proximal) to the annotated start site, however, would be more likely to only require bypassing of the annotated start codon which could easily occur due to leaky scanning (Kozak, 2005). For purposes of this comparison, we defined “proximal” truncations as starting within 40aa of the annotated start codon and “distal” truncations as starting greater than 40aa from the annotated start codon. To determine which truncated isoforms had evidence of a corresponding truncated transcript, we used TL-seq data to calculate a “TSS score” which is the ratio of the sum of TL-seq reads 200bp upstream of the truncation TIS over the sum of reads 200bp downstream of the truncation TIS. A higher ratio indicates stronger evidence of an alternative transcription 5’ end (likely generated by a downstream transcription start site) that is close to the truncation TIS. To assess statistical significance, we compared each ratio to an empirical distribution of TSS scores created by randomly sampling 200bp windows from within the same gene. Due to the differences in staging between the two meiotic time courses, we were unable to do high-resolution time point matching between the time courses. However, we were able to achieve a level of temporal resolution that was still compatible with the differences in staging by splitting timepoints into either early or late meiosis based on correlation coefficients between time points in matched mRNA-seq data for each time course, as well as expression patterns of key meiotic genes (Figure S3.3A-B). From this analysis we observed that distal truncations were much more likely to have a TSS than proximal truncations, with approximately two-thirds of distal truncations showing evidence of transcriptional regulation via a truncated transcript ($p<0.05$ for both early and late meiotic groups, Fisher’s Exact Test, Figure 3.3E).

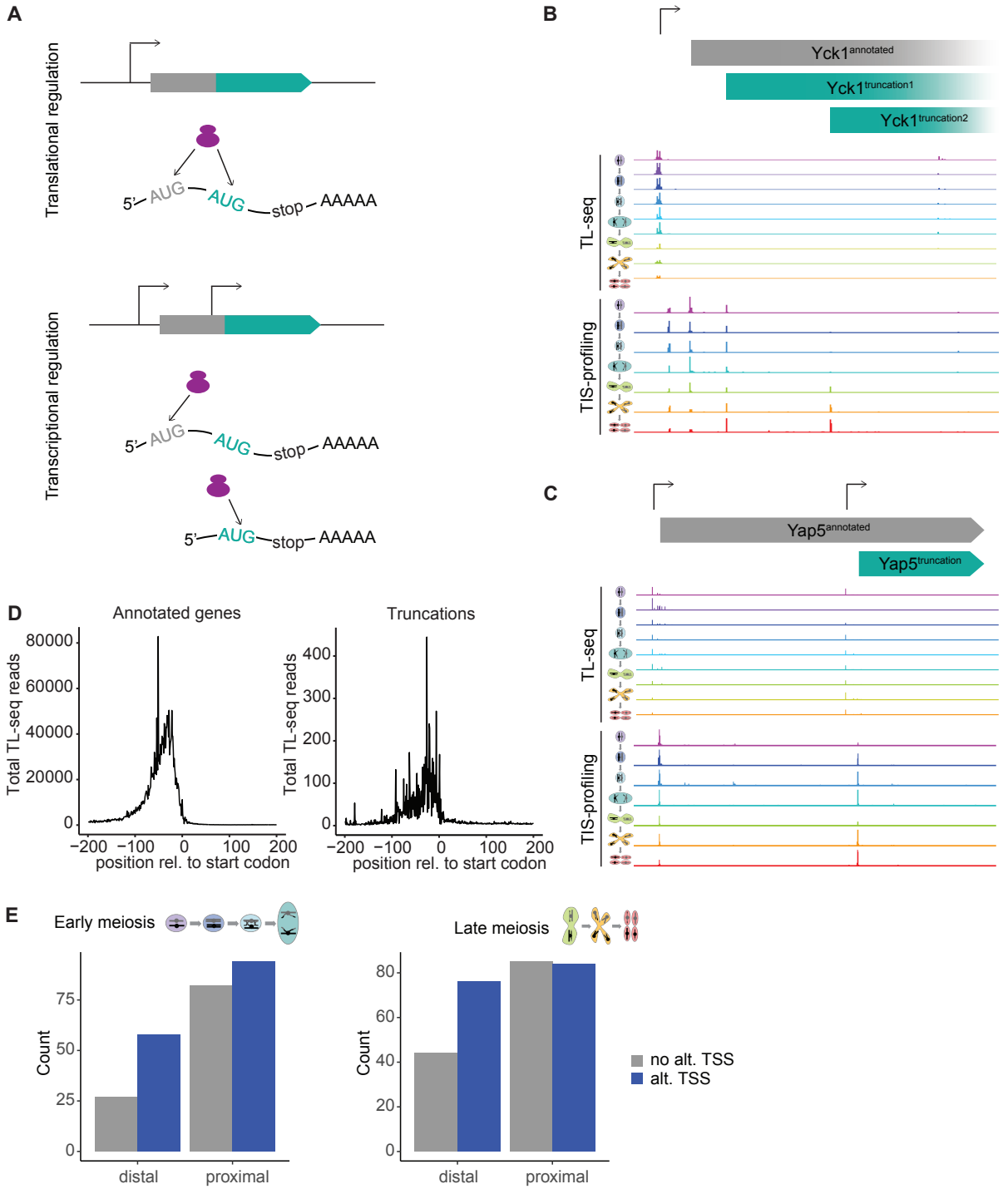


Figure 3.3 Truncated protein isoforms are often, but not always, produced from truncated transcript isoforms

(A) Schematic of potential regulatory mechanisms for truncated protein isoform production: (1) Translational regulation (top) in which a single transcript isoform is produced, and protein isoform production is determined by initiation site selection. (2) Transcriptional regulation (bottom) in which an annotated and truncated transcript isoform are translated into annotated and truncated protein isoforms, respectively. In each case, a schematic of the locus is shown with bent arrows representing transcription start sites above and resulting transcript(s) shown below.

(B) TL-Seq (above) and TIS-profiling (below) data for the *YCK1* locus showing production of the annotated isoform and two truncated protein isoforms but evidence for only the canonical transcript isoform. Cartoons at left represent meiotic timepoints (rainbow). Bent arrow represents the predicted transcription start site. The start peak upstream of the annotated start codon corresponds to an upstream open reading frame (uORF).
(C) TL-Seq (above) and TIS-profiling (below) data for the *YAP5* locus showing annotated and truncated transcript and protein isoform production. Cartoons at left represent meiotic timepoints (rainbow). Bent arrows represent predicted transcription start sites.
(D) Metagene plots of total TL-seq reads for the window -200 to +200bp surrounding translation initiation sites for annotated (left) and truncated (right) protein isoforms. Truncations with an annotated start beginning within the -200bp window were excluded to avoid including reads derived from annotated transcripts.
(E) Bar plot of number of distal (>40aa from annotated start) and proximal (≤40aa from annotated start) truncated isoforms with evidence of an alternative transcription start site ("alt. TSS") or not ("no alt. TSS") based on interpretation of 5' ends in approximately stage-matched TL-seq data collected in either early (left) or late (right) meiosis. Distal isoforms are significantly more likely to show evidence of an alternative transcript isoform ($p<0.05$ for both early and late groupings, Fisher's Exact Test).

3.2.6 Distal truncated protein isoforms for Yap5 and Pus1 exhibit condition-specific regulation

To better understand the functional relevance of distal, transcriptionally regulated truncations, we performed in-depth characterization of two examples, at the *YAP5* and *PUS1* loci. TIS-profiling data showed a very sharp start peak at an in-frame start codon within the *YAP5* locus, indicating translation of a truncated protein isoform ($Yap5^{truncation}$) across all meiotic time points (Figure 3.1C). Western blot analysis confirmed that the truncation is expressed across meiosis but is low during vegetative exponential growth (Figure 3.2A, 3.4A). Matched TL-seq data shows evidence of a transcript isoform with its 5' end positioned approximately 30bp upstream of the $Yap5^{truncation}$ start codon, suggesting that the truncated protein isoform is produced from an alternative transcript (Figure 3.3C). Given the dynamic regulation of $Yap5^{truncation}$ we wondered whether it had a role related to that of the full-length protein. Annotated Yap5 is an iron-response transcription factor that upregulates its target genes upon exposure to elevated iron in order to mitigate iron toxicity (Li et al., 2008). It constitutively occupies its target promoters and activates target gene transcription upon binding to Fe-S clusters in high iron conditions. The DNA binding domain of Yap5 is in the N-terminal half of the protein and is not present in $Yap5^{truncation}$; the Fe-S cluster binding domain lies in the C-terminal half and remains intact in the truncated isoform (Li et al., 2008). Previous work with artificial truncations of Yap5 showed that a version containing a comparable C-terminal region alone is capable of binding Fe-S clusters, suggesting that the natural truncation is also capable of this function (Rietzschel et al., 2015). We therefore hypothesized that $Yap5^{truncation}$ may play a role in Fe-S cluster homeostasis.

Fe-S clusters are key cofactors in the electron transport chain and thus important for mitochondrial function. It is therefore notable that $Yap5^{truncation}$ is induced during meiosis, a condition requiring respiration, but is lowly expressed in mitotic growth conditions, which favor fermentation. We also observed increased $Yap5^{truncation}$ expression in mitotic cells grown to saturation in rich media, a condition in which cells have undergone the diauxic shift from fermentation to respiration upon exhaustion of their fermentable

carbon sources (Figure 3.4A). We wondered whether the Yap5 truncation would also be induced in other respiratory conditions. To test this, we grew cells in dextrose-containing (fermentable) media, and after 4 hours of growth split the cells into either dextrose- or glycerol-containing (non-fermentable) media. We then assayed Yap5^{truncation} production by western blotting and observed upregulation of Yap5^{truncation} upon the switch to non-fermentable media (Figure 3.4B). These experiments suggest that dynamic production of Yap5^{truncation} occurs as part of a cellular response to conditions of increased respiratory activity.

Experimental validation, as well as comparison to TL-seq and standard ribosome profiling data, support the robustness of our approach for identifying hundreds of new truncations, a class of non-canonical coding region that has been difficult to annotate sensitively and reliably. We chose parameters specifically to minimize false positive detection, a major issue for this class of protein in our experience (Eisenberg et al., 2020). However, this also necessitated excluding some promising predicted truncations that did not pass some of our filters. As an example, TIS-profiling data indicated that the *PUS1* locus also encodes two isoforms, the annotated isoform and a truncated isoform (Pus1^{truncation}), but the data were noisier than cases like Yap5^{truncation}, and this resulted in our algorithm only calling this truncation at one timepoint, leading to its exclusion from our final list of truncated isoforms (*YAP5:* Figure 3.3C, *PUS1:* Figure 3.4C). In support of Pus1^{truncation} representing a true case, TL-seq data additionally showed evidence of an alternative transcript isoform beginning about 50-70bp upstream of the Pus1^{truncation} start codon, indicating that the truncated protein is likely to be translated from an alternative transcript (Figure 3.4C). Moreover, Pus1^{truncation} was readily detected by western blotting of cells expressing a C-terminally 3V5-tagged Pus1, and is dynamically expressed across meiosis, with levels peaking early in meiosis and again later in spores (Figure 3.2D). After confirming expression of Pus1^{truncation}, we next sought to understand its function. The annotated Pus1 protein is a pseudouridine synthase that is known to play key roles in translation through modification of tRNAs and snRNAs (reviewed in Rintala-Dempsey & Kothe, 2017). More recently, Pus1 has also been found to modify a subset of mRNAs (Basak and Query, 2014; Carlile et al., 2014; Lovejoy et al., 2014; Massenet et al., 1999; Motorin et al., 1998; Schwartz et al., 2014). Pus1^{truncation} contains regions involved in RNA binding, present at the C-terminus of the full-length protein, but lacks the N-terminal catalytic domain responsible for pseudouridylation in annotated Pus1 (Czudnochowski et al., 2013; Rintala-Dempsey and Kothe, 2017).

Our western blot data indicate that Pus1^{truncation} is most abundant in early meiosis and in spores, which are both times that display reduced levels of translation, as measured by polysome profiling (Brar et al., 2012). To assess whether decreased translation is associated with Pus1^{truncation} expression, we analyzed multiple additional conditions in which translation is reduced. First, we investigated the regulation of the *PUS1* locus in a standard ribosome profiling dataset collected in in meiotic cells lacking ribosomal protein Rpl40a, a condition which even more severely reduces translation early in meiosis (Cheng et al., 2019). Relative to WT, *rpl40a*Δ cells show reduced translation of the annotated Pus1 isoform and much more prominent translation of Pus1^{truncation} (Figure S3.4A).

To monitor Pus1^{truncation} levels in an orthogonal low-translation context, we treated mitotically growing cells with rapamycin, a general inhibitor of growth and ribosome biogenesis. Upon rapamycin treatment, we observed marked induction of Pus1^{truncation}, compared to undetectable levels in the untreated vehicle control (Figure 3.4D). Glucose starvation is another condition that reduces translation. By western blotting, we compared Pus1^{truncation} production in cells grown in normal rich media (2% glucose YEPD) to those grown in low-glucose media (0.2% glucose YEPD). After several hours in low-glucose media, we observed slight induction of the truncated isoform, compared to no induction in standard rich media (Figure 3.4E). Together, these data are consistent with a role for Pus1^{truncation} in conditions in which cellular translation is lowered, perhaps also explaining why this truncation has not been observed previously in the many studies focused on standard nutrient-rich growth conditions.

To further understand the role of Pus1^{truncation}, we strongly expressed the truncated isoform in vegetative exponential cells, a condition where the truncation is not normally present, and performed mRNA-seq to assess transcript level changes. We used an anhydrotetracycline-inducible allele to conditionally express Pus1^{truncation} and collected uninduced and induced samples in both a WT and deletion (*pus1*Δ) background (Figure S3.4B). Expression of Pus1^{truncation} alone had relatively little impact on overall gene expression, as measured by the number of differentially expressed genes relative to WT (n=61). Deletion of *PUS1* resulted in a slightly increased, but still modest, number of differentially expressed genes (n=210; $p<2.2e^{-16}$, Fisher's Exact Test). The combination of *pus1*Δ with Pus1^{truncation} expression, by contrast, yielded dramatically higher rates of differential gene expression, suggesting a synthetic phenotype between the two perturbations (n=1690; $p<2.2e^{-16}$, Fisher's Exact Test; Figure 3.4F, S3.4C).

We performed gene ontology (GO) analysis of the differentially expressed genes in *pus1*Δ cells expressing Pus1^{truncation} and found that the genes that were downregulated in the *pus1*Δ with Pus1^{truncation} expression condition were strongly enriched for GO terms related to translation and RNA modification substrates, including ncRNA production and processing, rRNA processing, and ribosome biogenesis (Figure S3.4D). To better understand the nature of the synthetic phenotype and visualize expression patterns across all our samples, we performed hierarchical clustering of all genes that were differentially expressed between WT and *pus1*Δ with Pus1^{truncation} expression (Figure 3.4G). Cluster C contains genes that are upregulated in both WT and *pus1*Δ cells upon Pus1^{truncation} induction. This gene set is generally upregulated upon aTC-induction of entirely unrelated genes and is therefore likely to be drug-dependent rather than Pus1^{truncation}-specific (data not shown). Cluster E contains genes that are strongly downregulated in *pus1*Δ cells but are unaffected by Pus1^{truncation} expression. This cluster is enriched for genes involved in translational regulation, RNA export, and rRNA processing, consistent with known roles of Pus1. In light of the observed synthetic effects of expression of Pus1^{truncation} in a *pus1*Δ background, two additional clusters are of particular note: (1) Cluster B shows very little change upon Pus1^{truncation} expression alone, modest downregulation occurring with the *pus1*Δ alone, and much more pronounced downregulation arising in the *pus1*Δ with Pus1^{truncation} expression. This

cluster is enriched for rRNA processing and ribosome assembly. (2) Cluster A, in contrast, contains genes that are largely unaffected with either of the individual perturbations but are strongly downregulated in the *pus1*Δ with Pus1^{truncation} expression. This cluster is strongly enriched for genes involved in rDNA heterochromatin formation, rRNA processing, splicing, and snoRNA metabolism. Together these results indicate that, while the loss of Pus1 alone has some impact on processes related to translation and ribosome biogenesis, the addition of Pus1^{truncation} expression expands the number of affected genes and increases the severity of the changes to translation-related transcript expression. Thus, Pus1^{truncation} appears to have subtle but global effects on translation that manifest primarily in the sensitized hypomodified context of *pus1*Δ cells, arguing that the role of the truncated isoform of Pus1 is not dependent solely on the function of the annotated full-length Pus1 isoform but may be related to the functions of other pseudouridine synthases that affect translation.

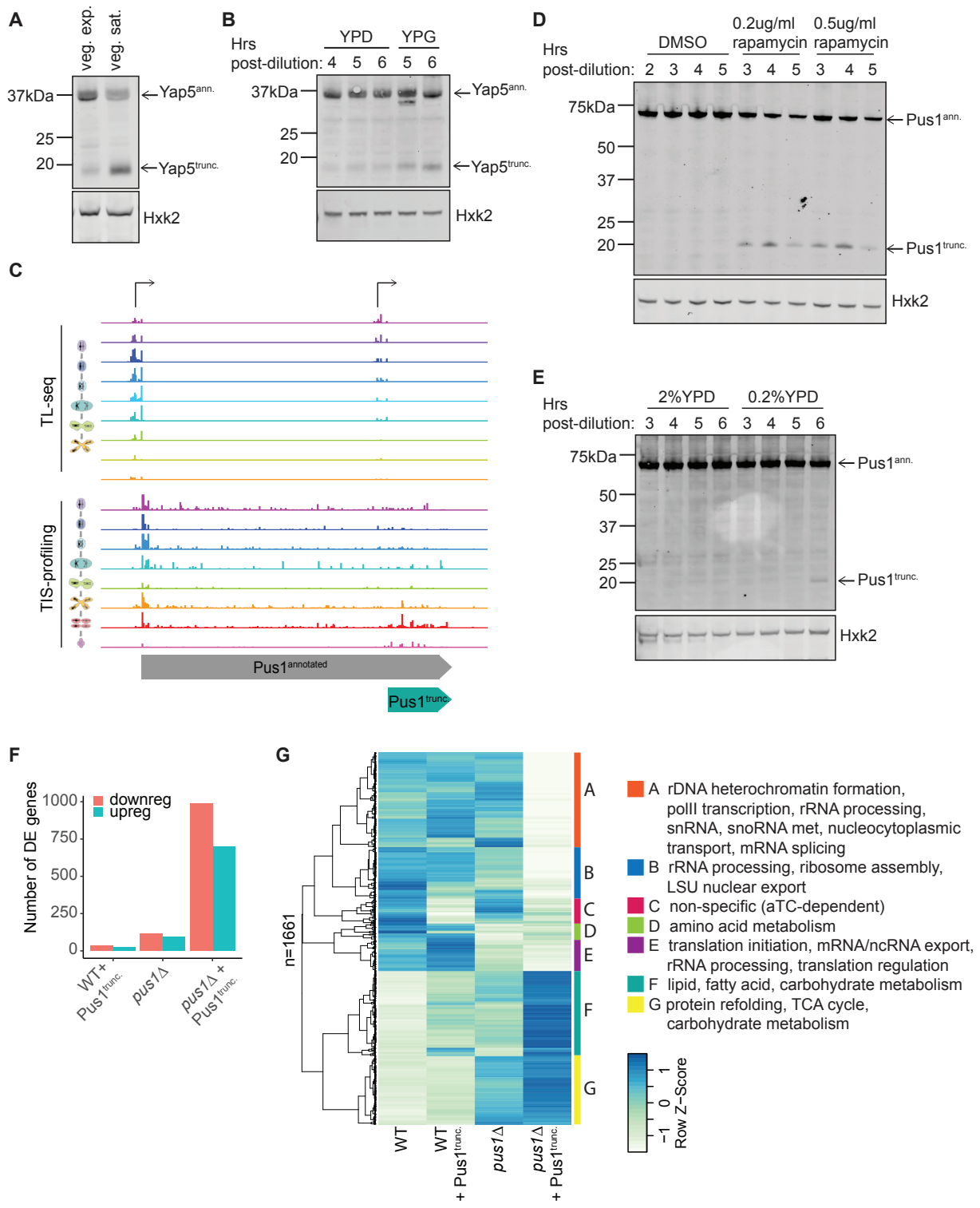


Figure 3.4 Condition-specific regulation of distal truncated protein isoforms
(A) Western blot for Yap5-FLAG. Samples were collected during vegetative exponential and vegetative saturated growth in YEPD. Hexokinase (Hxk2) is shown as a loading control.
(B) Western blot for Yap5-FLAG from cultures grown in rich media (YEPD) for 4 hours then transferred to either YEPD (fermentable) or YEPG (non-fermentable) media. Hexokinase (Hxk2) is shown as a loading control.

(C) TL-seq (above) and TIS-profiling (below) data at the Pus1 locus showing annotated and truncated transcript and protein isoforms. Cartoons at left of each data type indicate meiotic timepoints (rainbow). Bent arrows represent predicted transcription start sites. Note that very early TL-seq timepoints are not fully stage-matched to the first TIS-profiling timepoint, and late TIS-profiling timepoints do not have matched equivalents in the TL-seq time course, leading to patterns that appear to differ in timing when comparing patterns between the two sets of samples.
(D) Western blot for Pus1-3V5 from samples grown in rich media (YEPD) and treated with either DMSO or two different rapamycin concentrations for the times indicated. Hexokinase (Hxk2) is shown as a loading control.
(E) Western blot for Pus1-3V5 from samples grown in either standard (2%YEPD) or reduced-nutrient (0.2%YEPD) media for the timepoints indicated. Hexokinase (Hxk2) is shown as a loading control.
(F) Bar graph showing the number of genes differentially expressed in the following conditions: WT cells carrying a construct to allow anhydrotetracycline-inducible Pus1$^{trunc.}$ expression and treated with anhydrotetracycline (aTC; "WT + Pus1$^{trunc.}$"), *pus1*Δ cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with vehicle ("*pus1*Δ") or aTC ("*pus1*Δ + Pus1$^{trunc.}$"). In all cases differential expression is relative to WT cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with vehicle ("WT"). Differentially expressed genes were determined using DESeq2.
(G) Clustered heatmap of RNA-seq data for all genes that were differentially expressed between WT cells treated with vehicle ("WT") and pus1Δ cells treated with aTC ("*pus1*Δ + Pus1$^{trunc.}$"). Samples include WT cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with vehicle ("WT") or aTC ("WT + Pus1$^{trunc.}$"), cells deleted for PUS1 and carrying a construct to allow aTC-inducible Pus1^{trunc} expression and treated with vehicle ("*pus1*Δ") or aTC ("*pus1*Δ + Pus1$^{trunc.}$"). Values are the log of the average of two replicates. Seven discrete clusters are indicated to the right with colored bars and letters. Gene ontology terms enriched in each cluster are indicated to the right.

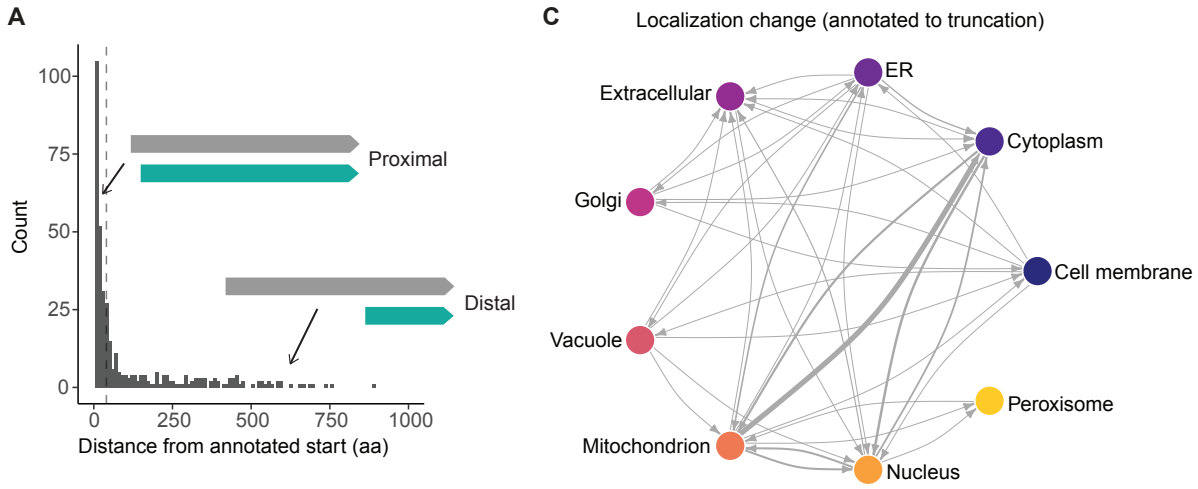
3.2.7 Proximal truncations are a general mechanism for encoding multiple localizations of protein products at a single locus

Although in-frame start codons are uniformly distributed throughout coding genes, their usage as translation start sites in our dataset is very non-uniform, with a strong bias for more N-terminal (proximal to annotated starts) start sites (Bazykin and Kochetov, 2011) (Figure 3.5A). In fact, nearly 60% of all truncations start within 40aa of the corresponding annotated start codon. We hypothesized that – unlike in the case of distal truncations, which are likely to encode isolated domains – for proximal truncations, the core functional domains of the protein should typically remain intact in the truncated isoform, but the missing N-terminal sequences could encode localization signals, resulting in differential localization of the truncated isoform. This hypothesis is supported by the fact that among the small set of previously known truncated isoforms in *S. cerevisiae*, nearly all serve the function of differentially localizing a subset of the protein (Table 3.1, 3.2, Figure 3.5B). We reasoned that our dataset would be an excellent opportunity to test whether this is indeed a broad phenomenon and to potentially identify additional differentially localized truncated isoforms.

We used a published algorithm, DeepLoc1.0, to perform localization prediction for our truncated isoforms and their annotated counterparts (Almagro Armenteros et al., 2017). DeepLoc1.0 is a deep learning algorithm trained on existing protein databases and uses sequence information alone to predict protein localization. It was important to use an algorithm that used sequence rather than homology, as the truncated and annotated isoforms would have very similar homology given that they share significant amounts of

sequence. Since localization signals are frequently encoded at the N-terminus, we reasoned that “artificial” truncations could easily generate differentially localized proteins by removing an N-terminal sequence. To determine the expected background rate of differential localization we generated sets of simulated truncations by randomly sampling in-frame start codons and performing localization prediction on those open reading frames compared to the annotated protein. We found that our set of truncations differentially localized at a higher rate than expected by chance, with just over 30% of truncated isoforms localizing to a different compartment than their respective annotated isoforms (Figure S3.5A).

Among differentially localized truncations, it was most common for the truncations to lose the predicted mitochondrial or nuclear localization of their annotated counterpart, while cytoplasmic or extracellular localization were the most likely localizations to be gained (Figure 3.5C, S3.5B). For initial validation of this prediction method, we compared the predicted and experimentally determined localizations for previously characterized truncations. For all but one of these genes, *GRX2*, the localization predictions match the characterized localization of the known isoforms, suggesting that the localization predictions were robust (Figure 3.5B).



B

Gene	Characterized localization	Predicted localization
GRX2	Mitochondria / Cytoplasm	ER / Cytoplasm
CCA1	Mitochondria / Cytoplasm	Mitochondria / Cytoplasm
SUC2	Extracellular / Cytoplasm	Extracellular / Cytoplasm
CRS1	Mitochondria / Cytoplasm	Mitochondria / Cytoplasm
MOD5	Mitochondria / Cytoplasm / Nucleus	Mitochondria / Cytoplasm / Nucleus
GLR1	Mitochondria / Cytoplasm	Mitochondria / Cytoplasm
FUM1	Mitochondria / Cytoplasm	Mitochondria / Cytoplasm
LEU4	Mitochondria / Cytoplasm	Mitochondria / Cytoplasm

Figure 3.5 Computational prediction of differentially localized truncated isoforms
(A) Distribution of distances between annotated and truncated isoform translation initiation sites among 388 truncations called by the algorithm, with cartoons above representing the “proximal” and “distal” categorization of truncations. The dotted line represents the 40aa cutoff between the proximal and distal categories.

(B) Table of previously characterized truncations called by our algorithm, comparing their experimentally characterized localization with their predicted localization using DeepLoc1.0. Discrepancies between predicted and characterized localization are indicated in red. Note that for one gene *CCA1*, our algorithm only called the larger of two known truncated isoforms, so the smaller isoform was not included in the localization predictions.
(C) Schematic of predicted subcellular localizations of annotated and truncated protein isoforms, with the blunt and pointed end of the arrows representing the annotated and truncated isoforms, respectively. Localization prediction was performed using DeepLoc1.0. Weight of line indicates number of truncations in each category.

We selected six predicted dual localized truncated proteins and sought to experimentally validate their localization using fluorescence microscopy (Figure S3.6A). Candidate proteins were primarily selected based on the quality of start-site profiling data at the relevant locus and their ability to be tagged and imaged. Five candidates had no known alternative isoforms in existing literature. The sixth, *GRX2*, had a characterized truncated isoform for which the localization predictions did not match the characterized localizations (Figure 3.5B). For each locus of interest, we inserted a fluorescently tagged transgene under the control of the gene's endogenous promoter. To determine the localization of each isoform separately, we generated three strains for each gene: (1) all isoforms, with unmutated start codons, (2) annotated only, with the truncation start codon mutated to isoleucine, and (3) truncation only, with the annotated start codon mutated to isoleucine (Figure 3.6A, S3.6B-D). Two genes, *IGO2* and *APE4*, were not possible to validate - the truncated isoform of Igo2 was not detectable by western blotting or microscopy, and although both isoforms of Ape4 could be detected by western blotting they were too low-abundance to detect by microscopy. For the remaining candidates, we performed fluorescence microscopy, collecting images at representative meiotic time points at which the respective truncations are expressed.

Bna3 is a kynurenine aminotransferase known to be dual localized to the mitochondria and cytoplasm but with only one characterized isoform and no known mechanism for its dual localization (Karniely et al., 2006; Wogulis et al., 2008). Since the annotated isoform was predicted to be mitochondrial and the truncated isoform was predicted to be cytoplasmic, we hypothesized that the mechanism of the previously observed dual localization could be via production of two isoforms. This prediction was indeed validated by microscopy, as the "annotated only" strain showed only mitochondrial localization and the "truncation only" strain showed cytoplasmic localization (Figure 3.6B).

Rex2 is an RNA exonuclease involved in snRNA, rRNA, and ncRNA processing with additional roles in mitochondrial DNA escape (Hanekamp and Thorsness, 1999; van Hoof et al., 2000). It is characterized as having mitochondrial localization, but its role in processing nuclear RNAs strongly suggested nuclear localization as well. This information aligned well with the DeepLoc1.0 prediction that the annotated isoform is mitochondrial, and the truncated isoform is nuclear. Upon imaging by microscopy, the strain containing only the annotated isoform showed mitochondrial localization and the strain containing only the truncated isoform showed nuclear localization, in support of the predicted localizations for both isoforms (Figure 3.6C).

Ath1 is an acid trehalase involved in extracellular trehalose degradation. Its localization has been a topic of debate, with one study showing it to be vacuolar and another showing it to be periplasmic (He et al., 2009; Huang et al., 2007). According to our predictions, the annotated isoform is Golgi-localized and the truncated isoform is extracellular. We hypothesized that the extracellular localization of the truncation could correspond to the characterized periplasmic localization that had previously been attributed to the annotated isoform. This was supported by the fact that the periplasmic localization had been observed using an endogenous C-terminal tag, which would result in tagging of both isoforms (He et al., 2009). The vacuolar localization, however, was observed with an N-terminal tag under a strong promoter, which would only tag the annotated isoform and could result in vacuolar signal simply due to protein degradation (Huang et al., 2007). Microscopy revealed that indeed the truncated isoform in isolation localizes to the periplasmic space (Figure S3.6E). Its levels also increase in spores, which aligns with its known function in breaking down trehalose, a storage carbohydrate that needs to be mobilized in spores. The annotated isoform, however, was not possible to visualize. This suggests that the truncated isoform is likely the main functional isoform, at least for the characterized function, and that the annotated isoform is either not produced or is at very low abundance under the conditions we assayed. This is consistent with published work showing that artificially periplasm-targeted Ath1 is sufficient for growth on trehalose while vacuole-targeted Ath1 is not (He et al., 2009).

Finally, we investigated the localization of Grx2, a glutaredoxin with two known in-frame start codons, one producing the annotated mitochondrial isoform and the second producing a truncated cytoplasmic isoform (Pedrajas et al., 2002). It was the only characterized gene with a truncated isoform for which our localization predictions did not match the characterized localizations, with the annotated isoform predicted to be ER-localized rather than the characterized mitochondrial localization (Figure 3.5B). To our surprise, when we imaged the annotated isoform alone, we saw clear ER localization, in line with our prediction but contradicting the established localization (Figure 3.5B). To reconcile this with the published data, we returned to the TIS-profiling data and noticed an additional in-frame start codon with clear signal, this time at a TTG codon, one of the most efficient of nine known near-cognate start codons (differing from ATG by one nucleotide; Figure 3.6D) (Kearse and Wilusz, 2017; Kolitz et al., 2008). We hypothesized that this could be the source of the missing mitochondrial localization, and indeed, the DeepLoc1.0 algorithm predicted mitochondrial localization for this isoform. To validate this prediction by microscopy, we generated strains that expressed each isoform alone by mutating the other two start codons, either to isoleucine (ATT) for the ATG starts or to a non-near-cognate leucine (TTA) for TTG. The three isoforms localized as predicted, with the annotated isoform in the ER, the TTG truncation in the mitochondria, and the ATG truncation in the cytoplasm (Figure 3.6E). Therefore, our truncation calling algorithm, in combination with DeepLoc1.0, correctly predicted the existence and localization of several previously characterized truncations and provided new insights into the regulation of the Grx2, a glutaredoxin that is localized to three different subcellular compartments by expression of three protein isoforms with slightly different N-termini. Together with the prediction of novel dual-localized isoforms, these

localization analyses revealed that production of proximal truncations may be a widespread strategy for targeting one protein to two or more subcellular localizations.

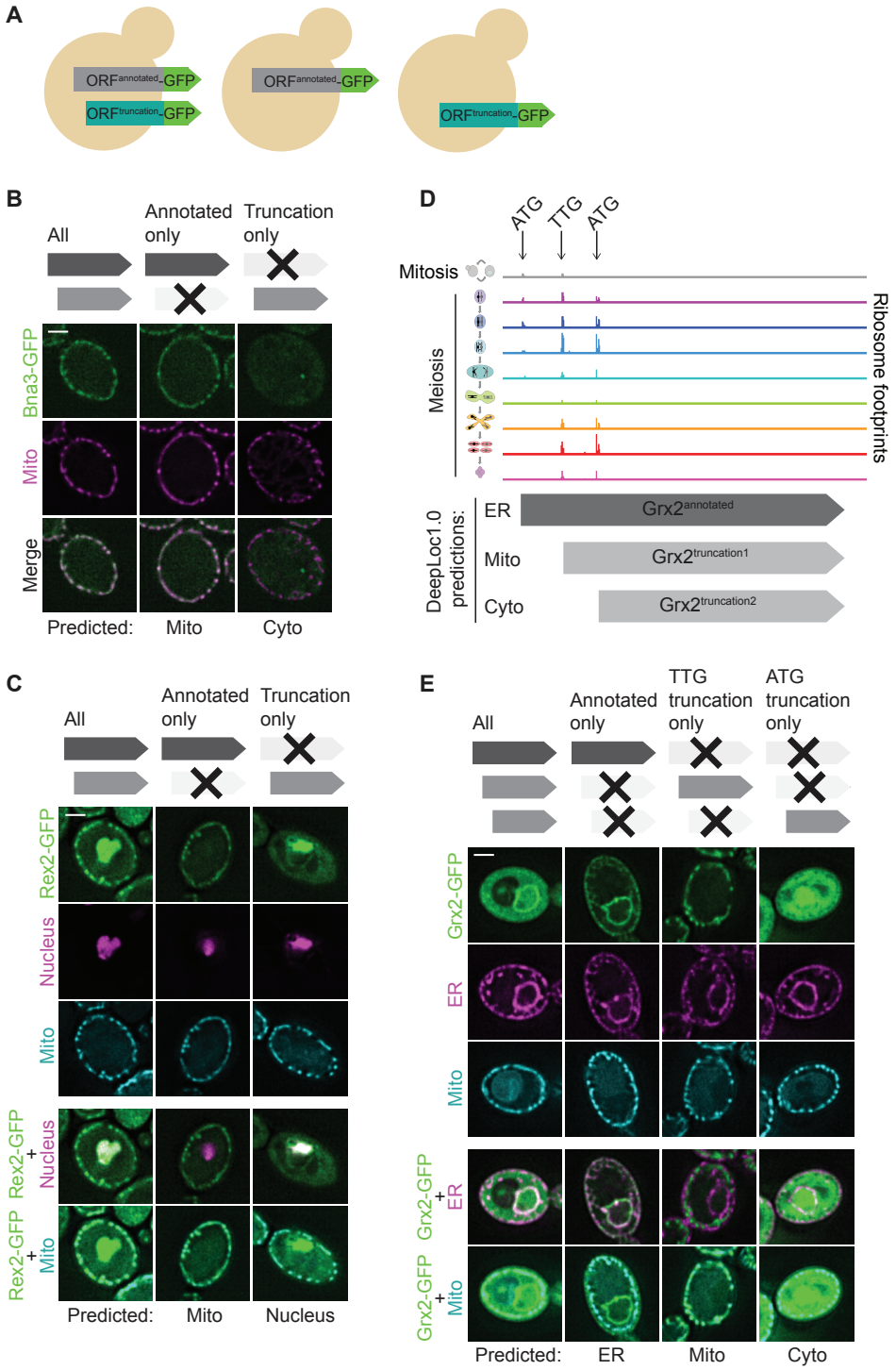


Figure 3.6 Experimental validation of differentially localized truncated isoforms
(A) Schematic of experimental approach for testing localization of candidate truncated and annotated isoforms. ORFs of interest were fused to GFP to create a C-terminal tagged version of both annotated and N-terminally truncated isoforms simultaneously. Strains were constructed that contained either (1)

WT start codons for both annotated and truncated isoforms, (2) the truncation start codon mutated to encode isoleucine, or (3) the annotated start codon mutated to encode isoleucine.
(B) Fluorescence microscopy of C-terminally GFP-tagged Bna3, collected at 6h in SPO, using the approach outlined in (A). Mitochondrial localization is indicated by Cit1-mCardinal. DeepLoc1.0-based predictions are shown below, constructs analyzed in each column are illustrated above. Scale bar is 2μm.
(C) Fluorescence microscopy of C-terminally GFP-tagged Rex2, collected at 1h in SPO, using the approach outlined in (A). Mitochondrial localization is indicated by Su9MTS-BFP; nuclear localization is indicated by Htb1-mCherry. Separate merged images for nuclear and mitochondrial signal are shown at the bottom. DeepLoc1.0-based predictions are shown below, constructs analyzed in each column are illustrated above. Scale bar is 2μm.
(D) TIS-profiling data at the *GRX2* locus across all meiotic time points. Cartoons to the left indicate whether track represents mitotic cells (gray) or meiotic timepoints (rainbow). Arrows indicate start codons for the annotated isoform and cartoons of the annotated and two predicted truncated isoforms are illustrated below, with predicted localizations for each isoform are to the left.
(E) Fluorescence microscopy of C-terminally GFP-tagged Grx2, collected at 0h in SPO, using an expanded version the approach outlined in (A). Strains were designed to express either all isoforms, annotated isoform only, TTG truncated isoform only, or ATG truncated isoform only. Mitochondrial localization is indicated by Su9MTS-BFP; ER localization is indicated by mCherry-HDEL. Separate merged images for ER and mitochondrial signal are shown at the bottom. DeepLoc1.0-based predictions are shown below. Constructs analyzed in each column are illustrated above. Scale bar is 2μm.

3.3 Discussion

To date, only a handful of truncated isoforms have been characterized, and most were identified in single-gene studies of very well-studied biological pathways (Table 3.1, 3.2). Global approaches to systematically identify these isoforms by ribosome profiling and mass spectrometry have been hampered by limitations imposed by their overlap with known proteins. Here we develop a novel algorithm for identifying truncated protein isoforms and report the translation of 388 truncated proteins in budding yeast, an organism generally considered to have little protein isoform diversity. Our analysis of this dataset represents a dramatic increase in the number of known truncated proteins in yeast and provides an unbiased picture of the prevalence and contexts in which these isoforms are present. Although functions for most remain to be investigated, we show evidence of functional activity for several truncations, with examples of proximal truncations serving to differentially localize proteins and distal truncations acting in condition-specific roles that may differ from the function of their annotated isoform. In addition to the rich global information it contains, we believe that this dataset will be a valuable resource for single gene studies, as it will provide more complete information about the gene products present at a given locus and may be helpful for generating functional hypotheses.

We orthogonally validated the production of numerous novel truncated isoforms by performing western blotting for C-terminally epitope tagged proteins (Figure 3.2A-G, S3.2A-F). Of the candidates we chose, 10 were detectable by western blotting, generally at time points consistent with the TIS-profiling data. The remaining five truncations that we tested, however, were not readily detectable by western blotting despite strong and specific signal in the TIS-profiling data. We showed that several truncated proteins are subject to proteasome-mediated degradation, which in some cases made their detection in untreated cells more challenging (Figure 3.2K-L, S3.2G-I). The observed degradation of this set of truncated proteins could suggest that they are

turned over because they are non-functional or deleterious. It is also possible, however, that in some cases instability is an important functional feature of the protein. For one previously characterized truncation, produced at the *KAR4* locus, the annotated isoform is stable and constitutively expressed, while the truncation is much less stable and is expressed in response to mating pheromone. The unstable truncated isoform serves to provide a large boost in protein expression during mating and is then rapidly degraded when no longer needed since high levels are toxic to the cell (Gammie et al., 1999). This illustrates a scenario where the lack of stability is in fact an important feature of the truncated isoform. Such regulation could exist at other loci with truncated isoforms, providing stable and unstable pools of protein that are either functionally identical (in the case of proximal truncations) or functionally distinct (in the case of distal truncations). In general, we observed that truncated proteins beginning closer to the annotated start codon are more robustly and stably expressed than more distal truncations, which tended to be much more difficult to detect despite often having strong signal in the TIS-profiling data. We suspect that more proximal truncations are more likely to be stable since they are quite similar in size and composition to the annotated protein. Given the outsized importance of the N-terminus of proteins for protein stability, however, even a slight difference between similarly sized isoforms could confer differences in stability.

We observed dynamic regulation for most truncated isoforms in the TIS-profiling data (Figure 3.1F), providing support for specific cellular functions for these newly identified proteins. While TIS-profiling can give an approximate picture of regulatory patterns, it is not a robustly quantitative measure, so specific peak heights should not be overinterpreted (Eisenberg et al., 2020). We did, however, observe consistent regulatory patterns by western blotting for several examples, indicating that the TIS-profiling can be informative for regulatory trends. In many cases, this dynamic regulation is likely facilitated by the presence of a truncated transcript rather than through translational regulation. We hypothesize that truncated proteins without a detectable truncated transcript are typically produced via start codon readthrough of the annotated start codon, although this remains untested in the cases presented here (Figure 3.3E). It remains possible, of course, that a subset of the truncated proteins apparently lacking a truncated transcript in fact arise from false positives in the TIS-profiling or false negatives in the TL-seq data; however, we expect that this would be a minor contribution.

A large subset of the proximal truncations that we identified seem to lead to otherwise functionally identical proteins being targeted to different subcellular locations, which provides a mechanism for how in so many cases, similar cellular functions are performed in multiple cellular compartments. For example, DNA replication, transcription, and translation all occur in both the mitochondria and either the nucleus or cytoplasm. These types of related but spatially separated functions can be encoded either by two separate genes (through gene duplication or functional convergence) or by a single gene (Danpure, 1995). Previous single-gene studies made it clear that a single locus can encode multiple differentially localized protein isoforms, but the full extent of the phenomenon was unknown, and the known examples were biased towards very well-studied biological pathways (Table 3.1, 3.2). For example, several previously

characterized truncated isoforms were amino acid tRNA synthetases, known to act on both cytoplasmic and mitochondrial tRNAs, and were therefore clear candidates for this type of regulation. Here, genome-wide data allowed us to see the extent of this type of regulation in yeast in a less biased way. Candidates that we validated for their predicted localization differences spanned a diverse range of functions and revealed a variety of ways in which knowledge of this cellular strategy can enhance our understanding of gene function and regulation.

In the case of Bna3, dual protein localization was already established, but its basis was unknown. The two isoforms at this locus follow a common trend seen among previously characterized truncations, in which the longer form has a mitochondrial signal sequence that is lost in the truncated isoform, causing it to default to cytoplasmic localization (Figure 3.6B). *REX2* presented a slightly different scenario, in which dual localization had not been established but was very likely given known functional information. Rex2 had been characterized as a mitochondrial protein by mitochondrial fractionation of overexpression strains (Hanekamp and Thorsness, 1999). Despite not being explicitly characterized as nuclear by microscopy, it was also shown to be involved in rRNA and snRNA processing, strongly suggesting a nuclear localization (van Hoof et al., 2000). Our predictions and validation support both localizations and suggest that the mechanism of dual localization is through production of two differentially targeted protein isoforms (Figure 3.6C). It is also an interesting example of one signal sequence being removed (mitochondrial) and another being unmasked (nuclear) in the truncated isoform. Interestingly, we did not find dual localization for *ATH1* and instead showed that the characterized function is likely carried out by the truncated isoform, and that the annotated isoform is likely not expressed at appreciable levels (Figure S3.6E). Identification of the truncated isoform and subsequent localization predictions were valuable for explaining inconsistencies in the existing literature.

The *GRX2* locus provides an additional demonstration of how TIS-profiling can help reconcile confusing information about a gene's regulation and function. Based on previous characterization, *GRX2* was thought to produce two protein isoforms, one mitochondrial and one cytoplasmic (Porras et al., 2006). In that study, however, three bands were observed by western blotting, which likely correspond to the three isoforms that we identified. The long isoform, however, was hypothesized to be mitochondrial and the intermediate isoform was attributed to processing of the mitochondrial targeting sequence from the longer isoform. A small amount of protein was also detected in the ER, but this was attributed to slow import kinetics into the mitochondria. With the additional insight provided by the TIS-profiling paired with localization prediction, we were able to identify a third isoform of the redox-regulator Grx2. Visualization of the protein structure using AlphaFold shows that all three isoforms, which localize to three different cellular compartments, still retain the structured, functional core of the protein (Figure S3.6F, 3.6D-E) (Jumper et al., 2021).

The identification of a TTG-initiated truncation at the *GRX2* locus raises the question of whether near-cognate start codons should have been included in our truncation calling algorithm. We chose to exclude them because visual analysis indicated that most near-

cognate truncations were false positives caused by background noise within genes and would likely require different treatment and calling thresholds than AUG start sites. While the *GRX2* locus shows that near-cognate-initiated truncations can be made, we still believe that they are very rare. Translation at near-cognate start codons alone is often not sufficient to stabilize a transcript, as shown in past work in which mutation of the annotated start codon of transcripts encoding a near-cognate-initiated extended protein isoform led to nonsense-mediated decay of the transcript; this was caused by efficient translation initiation at an out-of-frame AUG codon downstream following inefficient initiation at the in-frame near-cognate codon (Eisenberg et al., 2020). Therefore, we suspect that most cases of near-cognate truncations would need to arise in a context similar to *GRX2*, in which a truncated transcript bears an ATG truncation, whose translation ensures the stability of the transcript, paired with an upstream near-cognate-initiated isoform. In these cases, the near-cognate isoform is essentially behaving like an N-terminal extension within the context of the truncated transcript.

We observed dynamic and condition-specific regulation for two distal truncations that we investigated experimentally, Yap5^{truncation} and Pus1^{truncation}, which suggests functional relevance. Yap5^{truncation} contains the Fe-S cluster binding domain located in the C-terminal half of the annotated protein and is markedly similar in size to an artificial truncation of Yap5 that was shown to effectively bind Fe-S clusters (Rietzschel et al., 2015). We show that Yap5^{truncation} is specifically induced under multiple respiratory conditions: meiosis, saturated growth, and growth in non-fermentable media (Figure 3.2A, 3.4A-B). This suggests that it may be involved in responding to elevated respiratory activity, a role which could be related to its ability to bind Fe-S clusters, important cofactors in the electron transport chain. Further work will be necessary to elucidate its specific functional role.

Pus1^{truncation} is produced throughout meiosis and contains the positively charged residues involved in RNA binding of the full length protein (Czudnochowski et al., 2013). We show that this truncated isoform is likely nutrient-regulated since it is expressed in the low nutrient media that induces meiosis, as well as upon glucose starvation and rapamycin treatment (Figure 3.2D, 3.4D-E). This nutrient regulation is intriguing given that a number of Pus1-dependent modifications in mRNA are dynamically regulated during nutrient deprivation (Carlile et al., 2014; Schwartz et al., 2014). The induction of Pus1^{trunction} upon rapamycin treatment is also interesting; formation of some pseudouridines by other Pus proteins is known to be dynamically regulated by the TOR pathway, and a pseudouridine in the U6 snRNA is introduced by Pus1 during filamentous growth, also regulated by the TOR pathway (Basak and Query, 2014; Wu et al., 2016b, 2011).

mRNA-seq of cells expressing Pus1^{truncation} in WT and *pus1*Δ backgrounds under rich growth conditions revealed mild effects of either Pus1^{truncation} expression or *PUS1* deletion alone, and a much more dramatic effect when deletion of *PUS1* is combined with Pus1^{truncation} expression (Figure 3.4F-G). This more severe synthetic gene expression phenotype suggests that Pus1^{truncation} has effects that are not directly related to full-length Pus1 function, potentially also affecting targets of other pseudouridine

synthases as well. This is perhaps unsurprising given that the specificity of the enzyme is primarily conferred by the catalytic domain and there is little reason to think the RNA binding domain alone would be specific to Pus1 targets.

The precise reason for the strong downregulation of genes involved in ribosome biogenesis, rRNA processing, and ncRNA processing in cells expressing Pus1^{truncation} and deleted for *PUS1* is not immediately obvious (Figure S3.4D). Pseudouridine synthases as a group perform extensive pseudouridylation of rRNA, tRNA, snRNA, snoRNA, and mRNA targets, any of which could have important impacts on translation and RNA processing (reviewed in Rintala-Dempsey & Kothe, 2017). Pus1 itself has been shown to modify ribosomal mRNAs, including 5 subunits of the ribosomal large subunit, as well as RNase MRP which is involved in maturation of rRNA (Carlile et al., 2014; Schwartz et al., 2014). Given that Pus1^{truncation} contains regions involved in RNA-binding, we hypothesize that it could occlude target binding by Pus1 and potentially other pseudouridine synthases as well (Czudnochowski et al., 2013). The presence of a synthetic effect with deletion of *PUS1* is perhaps suggestive of a role in modulating pseudouridylation. While deletion of *PUS1* alone results in viable cells and only mild phenotypic affects, much more dramatic synthetic phenotypes have been observed when *PUS1* deletion is combined with deletion of other pseudouridine synthases or with mutations that compromise tRNA stability (Großhans et al., 2001; Khonsari and Klassen, 2020; Wu et al., 2016a). If Pus1^{truncation} has a role related to pseudouridylation, a synthetic phenotype with deletion of *PUS1* would be consistent. Further study will be necessary to understand the specific mechanistic role of Pus1^{truncation}.

Our hypotheses for Yap5^{truncation} and Pus1^{truncation} function were notably tied to the known functional characteristics of their annotated isoforms. Whether this is a valid approach is unclear, as many truncations – particularly distal truncations – lack key functional sequences of the annotated protein. The degree to which the annotated and truncated isoform differ in function likely varies depending on the extent of the truncation, with the proximal truncations being much more likely to share functional characteristics with the annotated isoform and only varying in typical N-terminally encoded characteristics such as stability and localization. Distal truncations, on the other hand, are more likely to be missing key functional domains and may not even contain any intact domains, making it much more difficult to generate a rational prediction for their functions.

Beyond the new regulation we uncovered for Pus1 as a result of identifying its truncated isoform, this case highlights a key point: our truncation-calling algorithm is stringent, likely excluding a number of real truncated isoforms in order to minimize false-positive calls, which previously plagued identification of N-terminal truncations. Systematic identification of this type of non-canonical protein is fundamentally distinct from other classes, including ORFs in upstream regions (uORFs), those that are short and intergenic (sORFs), those downstream of annotated ORFs (dORFs), and N-terminally extended ORFs. In all other cases, some or all of the novel ORF is non-overlapping with an annotated ORF. Thus, approaches based on standard ribosome profiling data that leverage initiation codon peaks resulting from cycloheximide pre-treatment, periodicity

resulting from elongation, or simply ribosome density, are not effective for stringently calling N-terminal truncations. Even approaches independent of standard ribosome profiling, like those that use evolutionary conservation to identify coding regions, are problematic for this class in particular; and those that rely on machine learning-based analysis of translation initiation site mapping in conjunction with standard ribosome profiling analysis generate an exceedingly high level of false positives and negatives, likely due to multiple of the factors noted above (Eisenberg et al., 2020). Our systematic identification of truncated proteins, in contrast, relied entirely on TIS-profiling data. Relative to standard ribosome profiling, the TIS-profiling data is much simpler to interpret since the reads are highly enriched at sites of translation initiation and signal is not obscured by elongating ribosomes from the overlapping annotated ORF, allowing us to very robustly identify truncated protein initiation sites (Figure 3.1A-B). Analysis of parallel standard ribosome profiling data provided clear validation of our calls, suggesting that future studies can rely on TIS-profiling for protein isoform identification (Figure 3.1E). Furthermore, multiple lines of experimental testing revealed that these isoforms can have localization and function distinct from their corresponding annotated ORF.

Non-canonical translation has previously been shown to be higher during meiosis and other stress conditions, and from our observations in this dataset, truncated isoforms are no exception (Brar et al., 2012; Cheng et al., 2018; Eisenberg et al., 2020). This could be evolutionarily beneficial, allowing cells to sample a greater proteomic diversity to adapt to new environments. Some truncated isoforms may not currently be “useful” to cells but may eventually over time become functional. While the use of different isoforms bears some similarity to gene duplication, it is markedly different in that the shared sequences between the two isoforms are unable to evolve separately. Only the region missing in the truncation can change independently between the two isoforms. N-terminal sequences, however, are often particularly important for gene function, as we have shown for localization and stability. Therefore, having a mechanism to test out different N-terminal sequences while still retaining protein production from the annotated start site could be beneficial. Future work on the prevalence and conservation of truncated isoforms across different stress conditions and other organisms will further elucidate both the functional relevance and evolutionary processes giving rise to truncated proteins.

3.4 Materials and Methods

3.4.1 Yeast strain construction

Strains were constructed in the SK1 background of *Saccharomyces cerevisiae*. Strains and plasmids used for this study are listed below.

BrÜn Strain No.	Genotype
13	MATa wild-type
14	MATalpha wild-type
15	MATa/alpha wild-type
5805	MATa/a wild-type

7318	MATa/alpha YAP5-FLAG::KanMX
30782	MATa/alpha MOD5-3V5::KanMX
30516	MATa/alpha PEX32-3V5::KanMX
21546	MATa/alpha PUS1-3V5::KanMX
30089	MATa/alpha SAS4-3V5::KanMX
30091	MATa/alpha SSP1-3V5::KanMX
30781	MATa/alpha YCK1-3V5::KanMX
30092	MATa/alpha TPO1-3V5::KanMX
30783	MATa/alpha PRP4-3V5::KanMX
30778	MATa/alpha GLK1-3V5::KanMX
30780	MATa/alpha ARI1-3V5::KanMX
30779	MATa/alpha RTT105-3V5::KanMX
30090	MATa/alpha SIW14-3V5::KanMX
32012	MATa/alpha his3::PUS1-3V5::Hyg; pus1::KanMX
32013	MATa/alpha his3::PUS1-M1I-3V5::Hyg; pus1::KanMX
32014	MATa/alpha his3::PUS1-M436I-3V5::Hyg; pus1::KanMX
32527	MATa/alpha his3::YAP5-FLAG::HIS3; yap5::KanMX
32528	MATa/alpha his3::YAP5-M1I-FLAG::HIS3; yap5::KanMX
32529	MATa/alpha his3::YAP5-M152I-FLAG::HIS3; yap5::KanMX
33093	MATa/alpha YCK1-3V5::KanMX; pdr5::KanMX
33092	MATa/alpha SAS4-3V5::KanMX; pdr5::KanMX
33090	MATa/alpha MOD5-3V5::KanMX; pdr5::KanMX
33089	MATa/alpha TPO1-3V5::KanMX; pdr5::KanMX
33091	MATa/alpha PRP4-3V5::KanMX; pdr5::KanMX
36451	MATa/alpha ARI1-3V5::KanMX; pdr5::KanMX
36450	MATa/alpha RTT105-3V5::KanMX; pdr5::KanMX
33088	MATa/alpha SIW14-3V5::KanMX; pdr5::KanMX
36736	MATa/alpha his3::pTetO7.1-altPus1-3V5::HIS3; ura3::pRNR2-TetR-Tup1, pTetO7.1-TetR::URA3
36735	MATa/alpha his3::pTetO7.1-altPus1-3V5::HIS3; ura3::pRNR2-TetR-Tup1, pTetO7.1-TetR::URA3; pus1::KanMX
34438	MATa/alpha trp1::BNA3-GFP::TRP1; CIT1-mCardinal::HIS3MX6
34439	MATa/alpha trp1::BNA3-M1I-GFP::TRP1; CIT1-mCardinal::HIS3MX6
34440	MATa/alpha trp1::BNA3-M13I-GFP::TRP1; CIT1-mCardinal::HIS3MX6
34397	MATa/alpha trp1::REX2-GFP::TRP1; HTB1-mCherry-HISMX6; 2micron_plasmid_KanMX_pGPD1-Su9-BFP
34398	MATa/alpha trp1::REX2-M1I-GFP::TRP1; HTB1-mCherry-HISMX6; 2micron_plasmid_KanMX_pGPD1-Su9-BFP
34399	MATa/alpha trp1::REX2-M41I-GFP::TRP1; HTB1-mCherry-HISMX6; 2micron_plasmid_KanMX_pGPD1-Su9-BFP
34400	MATa/alpha trp1::GRX2-GFP::TRP1; his3:pAro10-mCherry-HDEL::HIS3; 2micron_plasmid_KanMX_pGPD1-Su9-BFP
34401	MATa/alpha trp1::GRX2-M1I-M35I-GFP::TRP1; his3:pAro10-mCherry-HDEL::HIS3; 2micron_plasmid_KanMX_pGPD1-Su9-BFP

Deletion strains were created using pÜB81, and C-terminal 3V5 or FLAG-tagged strains were generated via Pringle tagging at the endogenous locus using pÜB81 or pÜB166 (Longtine et al., 1998). GFP-tagged strains for microscopy were generated using PmeI-digested single-integration plasmids constructed via Gibson assembly of PCR-amplified fragments containing the ORF of interest along with its own 5' leader region amplified from genomic DNA and backbone fragments containing either a GFP tag, *ADH1* terminator, and a *TRP1* selection marker (pÜB629) or an mCherry tag, *ADH1*

terminator, and a *HIS3* selection marker (pÜB1736). Start codon mutants were generated from single-integration plasmids described above by PCR amplifying fragments for Gibson assembly using primers containing the desired point mutation. Pus1^{truncation} overexpression strains were generated following the WTC_{846} system (Azizoglu et al., 2021). The truncated open reading frame sequence was inserted into a single-integration plasmid downstream of pTetO7.1 by Gibson assembly of PCR fragments from genomic DNA and backbone fragments from pUB2344. Transformants were crossed into strains containing pRNR2-TetR-Tup1 and pTetO7.1-TetR.

3.4.2 Yeast growth and sporulation

For vegetative experiments, strains were grown in YEPD at 30°C. Strains were inoculated and grown overnight to reach saturation (OD600 > 10), then back-diluted to an OD600 of 0.2 and grown to desired OD600. For meiotic time courses, strains were inoculated into YEPD supplemented with uracil and tryptophan (1% yeast extract, 2% peptone, 2% glucose, 22.4 mg/L uracil, and 80 mg/L tryptophan) and grown for 24h at RT to an OD600 ≥ 10, then diluted to an OD600 of 0.25 in buffered YTA (1% yeast extract, 2% bacto tryptone, 1% potassium acetate, and 50 mM potassium phthalate) and grown for 16h at 30°C to an OD_{600} ≥ 5. Cells were spun down and washed once with sterile MilliQ water before resuspension in sporulation media (SPO; 2% potassium acetate supplemented with amino acids (40 mg/L adenine, 40 mg/L uracil, 10 mg/L histidine, 10 mg/L leucine and 10 mg/L tryptophan)) at OD600 = 1.85 and shaken at 30°C, with timepoints collected at times indicated in figures.

3.4.3 Protein extraction and western blotting

Strains were grown in specified media and 2 or 3.3 OD600 equivalents of cells were collected for vegetative and meiotic cultures, respectively. Samples were incubated in 5% TCA for ≥10mins at 4℃ then spun down, washed once with TE, once with acetone, then dried overnight. Pellets were resuspended in 150ul of lysis buffer (50mM Tris-HCl, 1mM EDTA, 3mM DTT, 1.1mM PMSF (Sigma), and 1X cOmplete mini EDTA-free protease inhibitor cocktail (Roche)) and cells were lysed by bead-beating for 5min at RT. SDS loading buffer was added to 1X and samples were incubated at 50°C for 10min and beads were pelleted by centrifugation. Samples were run on a 4-12% Bis-Tris gel at 160V for 5min followed by 200V for 25min. Transfer to nitrocellulose membrane was performed using a semi-dry transfer system (Trans-Blot Turbo, BioRad) with a standard 30 min transfer. The membrane was blocked in 5% milk PBS-T for 1 hour at RT and incubated in primary antibody overnight at 4°C. Primary antibodies were diluted in 5% milk in PBS-T + 0.01% sodium azide (1:2,000 for mouse anti-GFP (Clontech) and mouse anti-3V5 (Invitrogen), 1:1000 for mouse anti-FLAG (Sigma), and 1:10,000 for rabbit anti-hexokinase (Rockland). Membrane was washed 3X in PBS-T then incubated in secondary antibody (1:15,000 anti-mouse 800 and anti-rabbit 680 in LI-COR PBS blocking buffer) for 1 hour at RT, then washed 3X in PBS-T before imaging on the LI-COR Odyssey Imager. Analysis and quantification was performed using ImageStudio Lite software.

3.4.4 Proteasome inhibition
Strains were constructed in a *pdr5*Δ background to confer drug sensitivity. Standard meiosis conditions were used as described above. At the designated time point, cultures were split into two cultures and 100uM MG132 or DMSO (vehicle) was added. Cells were collected for protein extraction and western blot as described above at time points indicated in figures.

3.4.5 Growth in non-fermentable media
Cells were grown to saturation overnight in YEPD then back-diluted to an OD600 of 0.2 in YEPD. At 4h post-dilution, cells were spun down and resuspended in either YEPD (fermentable) or YEPG (non-fermentable). Cells were collected for protein extraction and western blot as described above at time points indicated in figures.

3.4.6 Rapamycin treatment
Cells were grown to saturation overnight in YEPD then back-diluted to an OD600 of 0.2 in YEPD. At 2 hours, cultures were split and treated with either rapamycin (0.2ug/ml or 0.5 ug/ml) or DMSO (vehicle). Cells were collected for protein extraction and western blot as described above at time points indicated in figures.

3.4.7 Low glucose growth
Cells were grown to saturation overnight in YEPD then back-diluted to an OD600 of 0.2 in either YEPD with 2% dextrose (normal) or 0.2% dextrose (low). Cells were collected for protein extraction and western blot as described above at time points indicated in figures.

3.4.8 Pus1^{truncation} overexpression
Cultures were grown to saturation overnight in YEPD then back-diluted to an OD600 of 0.2 in YEPD and treated immediately with either 1ug/ml anhydrotetracycline (aTC) or DMSO (vehicle). Samples were collected 3h post-dilution.

3.4.9 RNA extraction
5ODs of cells were pelleted by centrifugation and flash frozen in liquid nitrogen. Cells were thawed on ice and resuspended in TES buffer (10 mM Tris pH 7.5, 10 mM EDTA, 0.5% SDS). An equal volume of acid phenol (pH4.3, Sigma-Aldrich) was added. Samples were shaken at 1400rpm for 30min at 65°C, then spun down at 4°C. The aqueous phase was transferred to a new tube containing 350ul chloroform. Samples were spun down and the aqueous layer was transferred to a new tube containing 100% isopropanol and with 350mM sodium acetate (pH5.2). Samples were precipitated overnight at -20°C. RNA was pelleted by centrifugation and pellets were washed with 80% ethanol, dried, resuspended in DEPC water for 10min at 37°C. Total RNA was quantified using the Qubit RNA BR Assay Kit (ThermoFisher).

3.4.10 Poly-A selection and RNA-seq
Poly-A selection was performed using the NEXTFLEX Poly(A) Beads 2.0 kit with 5ug total RNA (NOVA-512992). RNA-seq libraries were prepared from the resulting poly-A selected RNA using the NEXTFLEX Rapid Directional RNA-Seq Kit 2.0 (NOVA-5198-

02). Libraries were quantified and quality checked using the Agilent 4200 TapeStation (Agilent Biotechnologies Inc). Samples were sequenced on the NovaSeqX sequencer.

3.4.11 Live imaging

At designated time points 2ul of meiotic culture was placed on a glass slide and imaged immediately. Images were acquired using a DeltaVision Elite wide-field fluorescence microscope (GE Healthcare), a 100X/1.40 NA oil-immersion objective (DeltaVision, GE Healthcare, Sunnyvale, CA), and the following filters: FITC, mCherry, DAPI. 30 z-stacks were collected with 0.2uM spacing. Images were deconvolved using softWoRx imaging software (GE Healthcare).

3.4.12 Sequence alignment, quantification, and differential expression analysis

Sequencing data were aligned to the SK1 genome using STAR. A-site mapping for standard ribosome profiling and TIS-profiling data was performed as previously described (Eisenberg et al., 2020). Differential expression analysis for mRNA-seq data was performed using DESeq2. Hierarchical clustering was performed using complete-linkage clustering on the Pearson correlation of the log2-transformed average of 2 replicates. Genome browser visualization was performed using IGV.

3.4.13 Gene ontology enrichment analysis

GO analysis was performed using the PANTHER classification system (Mi et al., 2013).

3.4.14 Truncation calling algorithm

Analysis was performed on TIS-profiling data collected at 0h, 1.5h, 3h, 4.5h, 6h, 8h, 10h, and 22h after addition to sporulation media (SPO), as well as in vegetative exponential, vegetative saturated growth and a MATa/a non-meiotic starvation control collected at 4.5h in SPO, as described in (Eisenberg et al., 2020). To generate a list of putative truncation-generating start codons, we first found all in-frame start codons within annotated exons. For each potential start codon (ATG) at each timepoint, a “peak sum” was calculated by summing the reads at the three nucleotides corresponding to the start codon. To model the background reads for each gene at each time point, we generated an empirical distribution of peaks sums from sets of three independent nucleotides that were randomly sampled with replacement (10,000x). The empirical p-value for each putative start codon, including annotated start codons, was determined by comparing the peak sum for the codon of interest to the empirical distribution. Annotated and putative truncation start codons were then filtered with the following criteria: p-value ≤0.0015 and >11 reads for at least one nucleotide in the start codon. To be considered in the final set, each truncation was required to be called at 2 or more timepoints. Putative truncations were additionally required to start ≥5aa from the annotated start codon and have an ORF length >10aa. Cases of likely mis-annotation, where the “truncated” isoform is likely the dominant isoform, were also removed; this gene set was generating through computational filtering to identify genes where the annotated isoform was not called followed by manual curation through visualization in a genome browser.

3.4.15 Ribosome profiling metagene analysis

Reads were averaged across all truncations at all positions between -50bp and +100bp surrounding truncated isoform TISs. We excluded truncations that begin within 50bp of the annotated isoform to avoid including reads associated with annotated start peaks that would confound the 5' signal. To prevent the profile from being overpowered by single highly expressed genes, we excluded genes with a Z-score greater than 10 at any position.

3.4.16 TL-seq metagene analysis

Reads were summed across all truncations at all positions between -200bp and +200bp surrounding annotated or truncated isoform TISs. We excluded truncations that begin within 200bp of the annotated isoform to avoid including reads associated with annotated TSSs. To prevent the profile from being overpowered by single highly expressed genes, we excluded genes with a Z-score greater than 10 at any position.

3.4.17 TL-seq peak calling

Counts per site were extracted from published bigwig files using custom scripts (Chia et al., 2021). To call protein isoforms with 5' transcript ends upstream of their start codons, for each gene an upstream-to-downstream ratio was calculated, such that { ratio = sum(reads 200bp upstream)/sum(reads 200bp downstream) }. Each gene's upstream-to-downstream ratio was compared to an empirical distribution of 10,000 random ratios obtained by taking the ratio of the sums of two randomly sampled groups of 200 sites within the gene, sampled with replacement. Reads upstream of the annotated TIS were masked to avoid including reads derived from 5' ends of annotated transcript isoforms. A p-value was calculated using the empirical cumulative distribution function of these ratios, and a p-value cutoff of 0.1 was used. To exclude genes with very sparse or no coverage we additionally required a variance greater than 0.05 for the distribution of sample ratios.

3.4.18 Staging comparison between TIS-profiling and TL-seq time courses

Stage matching between time courses was performed using an mRNA-seq time course collected in parallel with the TL-seq time course and an mRNA-seq time course collected under matched strain and growth conditions as the TIS-profiling time course. Note that staging was performed differently for the two time courses – the TIS-profiling was synchronized naturally via starvation conditions, whereas the TL-seq time course was synchronized via inducible expression of meiotic master regulator transcription factors *IME1* and *NDT80*. See (Cheng et al., 2018; Chia et al., 2021) for details. Timepoints for each time course were split into either early-meiotic (TIS-profiling: 0h, 1.5h, 3h, 4.5h; TL-seq: 0h, 2h, 3h, 4h, 5h, 6h) or late-meiotic (TIS-profiling: 4.5h, 6h, 8h, 10h; TL-seq: 5h, 6h, 7h, 8h, 9h) based on Pearson correlation of log2-transformed RPKMs and expression patterns of key meiotic genes in the mRNA-seq time courses (Figure S3A-B).

3.4.19 Localization prediction

DeepLoc1.0 was run on the amino acid sequence of all called truncated isoforms as well as their corresponding annotated isoform (Almagro Armenteros et al., 2017). As a

control, we generated sets of simulated truncations by randomly sampling in-frame ATGs within annotated genes that are not called as real start sites. To ensure that the length distribution of the control set approximately matched the set of real truncations, for each real truncation we randomly sampled an in-frame start with the distance from annotated isoform within +/- 5 amino acids of that of the real truncation.

3.4.20 Resource availability

All reagents used in this study are available upon request from the corresponding author. Sequencing data will be made available at NCBI GEO. Custom analysis code will be made available on GitHub.

3.5 Supplemental Figures

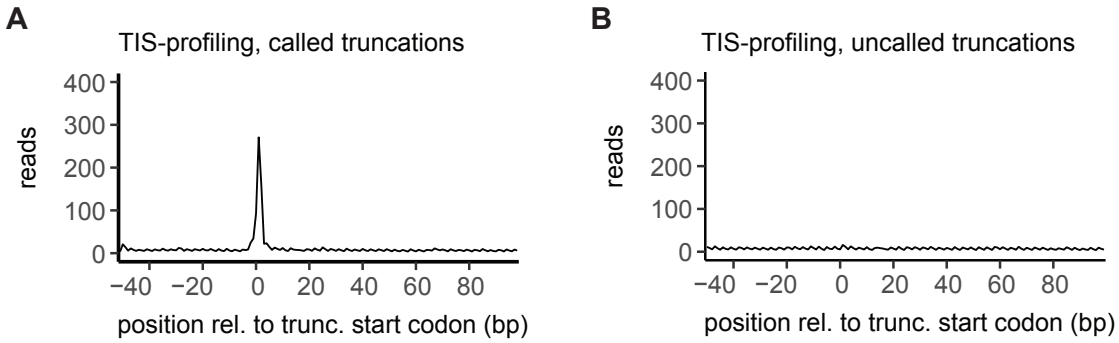


Figure S3.1 Metagene plots of TIS-profiling data for truncated isoforms
(A) Metagene plot of TIS-profiling data for all called truncated isoforms for the region between -50 and +100bp relative to the truncation start codon. Reads are summed across all timepoints.
(B) Metagene plot of TIS-profiling data for all uncalled truncated isoforms for the region between -50 to +100bp relative to the truncation start codon. Reads are summed across all timepoints.

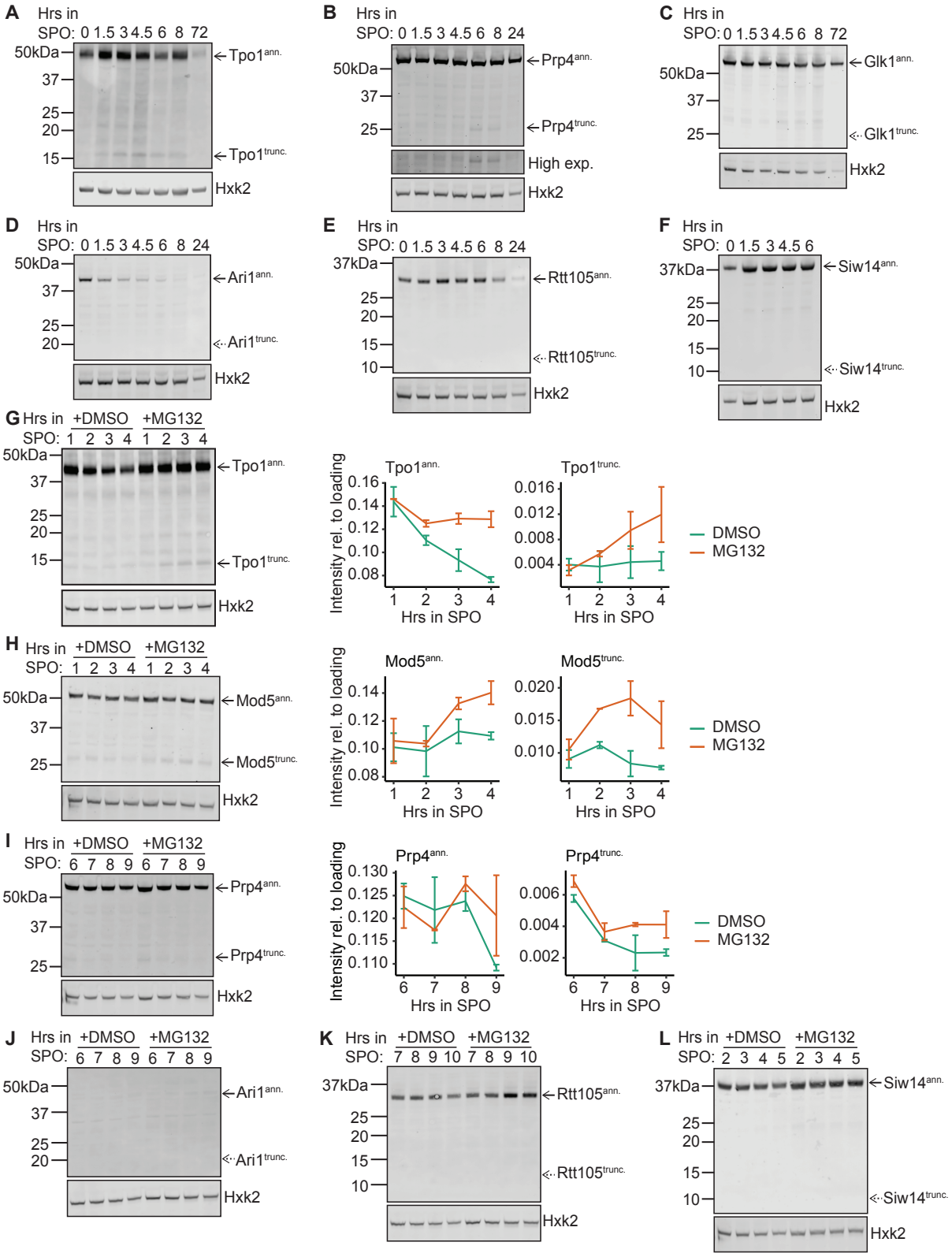


Figure S3.2 Western blots of additional tagged truncations and proteasome inhibition experiments
(A) Western blot of samples collected at various timepoints after transfer of cells to sporulation media (SPO). Hexokinase (Hkx2) is shown as a loading control. A high exposure (high exp.) panel is included for lowly expressed truncations. A different strain is used in each case, expressing the indicated C-

terminal epitope-tagged protein and enabling detection of annotated and truncated isoforms for (A) Tpo1-3V5,
(B) Prp4-3V5,
(C) Glk1-3V5,
(D) Ari1-3V5,
(E) Rtt105-3V5,
(F) Siw14-3V5,
(G) Representative western blot (left) and quantification (right) for cells treated with proteasome inhibitor MG132 or vehicle control DMSO. Quantification is based on 2 replicates and error bars represent standard error. Hexokinase (Hkx2) is shown as a loading control. Blots show stabilization of truncated isoforms for (G) Tpo1-3V5,
(H) Mod5-3V5,
(I) Prp4-3V5,
(J) Western blot analysis for cells treated with proteasome inhibitor MG132 or vehicle control DMSO. Hexokinase (Hkx2) is shown as a loading control. Blots show lack of stabilization of truncated isoforms for (J) Ari1-3V5,
(K) Rtt105-3V5,
(L) Siw14-3V5

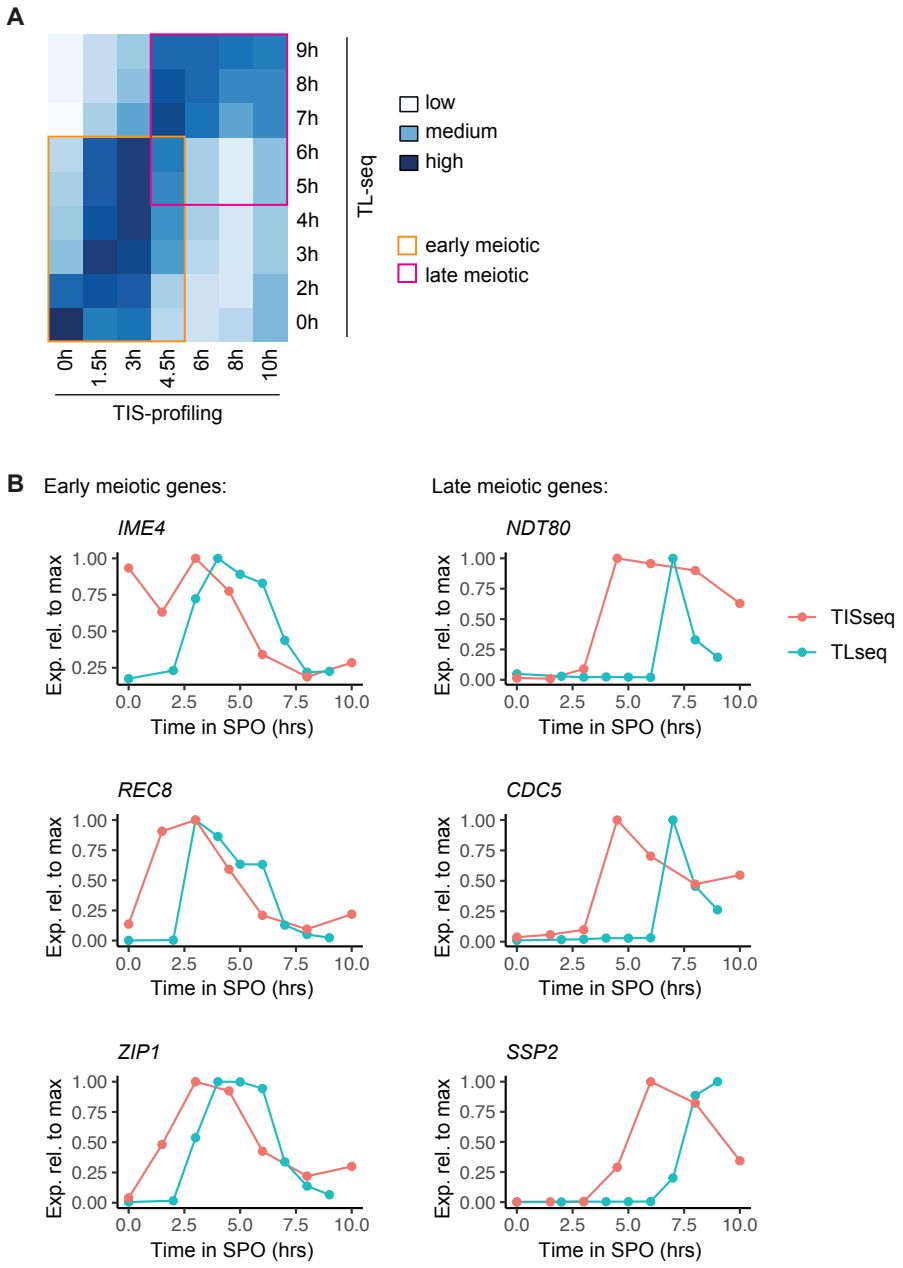


Figure S3.3 Time point matching between TL-seq and TIS-profiling datasets
(A) Heatmap of Pearson correlations between the two mRNA-seq time courses collected in parallel with the TL-seq and TIS-profiling data, used to compare meiotic staging between the two time courses. Meiotic time points are labeled along the x and y axes. Early and late meiotic time point groups are boxed in orange and pink, respectively.
(B) Plots of the expression relative to max for example early (left) and late (right) meiotic genes from mRNA-seq time courses collected in parallel with the TL-seq and TIS-profiling data.

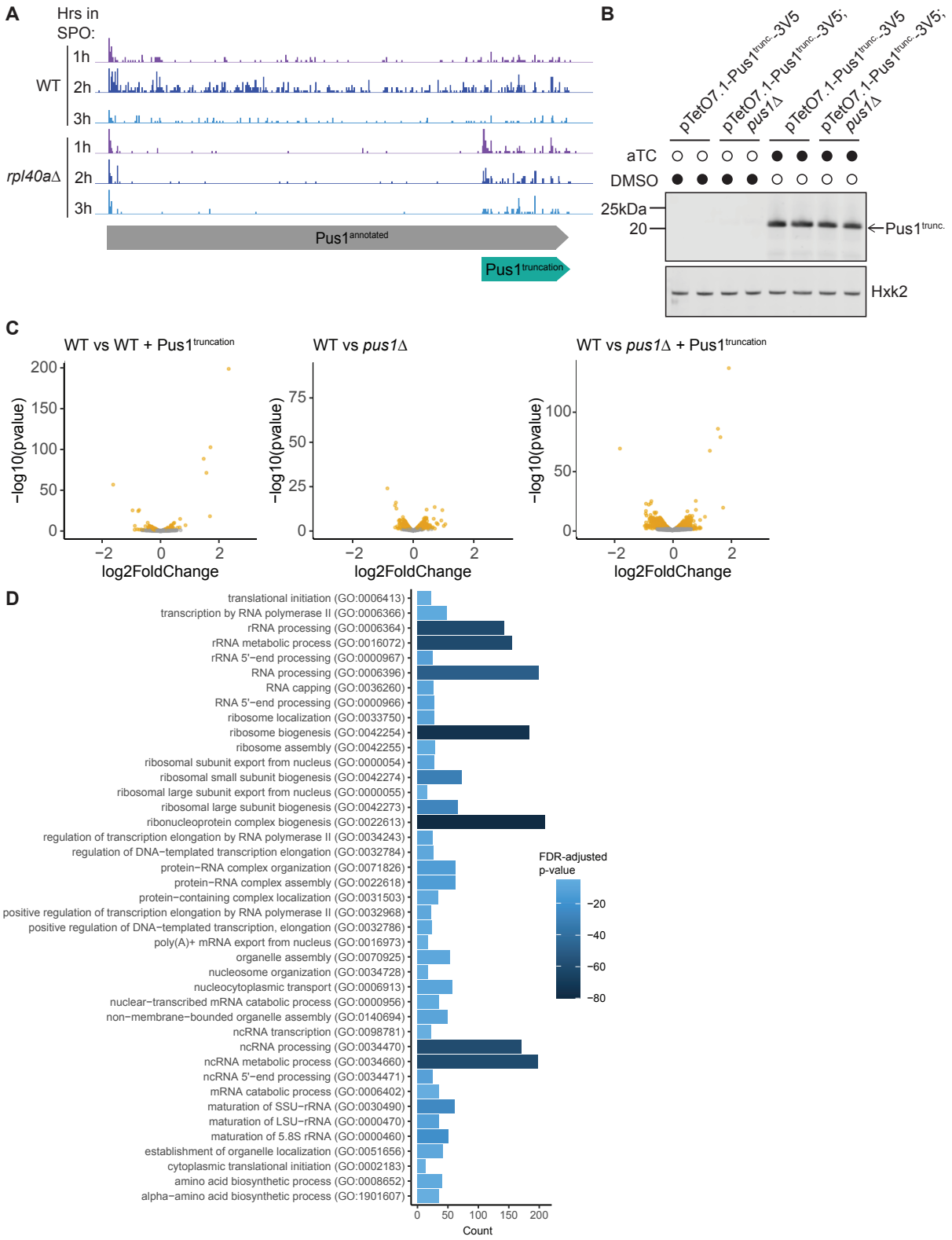


Figure S3.4 Pus1^{truncation} is naturally expressed in *rpl40a*Δ cells and its expression in vegetative exponential cells has broad effects on gene expression

(A) Standard ribosome profiling data of WT and *rpl40a*Δ cells at the *PUS1* locus. Cartoons below represent the annotated and truncated isoforms of Pus1. Samples were collected at indicated timepoints following transfer to sporulation media (SPO).
(B) Western blot confirming expression of Pus1^{truncation} upon aTC treatment. Samples were collected in WT and *pus1*Δ cells carrying a construct to allow anhydrotetracycline-inducible Pus1$^{trunc.}$ expression, treated with either vehicle (DMSO) or aTC.
(C) Volcano plot of DESeq2 analysis of mRNA-seq data for the following conditions: WT cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with aTC ("WT + Pus1$^{trunc.}$"), *pus1*Δ cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with vehicle ("*pus1*Δ") or aTC ("*pus1*Δ + Pus1$^{trunc.}$"). In all cases differential expression is relative to WT cells carrying a construct to allow aTC-inducible Pus1$^{trunc.}$ expression and treated with vehicle ("WT"). Points for significantly differentially expressed genes, as called by DESeq2 (p-adj < 0.1), are yellow and non-significant genes are gray.
(D) Top hits from GO term analysis of significantly downregulated genes between WT and *pus1*Δ + Pus1^{truncation} cells.

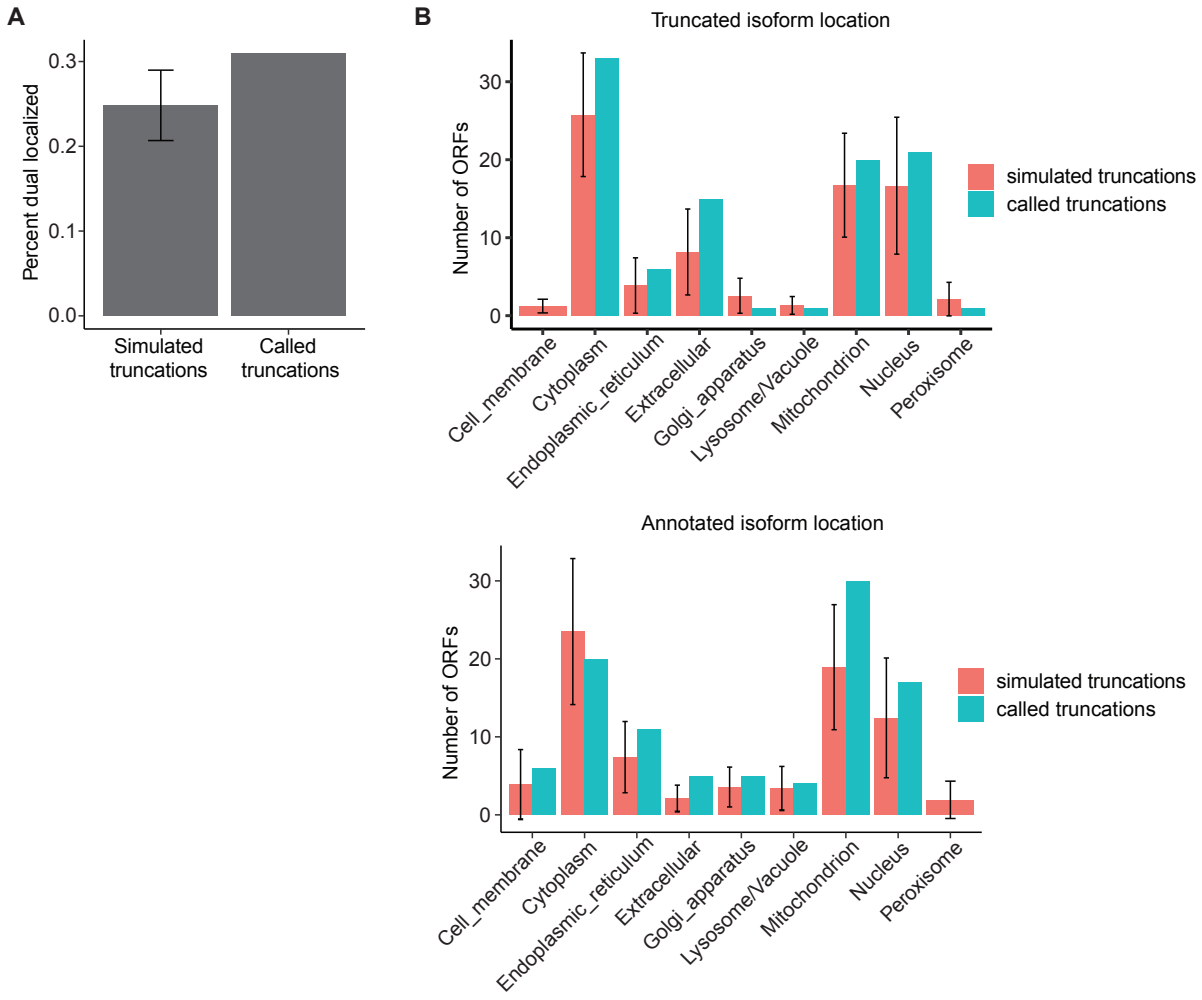


Figure S3.5 Differential localization of truncated protein isoforms
(A) Bar plot of percent of truncated isoforms that are differentially localized relative to their annotated isoform, compared to the percent of simulated truncations (randomly sampled in-frame start codons). Error bar represents 2 standard deviations.
(B) Bar plot of number of truncated (upper) or annotated isoforms (lower) localized to each subcellular compartment, compared to simulated truncations (randomly sampled in-frame start codons). Error bars represent 2 standard deviations.

A

Gene	Predicted localization	Notes
APE4	Mitochondria / Cytoplasm	Not detectable by microscopy
ATH1	Golgi / Extracellular	
BNA3	Mitochondria / Cytoplasm	
GRX2	ER / Mitochondria / Cytoplasm	
IGO2	Cytoplasm / Nucleus	Truncation not detectable by western or microscopy
REX2	Mitochondria / Nucleus	

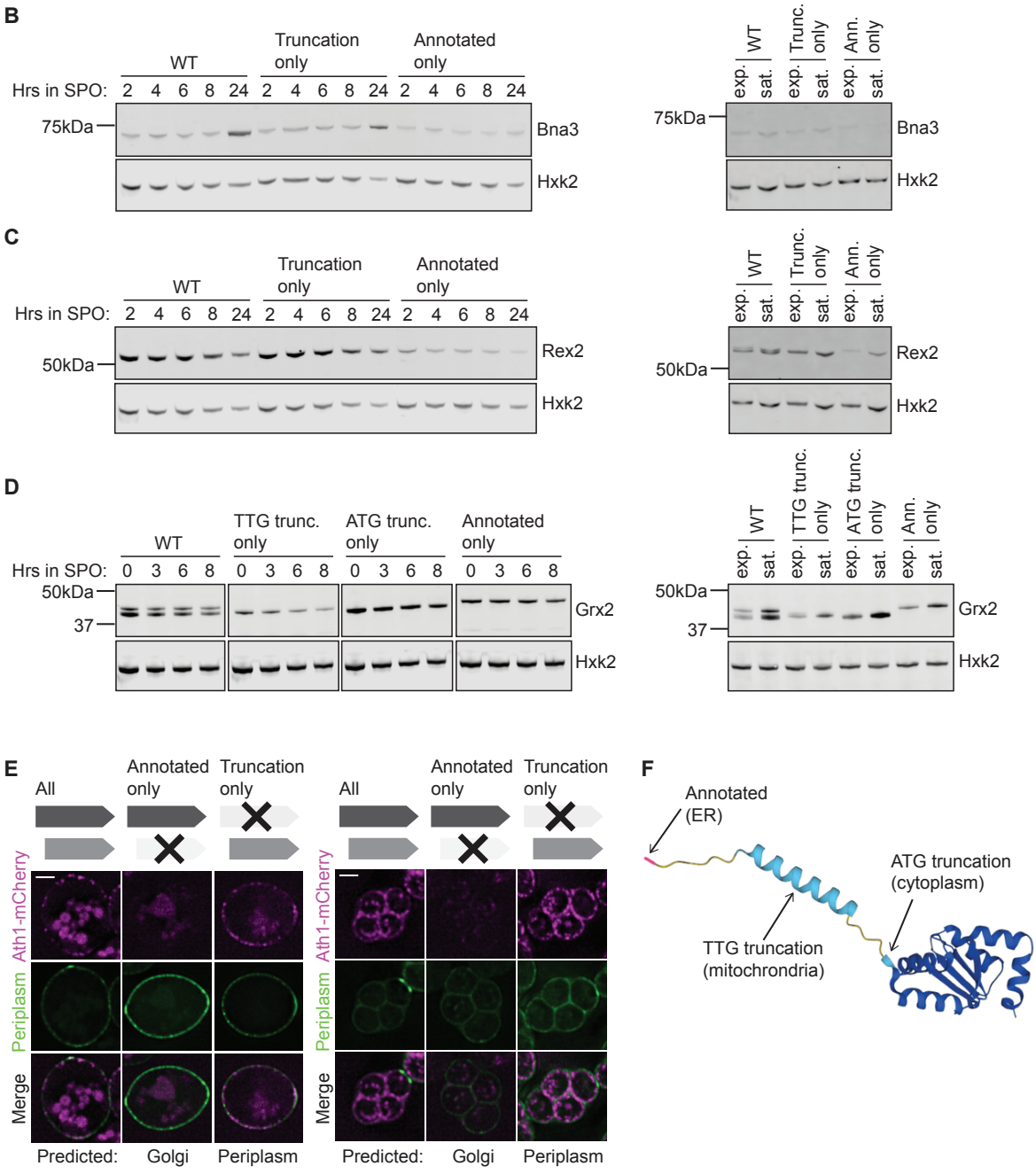


Figure S3.6 Western blots of microscopy validation strains and microscopy data for Ath1
(A) Table of candidates chosen for validation by microscopy, including their predicted localizations (annotated / truncation). Green represents localization predictions that were successfully validated, black represents unvalidated predictions.

(B) Western blot of samples collected at various timepoints after transfer of cells to sporulation media (SPO) or in vegetative exponential or vegetative saturated growth for strains described in Figure 3.6A. Hexokinase (Hxk2) is shown as a loading control. Blots are for (B) Bna3-GFP,
(C) Rex2-GFP,
(D) Grx2-GFP, for strains described in Figure 3.6E
(E) Fluorescence microscopy of C-terminally mCherry-tagged Ath1, collected at 3h (left) and 24h (right) in SPO, using the approach outlined in (6A). Periplasmic localization is indicated by Suc2-GFP. DeepLoc1.0-based predictions are shown below, constructs analyzed in each column are schematized above. Scale bar is 2μm.
(F) Alpha fold structural prediction for full-length Grx2, with arrows indicating the residues corresponding to the start codons of the annotated (ER), TTG truncation (mitochondrial), and ATG truncation (cytoplasmic) isoforms.

Chapter 4: Conclusions and Future Directions

Portions of this chapter were adapted from the following publication:
Higdon, A.L., Brar, G.A., 2020. Rules are made to be broken: a “simple” model organism reveals the complexity of gene regulation. Current Genetics 1–8. https://doi.org/10.1007/s00294-020-01121-8

4.1 What makes an ORF? Working towards a more inclusive definition

This work demonstrates that the true coding capacity of the yeast genome is much larger than previously appreciated and shows the importance of expanding our understanding of what makes a protein coding region. The extended and truncated protein isoforms identified in this work were likely missed previously due to a variety of factors: many are very similar in size to their annotated counterpart, some are produced only in meiosis and at very specific times, they may be produced from near-cognate start codons rather than the typical AUG start (primarily in the case of extensions), and for the yeast genome it has typically been assumed that there is only one protein isoform at a given locus. This reveals two prevalent biases in existing gene annotations: bias towards standard laboratory mitotic growth conditions and bias towards certain rules of translation that were defined by individual studies and then broadly generalized despite known exceptions.

Our concept of what defines an open reading frame is rigid, albeit for good reason. Even with its compact genome, *S. cerevisiae* still has thousands of genes, many of which have not yet been characterized in detail (Wood et al., 2019). To prioritize regions for study, it is useful to use certain rules to predict protein coding regions, namely that they start with an AUG, end with a stop codon, and are of a length capable of producing a stable peptide (reviewed in Dinger et al., 2008). These guidelines have served us well for many years, but with development of technologies for global empirical identification of coding regions, it may be time to revisit these rules to create a more inclusive definition of what constitutes an ORF.

For example, it has become increasingly clear that translation initiation at non-AUG start codons is a biologically relevant way of making protein isoforms (Kearse and Wilusz, 2017). *In vitro* reporter studies have shown that near-cognate initiation, while an order of magnitude less efficient than that at AUGs, can still produce protein (Kolitz et al., 2008). Prior to our work, only a handful of functional extended isoforms had been characterized in single-gene studies, but these include cases with clear and important biological function. The tRNA synthetase gene *ALA1*, for example, uses an upstream ACG codon in addition to an AUG start codon to produce two isoforms that localize to the mitochondria and cytoplasm, respectively, and are necessary for translation in both locations (Tang et al., 2004). In our study, we observed this and numerous other examples of near-cognate initiation, and similar studies in other mammalian systems have also revealed widespread near-cognate initiation (Fields et al., 2015; Ingolia et al., 2011). Although the vast majority of these near-cognate-initiated isoforms remain functionally uncharacterized, their prevalence and usage across very evolutionarily

diverged organisms suggests that near-cognate codons should be considered as possible ORF starts when annotating genes in the future.

Since examples of near-cognate initiation have been known for many years, should near-cognate codons have been included in annotations all along? Unfortunately, in the absence of empirical TIS usage data, it is simply not feasible to do so. A notable pitfall of expanding ORF definitions to include near-cognate start codons is that it creates a much more difficult computational prediction problem, by making the number of potential ORFs unrealistically large. In fact, our TIS-profiling data revealed that very few of the available in-frame near-cognate start codons in 5' leaders are actually used to initiate translation, and the factors contributing to this specificity are still largely unknown. Our study supports a role for eIF5A in modulating near-cognate usage, and other studies have suggested additional factors, like RNA structure, that may facilitate near-cognate initiation (Eisenberg et al., 2020; Guenther et al., 2018; Kozak, 1990). Careful integration of these different types of data, as well as experiments aimed at unraveling the interplay between multiple trans and cis factors, will be important for fully understanding why some start codons are chosen over others. Until this point, we will need to rely on empirical data to know which TISs are used. In turn, these data will likely inform our understanding and ability to predict TIS selection.

In the case of truncated protein isoforms, the start codon is typically a canonical ATG codon rather than a near-cognate. While most genes have multiple in-frame ATGs within the longer annotated ORF, only a small subset are used to produce truncated protein products, making it again important to have empirical data to help differentiate true start sites. Since many truncated proteins are produced from truncated transcripts, their identification can also be aided by the use of genome-wide transcript start site data such as TL-seq. These types of data alone have shown extensive transcript isoform heterogeneity, including during meiosis (Chia et al., 2021; Pelechano et al., 2013). Pairing these data with TIS-profiling data like our own provides additional information about the protein coding capacity of these truncated transcripts.

Empirical data also relieves us of our dependence on length restrictions in coding region prediction. While length cutoffs help significantly enrich for true protein-coding regions, they suffer from both false negatives - often missing smaller protein-coding ORFs - and false positives - erroneously categorizing non-coding RNAs as coding (reviewed in Dinger et al., 2008). The non-coding RNA Xist, for example, was initially thought to code for protein due to a nearly 300aa putative ORF that is in fact not translated (Brockdorff et al., 1992). On the other hand, a few critical proteins from short ORFs are known, including the ribosomal protein gene, *RPL41*, which is 25 codons long and conserved in humans (Suzuki et al., 1990; Yu and Warner, 2001). The largest casualty of length restrictions, however, may not be directly "functional" ORFs, but rather regulatory ones, like uORFs, which are typically very short but nonetheless can have important effects on downstream ORF translation (reviewed in Morris and Geballe, 2000; Renz et al., 2020; Zhang et al., 2019). Comprehensive identification of all translated ORFs, regardless of length, is necessary to create a truly complete genome annotation, whether the ORFs serve regulatory or protein-template function.

4.2 What is “normal”? The power of studying natural stress conditions

Our annotation of the genome and assignment of function to gene products draws heavily from studies of domesticated yeast strains under standard lab conditions. This skews our perception of “functional relevance” or “essentiality” towards nutrient-rich mitotic growth, which differs greatly from the conditions in which the wild ancestors of our domesticated lab strains evolved (reviewed in Engel et al., 2014; Liti, 2015). While truly understanding the evolutionary trajectory and life history of yeasts will require a population genetic approach and study of wild yeast species ecology, we can still glean tremendous insight into the diversity of their gene regulatory mechanisms from studying domesticated yeasts under a broad array of conditions. By collecting TIS-profiling data across a meiotic time course, for example, we were not only able to see dynamic regulation patterns but also detected many protein isoforms that are not produced in vegetative growth conditions.

The true capacity of gene expression regulation cannot be detected within the confines of standard laboratory growth conditions, and in fact, many regulatory strategies that appear illogical or inefficient only begin to make sense in the light of environmental pressures. An example of a seemingly wasteful phenomenon, first discovered in the context of yeast meiosis, is Long Undecoded Transcript Isoform (LUTI) production, which accounts for many of the cases where mRNA and protein levels are decoupled during meiosis (Chen et al., 2017; Cheng et al., 2018). Here, two transcript isoforms are produced from the same locus: a shorter transcript that produces functional protein and a longer (and often abundant) LUTI, whose coding sequence translation is repressed by uORF translation in the extended 5’ leader. This LUTI appears to serve no function beyond the co-transcriptional repression of the shorter transcript conferred by LUTI production (Chia et al., 2017). Making an extra transcript rather than just turning the other off seems wasteful, but in the context of the highly coordinated process of meiotic differentiation, this could provide a handy mechanism for simultaneously activating and inactivating sets of genes with the same transcription factor in a precisely temporally coordinated manner (reviewed in Otto and Brar, 2018; Tresenrider and Ünal, 2018). In another seemingly wasteful phenomenon, during vegetative growth, many transcripts are produced but spliced inefficiently, their intron-contained transcripts degraded, as a way to downregulate genes that are specific to meiosis or response to environmental stresses. This strategy, however, may allow them to remain primed to upregulate production of the spliced transcripts as soon as the necessary cues are in place (Juneau et al., 2007; Pleiss et al., 2007).

Studying stress conditions challenges our assumptions on the “normal” regulation or function of a gene. In our own work, we find ourselves relying on phrases such as “main isoform” or “annotated isoform” to distinguish between the previously known and newly identified isoforms. However, in many cases, we find that the new isoform is in fact more robustly produced, perhaps at more time points or with more dynamic regulation than the annotated isoform, calling into question an easy binary categorization between a “main” and “alternative” isoform. Indeed, across biology, we frequently categorize the

functions of a protein into their “main” and “moonlighting” roles, but “main” often just means the function that was discovered first, is most abundant during “normal” conditions, or has the most conventional regulation. Our increasingly nuanced understanding of transcript and protein isoform production suggests that it may be time to develop a less hierarchical naming system, and perhaps one that incorporates transcript and protein isoforms that serve a regulatory function rather than only a direct protein-template function.

4.3 Final thoughts

By studying the repertoire of proteins produced across the developmental process of meiosis in budding yeast, we have seen cells bend canonical rules of translation to produce an astounding diversity of protein products, especially during times of stress and upheaval. The apparent simplicity of budding yeast makes it an especially useful organism for exploring conserved complexities, and its strengths can complement those of similar efforts in other organisms . Decades of research have built off of certain rules of gene regulation, and even the things produced within that framework are mind-bogglingly complex and beautiful. Looking forward, however, we know that improvements in technology can allow us to go beyond those rules to observe yet more levels of complexity and seek to understand them.

References

Aitken, C.E., Lorsch, J.R., 2012. A mechanistic overview of translation initiation in eukaryotes. Nat. Struct. Mol. Biol. 19, 568–576. https://doi.org/10.1038/nsmb.2303

Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., Winther, O., 2017. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33, 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Amaral, C., Vicente, C.T., Caetano, S.M., Gaspar-Cordeiro, A., Yang, Y., Cloetens, P., Romão, C.V., Rodrigues-Pousada, C., Pimentel, C., 2021. An Internal Promoter Drives the Expression of a Truncated Form of CCC1 Capable of Protecting Yeast from Iron Toxicity. Microorganisms 9, 1337. https://doi.org/10.3390/microorganisms9061337

Ares, M., Grate, L., Pauling, M.H., 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. RNA 5, 1138–1139. https://doi.org/10.1017/S1355838299991379

Azizoglu, A., Brent, R., Rudolf, F., 2021. A precisely adjustable, variation-suppressed eukaryotic transcriptional controller to enable genetic discovery. eLife 10, e69549. https://doi.org/10.7554/eLife.69549

Baralle, F.E., Giudice, J., 2017. Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. 18, 437–451. https://doi.org/10.1038/nrm.2017.27

Basak, A., Query, C.C., 2014. A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast. Cell Rep. 8, 966–973. https://doi.org/10.1016/j.celrep.2014.07.004

Bazykin, G.A., Kochetov, A.V., 2011. Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Res. 39, 567–577. https://doi.org/10.1093/nar/gkq806

Beltzer, J.P., Chang, L.F., Hinkkanen, A.E., Kohlhaw, G.B., 1986. Structure of yeast LEU4. The 5' flanking region contains features that predict two modes of control and two productive translation starts. J. Biol. Chem. 261, 5160–5167.

Beltzer, J.P., Morris, S.R., Kohlhaw, G.B., 1988. Yeast LEU4 encodes mitochondrial and nonmitochondrial forms of alpha-isopropylmalate synthase. J. Biol. Chem. 263, 368–374.

Benne, R., Hershey, J.W., 1978. The mechanism of action of protein synthesis initiation factors from rabbit reticulocytes. J. Biol. Chem. 253, 3078–3087. https://doi.org/10.1016/S0021-9258(17)40805-2

Boguta, M., Hunter, L.A., Shen, W.C., Gillman, E.C., Martin, N.C., Hopper, A.K., 1994. Subcellular locations of MOD5 proteins: mapping of sequences sufficient for targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol. Mol. Cell. Biol. 14, 2298–2306. https://doi.org/10.1128/mcb.14.4.2298-2306.1994

Brar, G.A., Weissman, J.S., 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. 16, 651–664. https://doi.org/10.1038/nrm4069

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., Weissman, J.S., 2012. High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. Science 335, 552–557. https://doi.org/10.1126/science.1215110
Brent, M.R., 2005. Genome annotation past, present, and future: How to define an ORF at each locus. Genome Res. 15, 1777–1786. https://doi.org/10.1101/gr.3866105
Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., Rastan, S., 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71, 515–526. https://doi.org/10.1016/0092-8674(92)90519-i
Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., Gilbert, W.V., 2014. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515, 143–146. https://doi.org/10.1038/nature13802
Carlson, M., Botstein, D., 1982. Two differentially regulated mRNAs with different 5′ ends encode secreted and intracellular forms of yeast invertase. Cell 28, 145–154. https://doi.org/10.1016/0092-8674(82)90384-1
Celik, A., Baker, R., He, F., Jacobson, A., 2017. High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA 23, 735–748. https://doi.org/10.1261/rna.060541.116
Chang, C.-P., Chen, S.-J., Lin, C.-H., Wang, T.-L., Wang, C.-C., 2010. A single sequence context cannot satisfy all non-AUG initiator codons in yeast. BMC Microbiol. 10. https://doi.org/10.1186/1471-2180-10-188
Chang, K.-J., Wang, C.-C., 2004. Translation Initiation from a Naturally Occurring Non-AUG Codon in *Saccharomyces cerevisiae*. J. Biol. Chem. 279, 13778–13785. https://doi.org/10.1074/jbc.M311269200
Chatton, B., Walter, P., Ebel, J.P., Lacroute, F., Fasiolo, F., 1988. The yeast VAS1 gene encodes both mitochondrial and cytoplasmic valyl-tRNA synthetases. J. Biol. Chem. 263, 52–57.
Chen, J., Tresenrider, A., Chia, M., McSwiggen, D.T., Spedale, G., Jorgensen, V., Liao, H., van Werven, F.J., Ünal, E., 2017. Kinetochore inactivation by expression of a repressive mRNA. eLife 6, e27417. https://doi.org/10.7554/eLife.27417
Chen, S.-J., Lin, G., Chang, K.-J., Yeh, L.-S., 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. https://doi.org/10.1074/jbc.M706968200
Cheng, Z., Mugler, C.F., Keskin, A., Hodapp, S., Chan, L.Y.-L., Weis, K., Mertins, P., Regev, A., Jovanovic, M., Brar, G.A., 2019. Small and Large Ribosomal Subunit Deficiencies Lead to Distinct Gene Expression Signatures that Reflect Cellular Growth Rate. Mol. Cell 73, 36-47.e10. https://doi.org/10.1016/j.molcel.2018.10.032
Cheng, Z., Otto, G.M., Powers, E.N., Keskin, A., Mertins, P., Carr, S.A., Jovanovic, M., Brar, G.A., 2018. Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell 172, 910-923.e16. https://doi.org/10.1016/j.cell.2018.01.035
Chia, M., Li, C., Marques, S., Pelechano, V., Luscombe, N.M., van Werven, F.J., 2021. High-resolution analysis of cell-state transitions in yeast suggests widespread

transcriptional tuning by alternative starts. Genome Biol. 22, 34. https://doi.org/10.1186/s13059-020-02245-3
Chia, M., Tresenrider, A., Chen, J., Spedale, G., Jorgensen, V., Ünal, E., van Werven, F.J., 2017. Transcription of a 5' extended mRNA isoform directs dynamic chromatin changes and interference of a downstream promoter. eLife 6, e27420. https://doi.org/10.7554/eLife.27420
Clements, J.M., Laz, T.M., Sherman, F., 1988. Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. Mol. Cell. Biol. 8, 4533–4536. https://doi.org/10.1128/MCB.8.10.4533
Cox, J., 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. https://doi.org/10.1038/nbt.1511
Czudnochowski, N., Wang, A.L., Finer-Moore, J., Stroud, R.M., 2013. In Human Pseudouridine Synthase 1 (hPus1), a C-Terminal Helical Insert Blocks tRNA from Binding in the Same Orientation as in the Pus1 Bacterial Homologue TruA, Consistent with Their Different Target Selectivities. J. Mol. Biol., RNA Metabolism: Interactions Matter 425, 3875–3887. https://doi.org/10.1016/j.jmb.2013.05.014
Danpure, C.J., 1995. How can the products of a single gene be localized to more than one intracellular compartment? Trends Cell Biol. 5, 230–238. https://doi.org/10.1016/S0962-8924(00)89016-9
David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. U. S. A. 103, 5320–5325. https://doi.org/10.1073/pnas.0601091103
Davis, C.A., Grate, L., Spingola, M., Ares Jr., 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. https://doi.org/10.1093/nar/28.8.1700
Diaz de Arce, A.J., Noderer, W.L., Wang, C.L., 2018. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. Nucleic Acids Res. 46, 985–994. https://doi.org/10.1093/nar/gkx1114
Dinger, M.E., Pang, K.C., Mercer, T.R., Mattick, J.S., 2008. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. PLoS Comput. Biol. 4, e1000176. https://doi.org/10.1371/journal.pcbi.1000176
Douglass, S.M., Leung, C.S., Johnson, T.L., 2019. Extensive splicing across the Saccharomyces cerevisiae genome (preprint). Molecular Biology. https://doi.org/10.1101/515163
Eisenberg, A.R., Higdon, A.L., Hollerer, I., Fields, A.P., Jungreis, I., Diamond, P.D., Kellis, M., Jovanovic, M., Brar, G.A., 2020. Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast. Cell Syst. 11, 145-160.e5. https://doi.org/10.1016/j.cels.2020.06.011
Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., Weng, S., Wong, E.D., Lloyd, P., Skrzypek, M.S., Miyasato, S.R., Simison, M., Cherry, J.M., 2014. The Reference

Genome Sequence of Saccharomyces cerevisiae: Then and Now. G3 GenesGenomesGenetics 4, 389–398. https://doi.org/10.1534/g3.113.008995
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., Regev, A., Weissman, J.S., 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol. Cell 60, 816–827. https://doi.org/10.1016/j.molcel.2015.11.013
Fresno, M., Jiménez, A., Vázquez, D., 1977. Inhibition of Translation in Eukaryotic Systems by Harringtonine. Eur. J. Biochem. 72, 323–330. https://doi.org/10.1111/j.1432-1033.1977.tb11256.x
Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J., Brosch, M., 2012. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 22, 2208–2218. https://doi.org/10.1101/gr.139568.112
Fukasawa, Y., Tsuji, J., Fu, S.-C., Tomii, K., Horton, P., Imai, K., 2015. MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites*,. Mol. Cell. Proteomics 14, 1113–1126. https://doi.org/10.1074/mcp.M114.043083
Gammie, A.E., Stewart, B.G., Scott, C.F., Rose, M.D., 1999. The Two Forms of Karyogamy Transcription Factor Kar4p Are Regulated by Differential Initiation of Transcription, Translation, and Protein Turnover. Mol. Cell. Biol. 19, 817–825.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G., 1996. Life with 6000 Genes. Science 274, 546–567. https://doi.org/10.1126/science.274.5287.546
Gregio, A.P.B., Cano, V.P.S., Avaca, J.S., Valentini, S.R., Zanelli, C.F., 2009. eIF5A has a function in the elongation step of translation in yeast. Biochem. Biophys. Res. Commun. 380, 785–790. https://doi.org/10.1016/j.bbrc.2009.01.148
Großhans, H., Lecointe, F., Grosjean, H., Hurt, E., Simos, G., 2001. Pus1p-dependent tRNA Pseudouridinylation Becomes Essential When tRNA Biogenesis Is Compromised in Yeast *. J. Biol. Chem. 276, 46333–46339. https://doi.org/10.1074/jbc.M107141200
Guenther, U.-P., Weinberg, D.E., Zubradt, M.M., Tedeschi, F.A., Stawicki, B.N., Zagore, L.L., Brar, G.A., Licatalosi, D.D., Bartel, D.P., Weissman, J.S., Jankowsky, E., 2018. The helicase Ded1p controls use of near-cognate translation initiation codons in 5′ UTRs. Nature 559, 130–134. https://doi.org/10.1038/s41586-018-0258-0
Guisbert, K.S.K., Zhang, Y., Flatow, J., Hurtado, S., Staley, J.P., Lin, S., Sontheimer, E.J., 2012. Meiosis-induced alterations in transcript architecture and noncoding RNA expression in S. cerevisiae. RNA 18, 1142–1153. https://doi.org/10.1261/rna.030510.111
Gutierrez, E., Shin, B.-S., Woolstenhulme, C.J., Kim, J.-R., Saini, P., Buskirk, A.R., Dever, T.E., 2013. eIF5A Promotes Translation of Polyproline Motifs. Mol. Cell 51, 35–45. https://doi.org/10.1016/j.molcel.2013.04.021

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., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G.P., Haussler, D., Kuhn, R.M., Kent, W.J., 2019. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 47, D853–D858. https://doi.org/10.1093/nar/gky1095

Hanekamp, T., Thorsness, P.E., 1999. YNT20, a bypass suppressor of yme1 yme2, encodes a putative 3'-5' exonuclease localized in mitochondria of Saccharomyces cerevisiae. Curr. Genet. 34, 438–448. https://doi.org/10.1007/s002940050418

Harris, R.S., n.d. Improved pairwise alignment of genomic DNA (Ph.D.). The Pennsylvania State University, United States -- Pennsylvania.

He, S., Bystricky, K., Leon, S., François, J.M., Parrou, J.L., 2009. The Saccharomyces cerevisiae vacuolar acid trehalase is targeted at the cell surface for its physiological function. FEBS J. 276, 5432–5446. https://doi.org/10.1111/j.1742-4658.2009.07227.x

Henderson, A., Hershey, J.W., 2011. Eukaryotic translation initiation factor (eIF) 5A stimulates protein synthesis in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. 108, 6415–6419. https://doi.org/10.1073/pnas.1008150108

Hershey, J.W.B., Smit-McBride, Z., Schnier, J., 1990. The role of mammalian initiation factor eIF-4D and its hypusine modification in translation. Biochim. Biophys. Acta BBA - Gene Struct. Expr. 1050, 160–162. https://doi.org/10.1016/0167-4781(90)90159-Y

Heublein, M., Ndi, M., Vazquez-Calvo, C., Vögtle, F.-N., 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. https://doi.org/10.1016/j.jmb.2019.02.023

Hinnebusch, A.G., 2011. Molecular Mechanism of Scanning and Start Codon Selection in Eukaryotes. Microbiol. Mol. Biol. Rev. 75, 434–467. https://doi.org/10.1128/MMBR.00008-11

Homann, O.R., Johnson, A.D., 2010. MochiView: Versatile software for genome browsing and DNA motif analysis. BMC Biol. 8. https://doi.org/10.1186/1741-7007-8-49

Hood, H.M., Neafsey, D.E., Galagan, J., Sachs, M.S., 2009. Evolutionary Roles of Upstream Open Reading Frames in Mediating Gene Regulation in Fungi. Annu. Rev. Microbiol. 63, 385–409. https://doi.org/10.1146/annurev.micro.62.081307.162835

Hossain, M.A., Rodriguez, C.M., 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. https://doi.org/10.1093/nar/gkr497

Huang, J., Reggiori, F., Klionsky, D.J., 2007. The transmembrane domain of acid trehalase mediates ubiquitin-independent multivesicular body pathway sorting. Mol. Biol. Cell 18, 2511–2524. https://doi.org/10.1091/mbc.e06-11-0995

Hug, N., Longman, D., Cáceres, J.F., 2016. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. 44, 1483–1495. https://doi.org/10.1093/nar/gkw010

Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. 15, 205–213. https://doi.org/10.1038/nrg3645
Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S., 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324, 218–223. https://doi.org/10.1126/science.1168978
Ingolia, N.T., Lareau, L.F., Weissman, J.S., 2011. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell 147, 789–802. https://doi.org/10.1016/j.cell.2011.10.002
Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F., Baranov, P.V., 2011. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. 39, 4220–4234. https://doi.org/10.1093/nar/gkr007
Ivanov, I.P., Shin, B.-S., Loughran, G., Tzani, I., Young-Baird, S.K., Cao, C., Atkins, J.F., 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. https://doi.org/10.1016/j.molcel.2018.03.015
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2
Juneau, K., Nislow, C., Davis, R.W., 2009. Alternative Splicing of PTC7 in *Saccharomyces cerevisiae* Determines Protein Localization. Genetics 183, 185–194. https://doi.org/10.1534/genetics.109.105155
Juneau, K., Palm, C., Miranda, M., Davis, R.W., 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc. Natl. Acad. Sci. 104, 1522–1527. https://doi.org/10.1073/pnas.0610354104
Kaiser, C.A., Botstein, D., 1990. Efficiency and diversity of protein localization by random signal sequences. Mol. Cell. Biol. 10, 3163–3173. https://doi.org/10.1128/MCB.10.6.3163
Karniely, S., Rayzner, A., Sass, E., Pines, O., 2006. α-Complementation as a probe for dual localization of mitochondrial proteins. Exp. Cell Res. 312, 3835–3846. https://doi.org/10.1016/j.yexcr.2006.08.021
Kearse, M.G., Wilusz, J.E., 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 31, 1717–1731. https://doi.org/10.1101/gad.305250.117
Kemper, W.M., Berry, K.W., Merrick, W.C., 1976. Purification and properties of rabbit reticulocyte protein synthesis initiation factors M2Balpha and M2Bbeta. J. Biol. Chem. 251, 5551–5557. https://doi.org/10.1016/S0021-9258(17)33095-8
Khonsari, B., Klassen, R., 2020. Impact of Pus1 Pseudouridine Synthase on Specific Decoding Events in Saccharomyces cerevisiae. Biomolecules 10, 729. https://doi.org/10.3390/biom10050729

Kolitz, S.E., Takacs, J.E., Lorsch, J.R., 2008. Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. RNA 15, 138–152. https://doi.org/10.1261/rna.1318509
Kozak, M., 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361, 13–37. https://doi.org/10.1016/j.gene.2005.06.037
Kozak, M., 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299, 1–34. https://doi.org/10.1016/S0378-1119(02)01056-9
Kozak, M., 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234, 187–208. https://doi.org/10.1016/S0378-1119(99)00210-3
Kozak, M., 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc. Natl. Acad. Sci. 87, 8301–8305. https://doi.org/10.1073/pnas.87.21.8301
Kozak, M., 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. 12, 857–872. https://doi.org/10.1093/nar/12.2.857
Kozak, M., 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell 15, 1109–1123. https://doi.org/10.1016/0092-8674(78)90039-9
Kritsiligkou, P., Chatzi, A., Charalampous, G., Mironov, A., Grant, C.M., Tokatlidis, K., 2017. Unconventional Targeting of a Thiol Peroxidase to the Mitochondrial Intermembrane Space Facilitates Oxidative Protein Folding. Cell Rep. 18, 2729–2741. https://doi.org/10.1016/j.celrep.2017.02.053
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923
Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., Qian, S.-B., 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. 109, E2424–E2432. https://doi.org/10.1073/pnas.1207846109
Li, L., Bagley, D., Ward, D.M., Kaplan, J., 2008. Yap5 Is an Iron-Responsive Transcriptional Activator That Regulates Vacuolar Iron Storage in Yeast. Mol. Cell. Biol. 28, 1326–1337. https://doi.org/10.1128/MCB.01219-07
Lin, M.F., Jungreis, I., Kellis, M., 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27, i275–i282. https://doi.org/10.1093/bioinformatics/btr209
Liti, G., 2015. The fascinating and secret wild life of the budding yeast S. cerevisiae. eLife 4, e05835. https://doi.org/10.7554/eLife.05835
Longtine, M.S., Mckenzie III, A., Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., Pringle, J.R., 1998. Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast 14, 953–961. https://doi.org/10.1002/(SICI)1097-0061(199807)14:10<953::AID-YEA293>3.0.CO;2-U
Lopo, A.C., Lashbrook, C.C., Infante, D., Infante, A.A., Hershey, J.W.B., 1986. Translational initiation factors from sea urchin eggs and embryos: Functional properties are highly conserved. Arch. Biochem. Biophys. 250, 162–170. https://doi.org/10.1016/0003-9861(86)90713-7

Lovejoy, A.F., Riordan, D.P., Brown, P.O., 2014. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PloS One 9, e110799. https://doi.org/10.1371/journal.pone.0110799
Machkovech, H.M., Bloom, J.D., Subramaniam, A.R., 2019. Comprehensive profiling of translation initiation in influenza virus infected cells. PLOS Pathog. 15, e1007518. https://doi.org/10.1371/journal.ppat.1007518
Manjunath, H., Zhang, H., Rehfeld, F., Han, J., Chang, T.-C., 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. https://doi.org/10.1016/j.celrep.2019.10.129
Marston, A.L., Amon, A., 2004. Meiosis: cell-cycle controls shuffle and deal. Nat. Rev. Mol. Cell Biol. 5, 983–997. https://doi.org/10.1038/nrm1526
Martin, N.C., Hopper, A.K., 1994. How single genes provide tRNA processing enzymes to mitochondria, nuclei and the cytosol. Biochimie 76, 1161–1167. https://doi.org/10.1016/0300-9084(94)90045-0
Massenet, S., Motorin, Y., Lafontaine, D.L.J., Hurt, E.C., Grosjean, H., Branlant, C., 1999. Pseudouridine Mapping in the Saccharomyces cerevisiae Spliceosomal U Small Nuclear RNAs (snRNAs) Reveals that Pseudouridine Synthase Pus1p Exhibits a Dual Substrate Specificity for U2 snRNA and tRNA. Mol. Cell. Biol. 19, 2142–2154. https://doi.org/10.1128/MCB.19.3.2142
Melnikov, S., Mailliot, J., Shin, B.-S., Rigger, L., Yusupova, G., Micura, R., Dever, T.E., Yusupov, M., 2016. Crystal Structure of Hypusine-Containing Translation Factor eIF5A Bound to a Rotated Eukaryotic Ribosome. J. Mol. Biol., Ribosomes Structure and Mechanisms in Regulation of Protein Synthesis (II) 428, 3570–3576. https://doi.org/10.1016/j.jmb.2016.05.011
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., Carr, S.A., 2013. Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat. Methods 10, 634–637. https://doi.org/10.1038/nmeth.2518
Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8, 1551–1566. https://doi.org/10.1038/nprot.2013.092
Monteuuis, G., Miścicka, A., Świrski, M., Zenad, L., Niemitalo, O., Wrobel, L., Alam, J., Chacinska, A., Kastaniotis, A.J., Kufel, J., 2019. Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. Nucleic Acids Res. 47, 5777–5791. https://doi.org/10.1093/nar/gkz301
Morris, D.R., Geballe, A.P., 2000. Upstream Open Reading Frames as Regulators of mRNA Translation. Mol. Cell. Biol. 20, 8635–8642. https://doi.org/10.1128/MCB.20.23.8635-8642.2000
Motorin, Y., Keith, G., Simon, C., Foiret, D., Simos, G., Hurt, E., Grosjean, H., 1998. The yeast tRNA:pseudouridine synthase Pus1p displays a multisite substrate specificity. RNA N. Y. N 4, 856–869. https://doi.org/10.1017/s1355838298980396
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 320, 1344–1349. https://doi.org/10.1126/science.1158441

Natsoulis, G., Hilger, F., Fink, G.R., 1986. The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of S. cerevisiae. Cell 46, 235–243. https://doi.org/10.1016/0092-8674(86)90740-3
Nishimura, A., Nasuno, R., Yoshikawa, Y., Jung, M., Ida, T., Matsunaga, T., Morita, M., Takagi, H., Motohashi, H., Akaike, T., 2019. Mitochondrial cysteinyl-tRNA synthetase is expressed via alternative transcriptional initiation regulated by energy metabolism in yeast cells. J. Biol. Chem. 294, 13781–13788. https://doi.org/10.1074/jbc.RA119.009203
Otto, G.M., Brar, G.A., 2018. Seq-ing answers: uncovering the unexpected in global gene regulation. Curr. Genet. 64, 1183–1188. https://doi.org/10.1007/s00294-018-0839-3
Outten, C.E., Culotta, V.C., 2004. Alternative start sites in the Saccharomyces cerevisiae GLR1 gene are responsible for mitochondrial and cytosolic isoforms of glutathione reductase. J. Biol. Chem. 279, 7785–7791. https://doi.org/10.1074/jbc.M312421200
Pedrajas, J.R., Porras, P., Martínez-Galisteo, E., Padilla, C.A., Miranda-Vizuete, A., Bárcena, J.A., 2002. Two isoforms of Saccharomyces cerevisiae glutaredoxin 2 are expressed in vivo and localize to different subcellular compartments. Biochem. J. 364, 617–623. https://doi.org/10.1042/BJ20020570
Pelechano, V., Wei, W., Steinmetz, L.M., 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 497, 127–131. https://doi.org/10.1038/nature12121
Pleiss, J.A., Whitworth, G.B., Bergkessel, M., Guthrie, C., 2007. Transcript Specificity in Yeast Pre-mRNA Splicing Revealed by Mutations in Core Spliceosomal Components. PLoS Biol. 5, e90. https://doi.org/10.1371/journal.pbio.0050090
Porras, P., Padilla, C.A., Krayl, M., Voos, W., Bárcena, J.A., 2006. One single in-frame AUG codon is responsible for a diversity of subcellular localizations of glutaredoxin 2 in Saccharomyces cerevisiae. J. Biol. Chem. 281, 16551–16562. https://doi.org/10.1074/jbc.M600790200
Pujol, C., Maréchal-Drouard, L., 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. https://doi.org/10.1016/j.jmb.2007.03.015
Rai, R., Tate, J.J., Georis, I., Dubois, E., Cooper, T.G., 2014. Constitutive and Nitrogen Catabolite Repression-sensitive Production of Gat1 Isoforms. J. Biol. Chem. 289, 2918–2933. https://doi.org/10.1074/jbc.M113.516740
Renz, P.F., Valdivia-Francia, F., Sendoel, A., 2020. Some like it translated: small ORFs in the 5′UTR. Exp. Cell Res. 396, 112229. https://doi.org/10.1016/j.yexcr.2020.112229
Rietzschel, N., Pierik, A.J., Bill, E., Lill, R., Mühlenhoff, U., 2015. The Basic Leucine Zipper Stress Response Regulator Yap5 Senses High-Iron Conditions by Coordination of [2Fe-2S] Clusters. Mol. Cell. Biol. 35, 370–378. https://doi.org/10.1128/MCB.01033-14
Rintala-Dempsey, A.C., Kothe, U., 2017. Eukaryotic stand-alone pseudouridine synthases – RNA modifying enzymes and emerging regulators of gene

expression? RNA Biol. 14, 1185–1196. https://doi.org/10.1080/15476286.2016.1276150
Rose, A.M., Joyce, P.B., Hopper, A.K., Martin, N.C., 1992. Separate information required for nuclear and subnuclear localization: additional complexity in localizing an enzyme shared by mitochondria and nuclei. Mol. Cell. Biol. 12, 5652–5658.
Saini, P., Eyler, D.E., Green, R., Dever, T.E., 2009. Hypusine-containing protein eIF5A promotes translation elongation. Nature 459, 118–121. https://doi.org/10.1038/nature08034
Sapkota, D., Lake, A.M., Yang, W., Yang, C., Wesseling, H., Guise, A., Uncu, C., Dalal, J.S., Kraft, A.W., Lee, J.-M., Sands, M.S., Steen, J.A., Dougherty, J.D., 2019. Cell-Type-Specific Profiling of Alternative Translation Identifies Regulated Protein Isoform Variation in the Mouse Brain. Cell Rep. 26, 594-607.e7. https://doi.org/10.1016/j.celrep.2018.12.077
Schmitt, A.M., Chang, H.Y., 2017. Long Noncoding RNAs: At the Intersection of Cancer and Chromatin Biology. Cold Spring Harb. Perspect. Med. 7, a026492. https://doi.org/10.1101/cshperspect.a026492
Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., Liu, J.O., 2010. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat. Chem. Biol. 6, 209–217. https://doi.org/10.1038/nchembio.304
Schreier, M.H., Erni, B., Staehelin, T., 1977. Initiation of mammalian protein synthesis: I. Purification and characterization of seven initiation factors. J. Mol. Biol. 116, 727–753. https://doi.org/10.1016/0022-2836(77)90268-6
Schuller, A.P., Wu, C.C.-C., Dever, T.E., Buskirk, A.R., Green, R., 2017. eIF5A Functions Globally in Translation Elongation and Termination. Mol. Cell 66, 194-205.e5. https://doi.org/10.1016/j.molcel.2017.03.003
Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., León-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., Fink, G., Regev, A., 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell 159, 148–162. https://doi.org/10.1016/j.cell.2014.08.028
Sing, T.L., Conlon, K., Lu, S.H., Madrazo, N., Morse, K., Barker, J.C., Hollerer, I., Brar, G.A., Sudmant, P.H., Ünal, E., 2022. Meiotic cDNA libraries reveal gene truncations and mitochondrial proteins important for competitive fitness in Saccharomyces cerevisiae. Genetics 221, iyac066. https://doi.org/10.1093/genetics/iyac066
Spealman, P., Naik, A., McManus, J., 2021. uORF-seqr: A Machine Learning-Based Approach to the Identification of Upstream Open Reading Frames in Yeast, in: Labunskyy, V.M. (Ed.), Ribosome Profiling, Methods in Molecular Biology. Springer US, New York, NY, pp. 313–329. https://doi.org/10.1007/978-1-0716-1150-0_15
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., Mann, M., Ingolia, N.T., Weissman, J.S., 2012. Decoding Human Cytomegalovirus. Science 338, 1088–1093. https://doi.org/10.1126/science.1227919

Sugawara, K., Nishiyama, Y., Toda, S., Komiyama, N., Hatori, M., Moriyama, T., Sawada, Y., Kamei, H., Konishi, M., Oki, T., 1992. LACTIMIDOMYCIN, A NEW GLUTARIMIDE GROUP ANTIBIOTIC PRODUCTION, ISOLATION, STRUCTURE AND BIOLOGICAL ACTIVITY. J. Antibiot. (Tokyo) 45, 1433–1441. https://doi.org/10.7164/antibiotics.45.1433

Suomi, F., Menger, K.E., Monteuuis, G., Naumann, U., Kursu, V.A.S., Shvetsova, A., Kastaniotis, A.J., 2014. Expression and Evolution of the Non-Canonically Translated Yeast Mitochondrial Acetyl-CoA Carboxylase Hfa1p. PLoS ONE 9, e114738. https://doi.org/10.1371/journal.pone.0114738

Suzuki, K., Hashimoto, T., Otaka, E., 1990. Yeast ribosomal proteins: XI. Molecular analysis of two genes encoding YL41, an extremely small and basic ribosomal protein, from Saccharomyces cerevisiae. Curr. Genet. 17, 185–190. https://doi.org/10.1007/BF00312608

Suzuki, Y., Onge, R.P.S., Mani, R., King, O.D., Heilbut, A., Labunskyy, V.M., Chen, W., Pham, L., Zhang, L.V., Tong, A.H.Y., Nislow, C., Giaever, G., Gladyshev, V.N., Vidal, M., Schow, P., Lehár, J., Roth, F.P., 2011. Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. Nat. Methods 8, 159–164. https://doi.org/10.1038/nmeth.1550

Tang, H.-L., Yeh, L.-S., Chen, N.-K., Ripmaster, T., Schimmel, P., Wang, C.-C., 2004. Translation of a Yeast Mitochondrial tRNA Synthetase Initiated at Redundant non-AUG Codons. J. Biol. Chem. 279, 49656–49663. https://doi.org/10.1074/jbc.M408081200

Taussig, R., Carlson, M., 1983. Nucleotide sequence of the yeast SUC2 gene for invertase. Nucleic Acids Res. 11, 1943–1954.

Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A.-C., Vagner, S., 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol. Cell 95, 169–178. https://doi.org/10.1016/S0248-4900(03)00033-9

Tresenrider, A., Ünal, E., 2018. One-two punch mechanism of gene repression: a fresh perspective on gene regulation. Curr. Genet. 64, 581–588. https://doi.org/10.1007/s00294-017-0793-5

van Hoof, A., Lennertz, P., Parker, R., 2000. Three conserved members of the RNase D family have unique and overlapping functions in the processing of 5S, 5.8S, U4, U5, RNase MRP and RNase P RNAs in yeast. EMBO J. 19, 1357–1365. https://doi.org/10.1093/emboj/19.6.1357

van Werven, F.J., Amon, A., 2011. Regulation of entry into gametogenesis. Philos. Trans. R. Soc. B Biol. Sci. 366, 3521–3531. https://doi.org/10.1098/rstb.2011.0081

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484

Wogulis, M., Chew, E.R., Donohoue, P.D., Wilson, D.K., 2008. Identification of Formyl Kynurenine Formamidase and Kynurenine Aminotransferase from Saccharomyces cerevisiae Using Crystallographic, Bioinformatic and Biochemical Evidence. Biochemistry 47, 1608–1621. https://doi.org/10.1021/bi701172v

Wolfe, C.L., Lou, Y.C., Hopper, A.K., Martin, N.C., 1994. Interplay of heterogeneous transcriptional start sites and translational selection of AUGs dictate the production of mitochondrial and cytosolic/nuclear tRNA nucleotidyltransferase from the same gene in yeast. J. Biol. Chem. 269, 13361–13366.
Wood, V., Lock, A., Harris, M.A., Rutherford, K., Bähler, J., Oliver, S.G., 2019. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? Open Biol. 9, 180241. https://doi.org/10.1098/rsob.180241
Wu, G., Adachi, H., Ge, J., Stephenson, D., Query, C.C., Yu, Y.-T., 2016a. Pseudouridines in U2 snRNA stimulate the ATPase activity of Prp5 during spliceosome assembly. EMBO J. 35, 654–667. https://doi.org/10.15252/embj.201593113
Wu, G., Radwan, M.K., Xiao, M., Adachi, H., Fan, J., Yu, Y.-T., 2016b. The TOR signaling pathway regulates starvation-induced pseudouridylation of yeast U2 snRNA. RNA 22, 1146–1152. https://doi.org/10.1261/rna.056796.116
Wu, G., Xiao, M., Yang, C., Yu, Y.-T., 2011. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. EMBO J. 30, 79–89. https://doi.org/10.1038/emboj.2010.316
Wu, M., Tzagoloff, A., 1987. Mitochondrial and cytoplasmic fumarases in Saccharomyces cerevisiae are encoded by a single nuclear gene FUM1. J. Biol. Chem. 262, 12275–12282.
Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., Nusbaum, C., Thompson, D.-A., Friedman, N., Regev, A., 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. 106, 3264–3269. https://doi.org/10.1073/pnas.0812841106
Yogev, O., Pines, O., 2011. Dual targeting of mitochondrial proteins: Mechanism, regulation and function. Biochim. Biophys. Acta BBA - Biomembr., Including the Special Section: Protein translocation across or insertion into membranes 1808, 1012–1020. https://doi.org/10.1016/j.bbamem.2010.07.004
Yu, X., Warner, J.R., 2001. Expression of a Micro-protein. J. Biol. Chem. 276, 33821–33825. https://doi.org/10.1074/jbc.M103772200
Zalatan, J.G., Coyle, S.M., Rajan, S., Sidhu, S.S., Lim, W.A., 2012. Conformational Control of the Ste5 Scaffold Protein Insulates Against MAP Kinase Misactivation. Science 337, 1218–1222. https://doi.org/10.1126/science.1220683
Zhang, H., Wang, Y., Lu, J., 2019. Function and Evolution of Upstream ORFs in Eukaryotes. Trends Biochem. Sci. 44, 782–794. https://doi.org/10.1016/j.tibs.2019.03.002
Zhou, S., Sternglanz, R., Neiman, A.M., 2017. Developmentally regulated internal transcription initiation during meiosis in budding yeast. PLOS ONE 12, e0188001. https://doi.org/10.1371/journal.pone.0188001

Appendix

A.1 Interplay between Yap5^{truncation} expression and elevated iron conditions

A.1.1 Induction of Yap5^{truncation} in high iron

Since Yap5^{truncation} contains an Fe-S cluster binding domain, we hypothesized that it could be induced upon elevated iron to sequester Fe-S clusters and mitigate iron toxicity. We performed western blotting in Yap5-FLAG strains in low and high iron conditions and examined the expression of the truncated isoform. We included a Ccc1-3V5 tagged strain as well as a positive control to show that the high iron conditions were effective at causing high-iron response. Ccc1 is a vacuolar iron transporter that pumps iron into the vacuole and is upregulated under elevated iron conditions. We observed the expected upregulation of Ccc1 (Figure A.1, right) and saw upregulation of Yap5^{truncation} as well (Figure A.1, left), indicating that it may be upregulated in response to high iron. These results, however, replicated inconsistently, suggesting that further refinement of the media and treatment conditions may be necessary to consistently observe this effect.

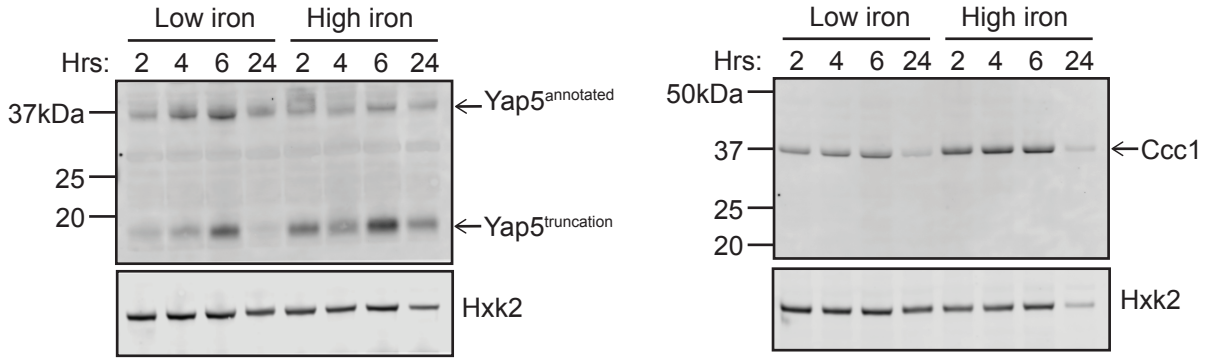


Figure A.1 Yap5^{truncation} may be induced under high iron conditions
Western blot of vegetative samples collected after addition of iron (high iron) or vehicle control (low iron) for strains containing Yap5-FLAG (left) or Ccc1-3V5 (right). Hexokinase (Hxk2) is shown as a loading control. Cells were grown overnight at 30C in YPD then diluted to OD^{600}=0.2 in SC media (6.7g Difco YNB w/o aa, 5g dextrose, 2g US Biological Drop-Out Mix Complete). Either 1000uM iron and 160uM BPS (high iron) or equal volume of 0.1M HCl (vehicle control for iron; low iron) were added immediately upon dilution. Iron media was made using a freshly-prepared 50mM stock solution of ammonium iron(II) sulfate hexahydrate dissolved in 0.1MHCl and a freshly-prepared 1M ascorbate stock solution dissolved in water. Samples were collected for TCA extraction and western blotting at designated time points. Note: to efficiently pellet cells grown in SC, 15mL collection tubes should be washed once with ~5mL of YPD to prevent cells from adhering the walls of the tube. Aspirate to remove all YPD, then add samples and spin down as normal.

A.1.2 Effect of Yap5^{truncation} overexpression on growth in high iron

If the expression of Yap5^{truncation} is part of a cellular response to high iron, it could follow that it is capable of reducing iron toxicity and therefore improving cellular growth in high iron conditions. To test this hypothesis, we grew cells in a plate reader under various

high iron conditions and measured OD^{600} to determine growth rate. Although we were able to see the effects of high iron on growth at the higher concentrations (50uM and 100uM), there was no evidence of a rescue of growth rate upon overexpression of $Yap5^{truncation}$ (Figure A.2).

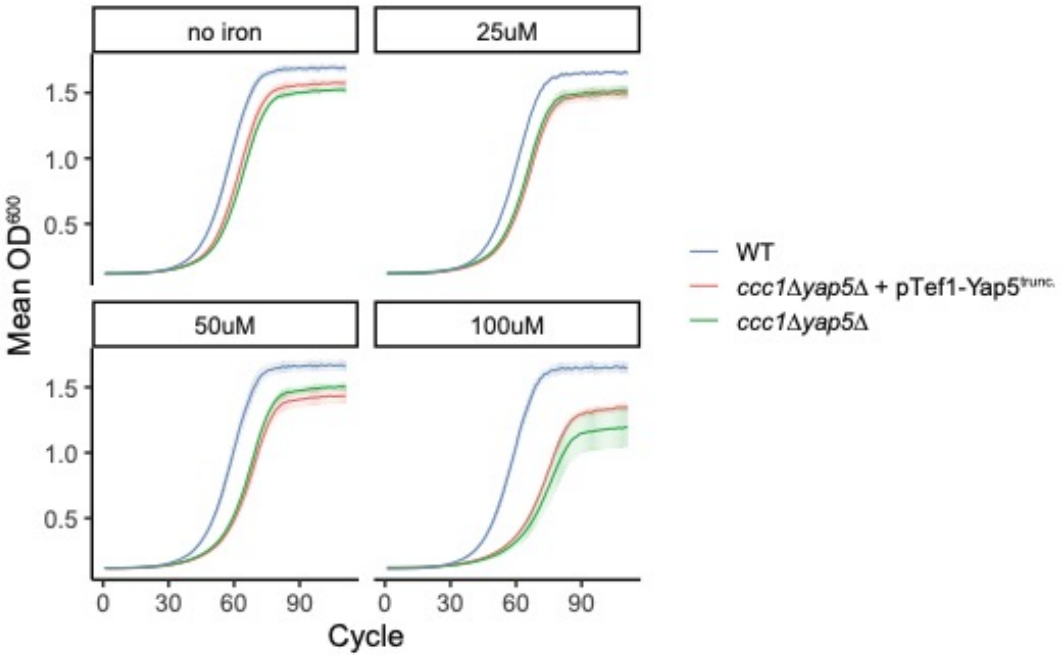


Figure A.2 Overexpression of $Yap5^{truncation}$ does not rescue high iron growth defects
Plate reader measurements of OD^{600} for WT, *ccc1*Δ*yap5*Δ, and *ccc1*Δ*yap5*Δ + pTef1-$Yap5^{truncation}$ overexpression at varying iron concentrations. Cells were grown overnight in at 30C in YPD, then 20-fold diluted into 1800ul of YPD pH4 (low pH improves iron solubility and absorption by cells) and sonicated. 1mL was used to find the OD of diluted culture, then the remainder was used to dilute to OD^{600}=0.005 in 1mL of designated media. 200uL of culture was added to each well of a flat-bottom 96 well plate. 3 replicates were used for each strain under each media condition. Media conditions were: YPD pH4 with 1mM ascorbate plus one of the following iron concentrations: 0uM, 25uM, 50uM, or 100uM. Iron media was made using a freshly-prepared 50mM stock solution of ammonium iron(II) sulfate hexahydrate dissolved in 0.1MHCl and a freshly-prepared 1M ascorbate stock solution dissolved in water.

A.2 Spore viability and sporulation efficiency of $Pus1^{truncation}$ null strains

Among the mitotic and meiotic conditions we assayed, $Pus1^{truncation}$ is notably abundant in spores (Figure 3.2D). To test whether it has a role in efficient sporulation and/or germination, we tested the sporulation efficiency and spore viability of strains lacking $Pus1^{truncation}$ as compared to WT strains (Figure A.3). In both assays, we did not see a significant difference between the mutant and WT, indicating that if $Pus1^{truncation}$ has a functional role, the phenotype is not severe enough to be detected under these conditions. Further work with either sensitized backgrounds or more fine-grained assays (for example RNA-seq) could be worth pursuing.

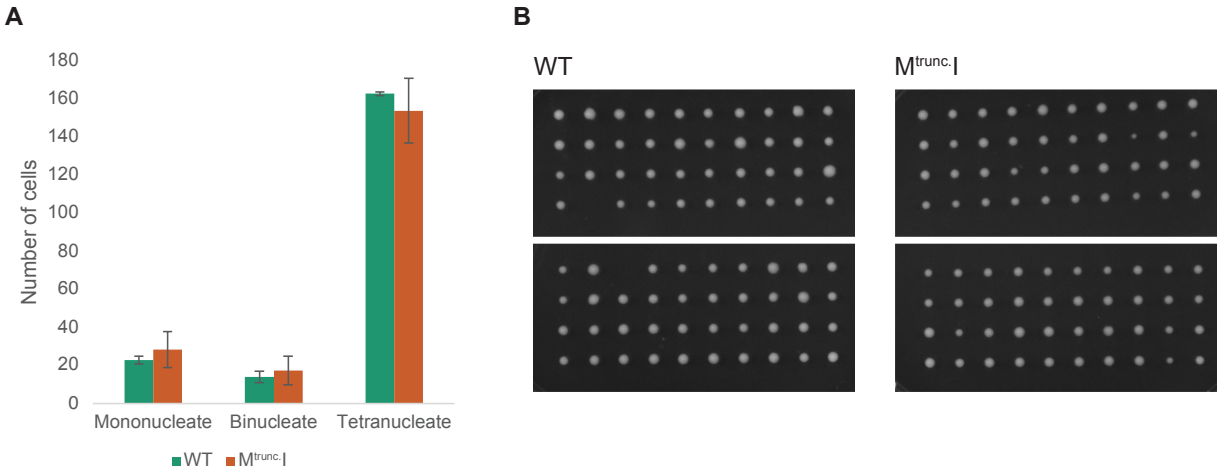


Figure A.3 Strains lacking Pus1^{truncation} do not show sporulation or spore viability defects
(A) Sporulation efficiency for WT (*pus1*Δ; his3::Pus1-3V5::Hyg) and M$^{trunc.}$I (*pus1*Δ; his3::Pus1-M$^{trunc.}$I-3V5::Hyg) strains. Strains were taken through the standard meiosis protocol described in 3.4.2 and counted at 24hrs after addition to SPO. n=3 sets of 100 cells, counted from a single biological replicate for each strain. Error bars represent standard error.
(B) Spore viability for WT (*pus1*Δ; his3::Pus1-3V5::Hyg) and M$^{trunc.}$I (*pus1*Δ; his3::Pus1-M$^{trunc.}$I-3V5::Hyg) strains. Strains were taken through the standard meiosis protocol described in 3.4.2 and dissected on 2%YPD plates at 24hrs after addition to SPO. Each row is four spores from the same tetrad, n=20 tetrads for each strain.