Title
Defining non-canonical modes of gene regulation in budding yeast meiosis

Author
Eisenberg, Amy R

Publication Date
2020

Peer reviewed|Thesis/dissertation

Defining non-canonical modes of gene regulation in budding yeast meiosis

By

Amy R. Eisenberg

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gloria A. Brar, Chair
Professor David Drubin
Professor Nicholas Ingolia
Professor Britt Glaunsinger

Summer 2020

Abstract

Defining non-canonical modes of gene regulation in budding yeast meiosis

by

Amy R. Eisenberg

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Gloria A. Brar, Chair

Gene regulation in budding yeast meiosis is incredibly complex, involving a number of non-canonical strategies. Defining the different modes of regulation is key to fully understanding how the cell functions, especially under changing conditions. Here, we examined two non-canonical gene regulation strategies employed by yeast during meiosis: alternative translation initiation and regulated turnover of ribosomal proteins.

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

In chapter three, we evaluated the role of protein degradation during yeast meiosis, and found that ribosomes are degraded and resynthesized in spores, revealing an interesting mode of regulation of an important molecular machine. Protein degradation is known to be a key component of expression regulation for individual genes, but its global impact on gene expression has been difficult to determine. We analyzed a parallel gene expression dataset of yeast meiotic differentiation, identifying instances of coordinated protein-level decreases to identify new cases of regulated meiotic protein degradation, including degradation of ribosomes. Comparison of protein and translation measurements over time revealed that ribosomes are degraded and resynthesized in spores, making the biological purpose of ribosome resetting an interesting area of follow-up.

TABLE OF CONTENTS

LIST OF FIGURES

Acknowledgements

There are so many people who have helped make this dissertation possible. Some people have contributed in very large and direct ways, while others have had more subtle, but still significant effects. I would like to acknowledge all the individuals who got me to where I am now and contributed to my scientific journey.

I first and foremost need to thank Gloria Brar for being the best mentor, motivator, and role model throughout my entire PhD. Without her guidance and support, the work done here would not have happened. Through the many stages of my project, Gloria always seemed to know exactly what I needed and was incredibly thoughtful and encouraging.

The entire Brar and Ünal lab have been the best team to work with for the last 5 years, and I've enjoyed and appreciated every single lab member's contribution. Having a great place to work with people who care about me has been incredibly important to completing my project, and I am so grateful for having lab mates who are there for me both in science and in life.

I have to thank Ze Cheng, who mentored me when I was a rotation student and when I joined the lab. He taught me everything I needed to know about yeast, and was always able to answer my questions in a way that was precise and efficient. I valued Ze's mentorship greatly, and strive to mentor others as well as he mentored me. Andrea Higdon was one such mentee, who also became an amazing collaborator, as she learned and took on projects that I had no expertise in. Her help was essential, as she did several key analyses that added significantly to my project. I always enjoyed being able to talk through concepts and know that there was never any judgment from her, helping me think expansively and creatively. She has been such a pleasure to work with, and is always so kind and inquisitive.

From joining the lab within months of each other to taking our qualifying exams a day apart, teaching and going to conferences together, eating lunch most days and hanging out most Wednesdays, George Otto has been my most constant and trusty friend, and I've truly loved all the time we've spent together. Helen Vander Wende and Emily Powers have both been so supportive and genuine, and it is incredible how we can so easily switch between talking about science and yoga/skiing (respectively). The two postdocs in our lab, Tina Sing and Ina Hollerer, have been incredibly helpful and knowledgeable, and I really enjoyed our backpacking trips and lab hangouts. Special shout outs also to Vicki Jorgensen for her enthusiasm and love for apples, Kate Morse for going to yoga with me, Jing Chen for so much guidance, Kelsey Van Dalfsen for her positive attitude, Grant King for being so authentic, and Elçin Ünal for being the second best PI, only after Gloria of course. Chris Mugler and Leon Chan were also both incredibly helpful during the first few years of my PhD, and I'd like to thank them for giving me great advice and being willing to talk whenever.

Our lab manager, Christiane Brune, does not get enough credit for keeping the lab running so smoothly, and I want to acknowledge all of the things she does that make it

so we don't even notice that there could be an issue. There are other people who keep everything running smoothly that I'd like to acknowledge, including Berta, the media facility, the store room staff, the building managers, and the cleaning and janitorial staff. Many of the other labs in the building and the yeast community have provided support in the form of reagents, ideas, and overall morale. People in the Rine and Ingolia labs, especially Davis Goodnight, Ellie Bondra, and Paige Diamond, have been super helpful for yeast-related science things, as well as lots of fun things.

I was really fortunate to have the opportunity to mentor two undergraduates, Sarah Guo and Sasha Saias, both of whom were incredibly fast learners, which made me ensure I understood everything I was teaching them. Training the next generation of scientists is very important, and I am thankful that I got the opportunity to contribute to this goal.

I'd like to thank my thesis committee for their insight and expertise, especially Nick Ingolia for help in thinking about all things relating to the ribosome and translation.

The MCB entering class of 2014 has been such a close-knit family since the week we started. Everyone in the class has improved my experience in grad school, especially Charlotte Nixon for being so open and big-hearted, Lydia Lutes for going on adventures and her optimism, Rob Nichols for keeping it real, Fernando Rodriguez-Perez for salsa and being easygoing, Tim Turkalo for being the best brewing partner, and Rose Hill, David McSwiggen and Dan Kramer for making me laugh during follies all of third year.

Shally Margolis deserves so much thanks for keeping me sane and being my go-to person for literally everything. Whether things were good or bad, Shally was the first person I'd tell and we went through a lot of life together doing urban kick and backpacking and valentaco and follies and BUS and marches and so much more.

It is very fitting that I'm acknowledging my beach house roommates as I sit on the beach while spending time with all of them: Janice Chen, Dan Kramer, Perri Callaway and Melissa Metcalf, and honorary beach house resident Orestes Mavrothalassitis. The fact that I am spending time with them, just weeks before finishing my dissertation, says a lot about how close we all are. Each one of them made coming home enjoyable, and I'm appreciative of all the time we've spent together- at beach house, the beach, and all the other places in between.

Throughout writing my dissertation during a pandemic, Robert McDonald has been such a wonderful partner, both encouraging me along the way and giving me a reason to take breaks, and I'm so thankful he's been there to keep me calm and cheerful.

Lastly, I'd like to thank my family, who are so loving and supportive and always believed in me. They are the reason I'm here, and my love for science, as well as my perseverance, was all from them. Also thanks to my sister Emily for being so strong and considerate, and inspiring me to fight for the things I believe in.

CHAPTER 1

General Introduction

1.1 Why do cells need to regulate gene expression?

Proteins perform the functions that every cell needs to survive. The types and amounts of protein in the cell at any given time dictate how the cell grows, develops, and adapts to different conditions. Proteins perform a wide range of cellular roles, with some being required at all times and others only at very specific times. Regulating how each protein is made is incredibly important and is accomplished by a variety of cellular strategies. The available DNA sequence limits the set of proteins that can be made, as the central dogma of biology is that DNA is transcribed into mRNA, which is then translated into protein. This is often thought of as a linear and unchanging series of steps, and although all proteins are indeed made from mRNA, we are learning that this dogma is more intricate and non-linear than previously thought.

We have greatly improved our ability to sequence DNA quickly and with high accuracy, creating a theoretically complete picture of the possible proteins used by the cell. Segments of DNA have been identified as regions that encode proteins, as well as regions that are regulatory, such as promoter regions. Although we need to know the DNA sequence to understand what can be produced, knowing the sequence alone cannot tell us how much will be made, what function it will have, and what the downstream effect on the cell will be. To understand these aspects, we have to look more in depth to decipher the pathways and how the cell coordinates itself. Further, when mutations or other stressors arise, the cell can change and adapt, and by studying what happens in these cases we can better understand disease and misregulation states. Knowing the possible ways the cell can alter its protein content also gives us the power to modulate cells synthetically to perform functions they would otherwise not be capable of.

1.2 Translation and start codon usage

The steps of translation have been well characterized and have provided significant insight into how proteins are produced. Ribosomes are the key molecular machines that recognize the mRNA sequence and translate it into the corresponding amino acids, which are combined to create proteins. The first step of translation is ribosome scanning, where the small subunit of the ribosome scans along an mRNA molecule, looking for the position where translation will begin (Hinnebusch, 2014). Often, this is the first AUG on the transcript that the ribosome encounters. This site is called the translation initiation site (TIS), and when this site is reached the large subunit of the ribosome joins so that translation can begin. Initiation is the step at which the first tRNA molecule brings the first amino acid into the ribosome, which will be the beginning, or N-terminus, of the protein (J. E. Wilson et al., 2000). Various initiation factors, which are other proteins that associate with the ribosome during initiation, aid in the process of

scanning and TIS recognition (Jackson et al., 2010). The subsequent elongation phase involves repeated amino acid addition and tRNA translocation within the ribosome (Dever and Green, 2012). This process is facilitated by a number of elongation factors, some of which use GTP exchange to catalyze the steps of new amino acid incorporation. Each amino acid has a corresponding three-nucleotide mRNA sequence, so the ribosome recognizes and moves along the mRNA in one of the three possible frames that is set relative to the frame of the initiation site. When the ribosome reaches an in-frame stop codon in the mRNA sequence, this signals for the forming peptide to be released, and subsequently for the ribosome to dissociate into its two subunits and fall off the mRNA.

The ribosome is the key component of protein synthesis, and therefore of cell function overall. The proteins that make up the ribosome have been studied in detail, and are tightly regulated in abundance and quality to ensure accurate protein synthesis (D. N. Wilson and Doudna Cate, 2012). The ribosome is also aided by initiation, elongation and termination factors, which are often essential to execute their respective stages of translation. The eukaryotic ribosome is made up of two subunits, the small (40S) and large (60S) subunits, which are combined to make up the full 80S ribosome. There are 80 proteins and 4 ribosomal RNA molecules that make up the eukaryotic ribosome, and structural and functional studies have determined much about this incredible molecular machine (Schmeing and Ramakrishnan, 2009).

The initiation step is especially important and well controlled, as it is the first and committal step to generating a protein. Conformational changes and auxiliary factor binding aid in this step, which take place once the small subunit of the ribosome stops at the TIS (Jackson et al., 2010). Specific steps of initiation have been determined by mutating initiation factors, or alternatively, by inhibiting initiation with drug treatment (Lin et al., 2018; Vazquez-Laslop et al., 2008). Several drugs, including harringtonine and lactimidomycin (LTM), preferentially inhibit the ribosome at the initiation site (Fresno et al., 1977; Sugawara et al., 1992). LTM has a similar structure to cycloheximide (CHX), which inhibits both initiating and elongating ribosomes. However, the additional ring structure on LTM blocks the site of the ribosome where the initial tRNA binds, making it specific to blocking that step of translation (Figure 1.1). These drugs have helped elucidate the conformation of the ribosome at this stage, and the effects of blocking initiation on the cell (Garreau de Loubresse et al., 2014). Inhibiting initiation generally causes severe phenotypes, as protein synthesis is halted.

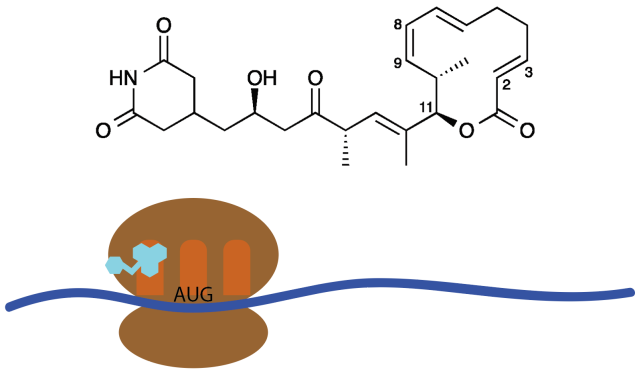

Figure 1.1: The chemical structure of lactimidomycin (LTM), which has a 12-membered lactone ring. LTM binds to the ribosome and inhibits initiation by blocking the site that the initial tRNA binds.

The sites of initiation are most often defined as being at the mRNA sequence AUG (Hinnebusch, 2011). This is primarily because the first amino acid of most proteins, methionine, has a sequence that matches AUG mRNA sequences. It has also been shown that the sequences around the TIS can influence the ribosome's recognition and therefore efficiency of initiation. This sequence, which was identified and investigated in depth by Marilyn Kozak, is called the Kozak context sequence, and can vary slightly in different species (Kozak, 2002; 1984). While many identified proteins have Kozak sequences around them, they do not appear to be required for all proteins, and there are examples of species and proteins that have high efficiency of initiation but no recognized Kozak motif (Li et al., 2017).

While the majority of proteins begin at an AUG initiation site, there are known cases of proteins beginning at non-AUG sites (Kearse and Wilusz, 2017). These are also called near-cognate sites, as they differ by one nucleotide from the cognate start codon, AUG. Certain near-cognate codons have been shown to be used more efficiently than others, but all are still significantly less efficient than AUG (Clements et al., 1988; Zitomer et al., 1984). In yeast, there is only one known example of a gene that has a near cognate (in this case AUU) site as the primary site of initiation for a long protein (Suomi et al., 2014). Several other genes initiate translation at non-AUG sites in addition to the canonical AUG site, such as *ALA1*, the alanyl tRNA synthetase gene. This gene uses an upstream, in-frame ACG site to generate an extended form of the protein (Tang et al., 2004). The extension portion contains a mitochondrial targeting sequence, which is responsible for mitochondrial localization of this isoform, while the canonical form is cytoplasmically localized. This strategy for dual localization by alternative translation initiation sites has only been shown experimentally to occur for a handful of genes, but is a very effective way to generate multiple protein isoforms from the same locus (Heublein et al., 2014; Kritsiligkou et al., 2017; Monteuuis et al., 2019; Tzani et al., 2016).

1.3 Mechanisms of gene regulation

While translation initiation from near-cognate codons is one specific way to generate alternative protein products, the cell uses many other strategies to regulate protein production. An obvious mechanism of control, based solely on the direct mRNA to protein pathway, is to alter the amount of mRNA that is made such that more or less protein can be translated from that mRNA. This can either be done by transcribing mRNA at a different rate or by altering the amount of mRNA degradation. There are many pathways in the cell dedicated to mRNA decay, including specialized pathways such as nonsense-mediated decay, which degrades mRNAs that have stop codons that are earlier than they should be, often due to mutations that generate an early stop codon (Nasif et al., 2018).

While it has been generally thought that an increase in mRNA would lead to an increase in protein, a recently identified mechanism of regulation shows that this is not always the case. In certain contexts, a long undecoded transcript isoform (LUTI) can be

produced that contains an upstream open reading frame (uORF) that is translated instead of the canonical ORF, leading to a decrease in synthesis of the protein from that locus (Jingxun Chen et al., 2017; Cheng et al., 2018; Chia et al., 2017; Hollerer et al., 2019; Jorgensen et al., 2020). This is an unexpected mechanism of controlling protein levels by increasing levels of a corresponding but longer mRNA, highlighting the idea that not all changes at the mRNA level can be assumed to affect the protein level in the same direction. Cells might use this strategy in order to up- and down-regulate different targets simultaneously with the same transcription factor. This has been well studied in the context of yeast meiosis, a developmental program that relies on waves of genes being turned on and off in a coordinated fashion to allow successful production of gametes.

The principle of regulating synthesis and degradation of mRNAs also applies to proteins, where the regulation of each protein's synthesis and degradation leads to the overall level of protein in the cell. Additionally, the ribosome itself can be controlled, where degradation of certain subunits or the whole ribosome can be used to regulate the amount or types of proteins that are made. Modifications to existing proteins can also be used to regulate the function, which can be in the form of chemical, structural, binding state and domain changes. There has also been recent interest in the translation of non-canonical proteins, or proteins that do not look like the typical proteins people have observed classically (Kochetov, 2008; Mouilleron et al., 2015). These include cases where the protein is very short or is made from a DNA region not previously believed to encode a protein (Andrews and Rothnagel, 2014; Jin Chen et al., 2020; Crappé et al., 2013; Delcourt et al., 2017; Slavoff et al., 2013).

1.4 Genome annotations

Historically, the annotation of protein coding regions within genomes has been done computationally (Brent, 2005). There are some rules that were used to determine what makes up an open reading frame (ORF), the region of the genome that encodes a protein. Many of these rules are based on known mechanisms of translation and shared features of known proteins. These rules include that translation begins at an AUG start codon and ends at an in-frame stop codon, which is true for most known proteins. Many initial gene predictions also imposed a length requirement, which was set somewhat arbitrarily at 100 amino acids long (Dinger et al., 2008). While these rules have helped predict many real proteins from the genome sequences provided, it is becoming increasingly clear that there are many exceptions that were not accounted for in prior genome annotations. Some of these exceptions have been identified on an individual basis, where a specific gene was studied and found to act differently than what had been previously thought. Genome-wide experimental studies have also identified more widespread exceptions to the rules regarding what makes an ORF, especially by looking to see which regions of an mRNA are being translated, as opposed to computational predictions. Advances in mass spectrometry, or identifying protein content of cells directly by detecting amino acid sequences, has also helped augment our view of what a protein is (Iwasaki and Ingolia, 2017).

Conservation analysis has also aided in identifying relevant genomic regions, because the regions that change the least over evolutionary time are most likely to encode segments important to the organism. For example, there are over a thousand genes that are conserved between budding yeast and humans (Kachroo et al., 2015). These genes are often essential for cell function, and thus are readily identifiable by aligning the genomes of diverse species. Different species and even different strains within the same species can have different sets of genes, making conservation helpful in identifying many, but not all, important genome features. Understanding the evolutionary relationships between species can inform how genes are likely to have changed, with more closely related species having more similar gene structures.

For any newly identified species, sequencing the genome is the critical, and now very feasible first step to determining the possible proteins that the organism is able to make. Then, computational and conservation analysis is helpful in initially annotating which regions are likely to encode for proteins. While having the genome sequences is necessary for any further determination, it is important to keep in mind that there can still be features that are not detectable without experimental evidence. Additionally, just knowing that the gene can be made into protein doesn't help predict when and how much will be made in the cell in different conditions.

1.5 Ribosome profiling and modifications

The development of ribosome profiling has helped fill these gaps in our understanding of gene expression by allowing experimental determination of where, when, and how much translation is taking place (Ingolia et al., 2009). This strategy relies on the fact that the ribosome strongly protects the piece of mRNA it is translating, such that the cell can be lysed and all of the unprotected RNA degraded by RNAse treatment. Then the remaining ribosome protected fragments (rpfs) can be sequenced, and these sequences can be aligned to the genome, with sequencing reads corresponding to the position and amount of translation across the genome. This protocol was optimized to detect the positions of all ribosomes on their transcripts, and has been used in many conditions leading to a wealth of information that was previously inaccessible (Brar and Weissman, 2015; Jackson and Standart, 2015). Ribosome profiling has addressed questions of how translation changes when cells undergo developmental processes, have mutations, and even when they are infected with coronaviruses (Brar et al., 2012; Irigoyen et al., 2016; Stern-Ginossar and Ingolia, 2015). These studies have given great insight into the changes that take place at a translational level and have led to important discoveries on non-canonical gene regulatory mechanisms.

One such seminal study was performed in budding yeast cells as they undergo meiosis (Brar et al., 2012). Prior to this, meiosis was known to have coordinated transcriptional signaling, with many key genes identified to control the timing, progression and completion of meiosis. By looking at the translational profile of cells during meiosis, however, it was realized that nearly all yeast genes are turned on at some point during the process, and the extent of the regulation was greater than previously appreciated. It was also clear that meiotic cells had increased ribosome densities in regions upstream

of canonical genes, regions often referred to as untranslated regions (UTRs). Much of this translation seems to be occurring at upstream open reading frames (uORFs), which are short sequences that have a stop codon prior to the gene they are upstream of. Translation upstream of canonical genes, if it is in frame and has no stop codon, can also lead to N-terminally extended versions of canonical proteins. Since many genes had several possible locations where translation could be starting that could lead to either uORFs or N-terminal extensions, it was difficult to discern exactly which TISs were being used. Systematic identification of these different ORFs, especially when they are overlapping, was challenging from standard ribosome profiling analysis.

A modification to ribosome profiling that includes use of a translation initiation inhibitor allows for more specific detection of initiation sites without complicating reads from elongating ribosomes from overlapping ORFs (Ingolia et al., 2011). Both LTM and Harringtonine have been used in combination with ribosome profiling for this purpose, and have aided in initiation site mapping in several organisms, including mice and human cell lines (Ingolia et al., 2011; Lee et al., 2012). Both drugs bind to the ribosome and preferentially inhibit initiating ribosomes while elongating ribosomes can run off of transcripts, so that when reads are mapped, they are primarily (or ideally exclusively) at initiation sites. This can then be used to see how initiation site usage is changing under different conditions.

1.6 Budding yeast meiosis as a model system

The single-celled eukaryote budding yeast (*Saccharomyces cerevisiae*) has been used as a model system for over 90 years and has been studied extensively in many aspects (Duina et al., 2014). Yeast have had a much longer shared history with humans, through its use in making beer and bread. Many genetic tools have been developed to manipulate and uncover basic biological principles. Because yeast are eukaryotes, much of the knowledge learned from studies in this relatively simple organism can be used to understand higher eukaryotes, including humans. Approximately a third of human genes have homologs in yeast, and at least 200 human genes are able to function if put in place of their orthologous yeast gene (Kachroo et al., 2015). While many important genes are similar, the yeast genome is less complex, making it straightforward to study gene regulatory mechanisms in a less complicated system. Chiefly, yeast have few genes with even one intron, meaning alternative splicing is not a major factor for most genes (Davis et al., 2000).

In addition to their genomic simplicity, yeast are able to grow quickly, with a doubling time around 1.5 hours. They can also be grown in large quantities, which is beneficial for taking multiple types of measurements on one sample. For instance, from a single culture, mRNA sequencing, ribosome profiling, mass spectrometry and microscopy samples can be taken (Cheng et al., 2018). These samples can also be taken at different time points or conditions, allowing direct comparisons of changes from the same exact population of cells. This has been especially helpful in studying meiosis, where cells can be synchronized such that they complete the stages of cell division at the same times. In comparison to human or mouse meiosis studies which generally

need to be done in single cells, this has made yeast an incredibly useful system for studying meiosis (Susor et al., 2015).

Meiosis is an important process for all sexually reproducing organisms, and understanding where issues can arise is relevant for infertility and chromosome missegregations leading to trisomies, such as trisomy 21 which causes down syndrome (Lamb et al., 2005). In addition to meiosis being important for human health, it is also a system that requires the cell to utilize gene regulatory mechanisms that are not used in other conditions. We and others have observed non-canonical modes of regulation that are enriched in meiosis relative to mitotic growth, including alternative transcription, alternative organelle segregation and translation of non-canonical proteins (Brar et al., 2012; Cheng et al., 2018; Kim Guisbert et al., 2012; King et al., 2019; Sawyer et al., 2019). By identifying and dissecting these mechanisms that the cell has available to use, we can further our understanding of the possible ways to regulate genes. This can then be applied and modulated in other systems to both improve our overall understanding of gene regulation.

CHAPTER 2

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

Part of the work presented in this chapter has previously been published in the following manuscript: Eisenberg, A.R., Higdon, A.L., Hollerer, I., Fields, A.P., Jungreis, I., Diamond, P.D., Kellis, M., Jovanovic, M., and Brar, G.A., 2020. Translation initiation site profiling reveals widespread synthesis of non-AUG-initiated protein isoforms in yeast. *Cell Systems,* 11, 1-16.

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 (*Saccharomyces cerevisiae*) has relatively simple transcript architectures with fewer known cases of complexity, such as from alternative splicing, despite extensive analyses of its transcriptome (Davis et al., 2000; Hossain et al., 2011; Juneau et al., 2009; Kim Guisbert et al., 2012; Yassour et al., 2009). This simple architecture allows for investigation of TISs to be more directly informative, as identification of the start codon alone can generally be used to define an ORF.

We developed a TIS identification approach for budding yeast, both in vegetative and meiotic conditions, with the goal of characterizing ORF types that were previously challenging to identify systematically by standard ribosome profiling. The class of ORFs that we were most interested in assessing, due to their potential to modulate the function of well-characterized genes, were those encoding alternate protein isoforms that result from translation initiation at non-AUG codons upstream of the characterized start codon. Several individual examples of N-terminally extended proteins isoforms have been identified in an ad hoc manner using classical approaches (K.-J. 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 2.S1A; Suzuki et al., 2011). However,

this strain could not efficiently undergo meiosis, precluding its use for our experiments (data not shown).

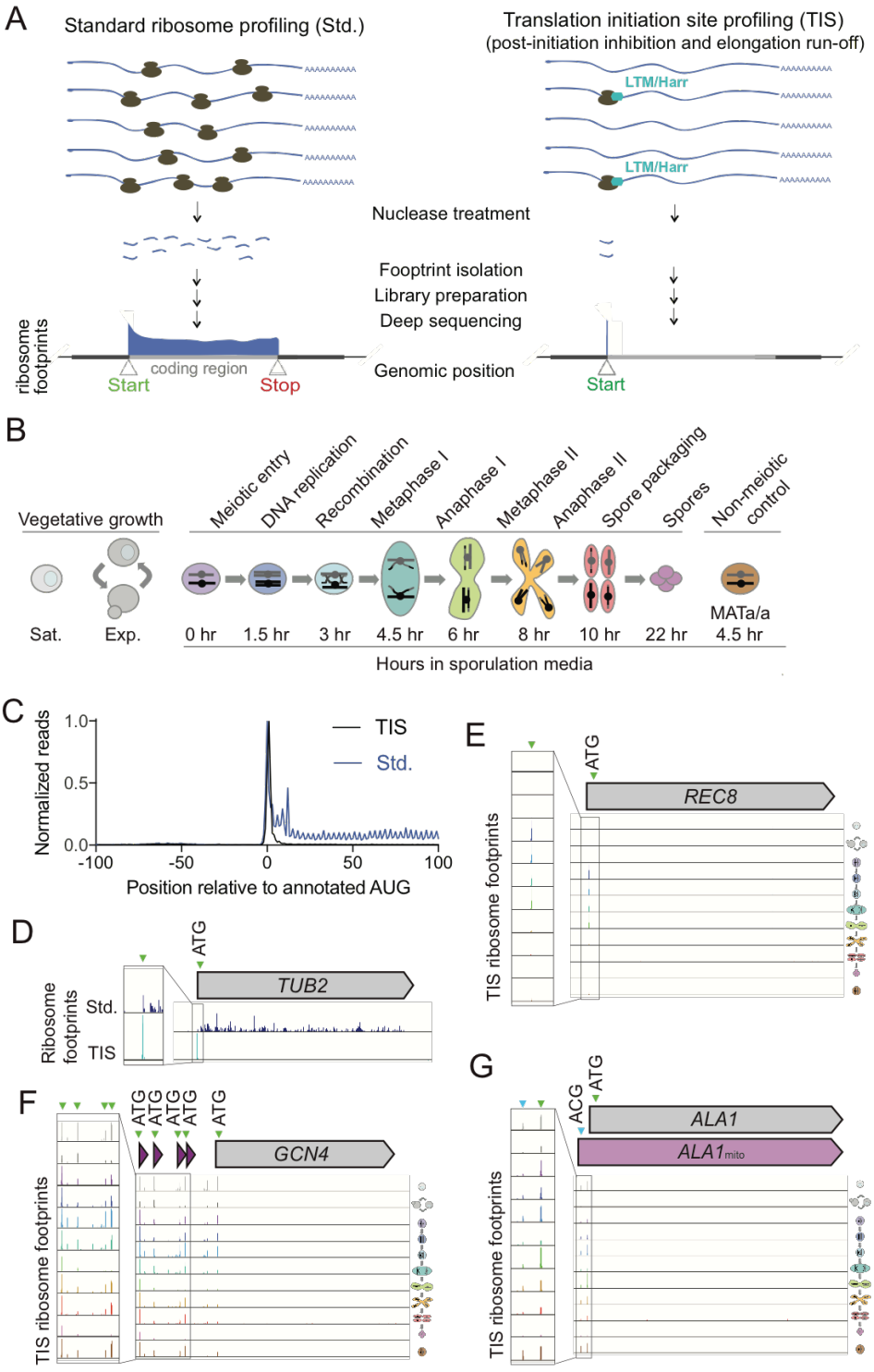

Figure 2.1: Translation initiation site ribosome profiling in mitotic and meiotic yeast cells
(A) Cartoon comparing standard (Std., left) and translation initiation site (TIS, right) ribosome profiling, with representative ribosome footprint profiles for a typical ORF.

(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-G) TIS-profiling of *REC8* (E), *GCN4* (F) and *ALA1* (G), 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.

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 2.S1B; Schneider-Poetsch et al., 2010). LTM concentrations 25-fold less than those used for TIS mapping in mammalian cells (Lee et al., 2012) still caused a growth defect in yeast (Figure 2.S1C) and resulted in strong TIS enrichment of ribosome footprints (Figure 2.S1D). This suggests that post-initiation ribosomes are more sensitive to LTM-based inhibition than elongating ribosomes. We selected an LTM concentration of 3 μM and a 20 minute incubation prior to harvesting to allow sufficient run-off time for elongating ribosomes. We performed translation initiation site profiling (TIS-profiling) for eight meiotic time points to assess translation initiation globally during meiosis (Figure 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.

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

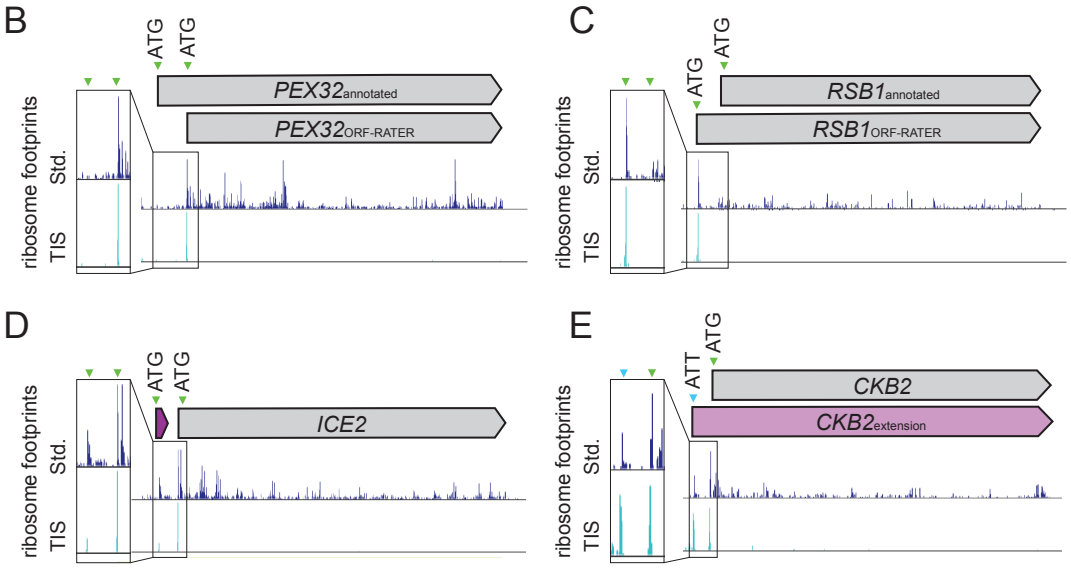

Figure 2.2: ORF-RATER annotations of TIS-profiling
(A) Numbers of different types of ORFs called by ORF-RATER at two different score cutoffs - a high score cutoff (> 0.5) and a low score cutoff (> 0.1). Truncation and internal out-of-frame numbers are likely overestimates due to high rates of false positives, indicated with a *.
(B-E) 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.

2.2.2 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 2.S2A, 2.S2B). 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 2.S2C). Finally, we estimate that approximately 15% of uncalled canonical annotated ORFs (representing 5% of total annotated ORFs) are false negatives, like *RIM11,* for which ORF-RATER did not call an ORF despite an observable peak at the annotated start site in the TIS-profiling data (Figure 2.S2D).

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 2.S2E, 2.S2F), and the fact that there were a median of 16 internal ORFs called per annotated gene (score >0.1; Figure 2.S2G). This high rate of apparent false positives is likely due to residual translation elongation inhibition at the concentration of LTM used in our method, resulting in background ribosome footprints within translated ORFs that erroneously result in internal TIS calls. While real internal initiation sites are expected to exist within these calls, the experimental and detection conditions here were not able to systematically separate true from false positives. In contrast to internally-initiated ORFs, manual visual analysis of the data for extensions and downstream ORFs called by ORF-RATER suggested that ORF-RATER calls of these classes of non-canonical ORFs are highly specific. We concluded that our analytical conditions are suitable to detect both canonical and non-canonical ORFs, with the exception of internal ORFs. We therefore excluded both out-of-frame internal ORFs and in-frame 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 2.S3A, Almagro Armenteros et al., 2019; Fukasawa et al., 2015). Excluding AUG-initiated extensions, many of which are likely to represent misannotations (as for *RSB1*, Figure 2.2C), left 231 extensions, representing 160 unique genes, as some genes contained multiple predicted extensions (Figure 2.S3B; this number was ultimately adjusted to 149 based on misannotations discovered through conservation analysis).

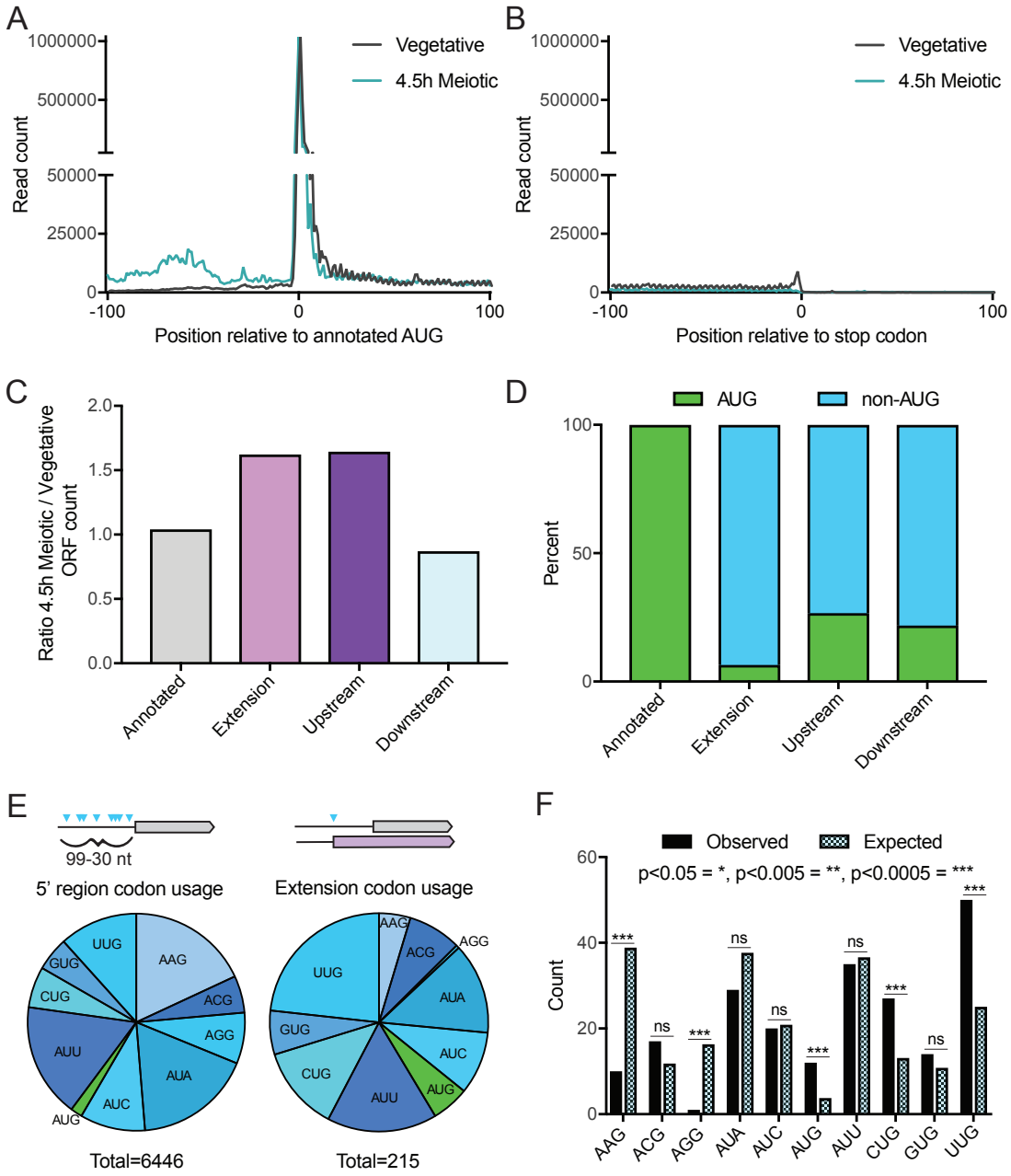

Figure 2.3: Specificity of uORF and N-terminal extension translation is partly dependent on condition and start codon identity

(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.

2.2.3 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 2.S2D), 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 (>10 additional amino acids; 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., 2009). The preference for specific near-cognate codons alone could not explain the small percentage of potential start codons in 5' leaders used to translate extended ORFs, so we also searched for evidence that start codon context influenced the set of used versus theoretically possible TISs. We found only weak enrichment for the optimal (Kozak-like) motif found around annotated AUG-initiated ORFs (Figure 2.S4A; Kozak, 2002; 1999; 1984; 1978), which is consistent with previous reports of differences between optimal contexts around near-cognate and AUG start codons (C.-P. 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 2.S3A). Of the 160 unique genes searched in this way, seven showed at least one peptide originating from the extension. Three of the seven had ORF-RATER scores well below the high score cutoff of 0.5 (Figure 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.

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	

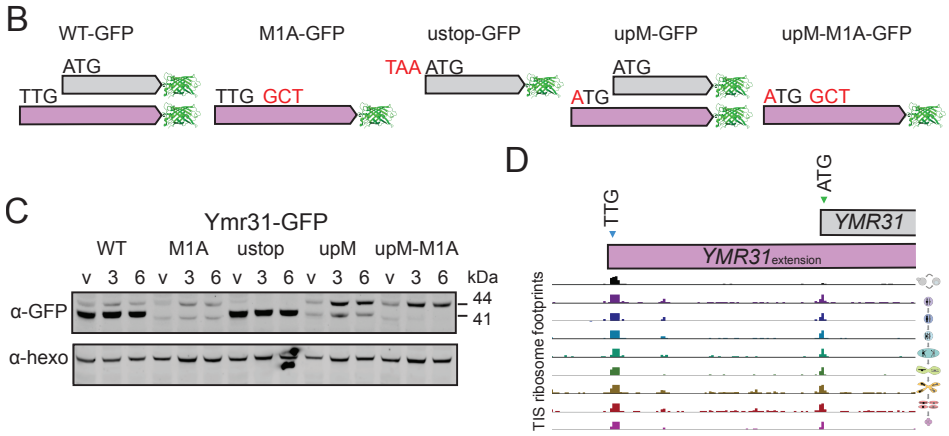

Figure 2.4: The abundance of near-cognate-initiated isoforms is not reflective of TIS-profiling peak height

(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.

2.2.6 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, 2.S5A). The extended form of Ymr31 was ten times lower in abundance than the canonical form in *WT* cells by western blot analysis (Figure 2.S5A), which is in marked contrast with the TIS-profiling data showing over eight times higher ribosome footprint read density at the near-cognate initiation site than at the canonical start codon (Figure 2.4D, 2.S6A).

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., 2009). This result also suggested that the peak height observed by TIS-profiling at near-cognate and AUG codons may not be comparable. This may be due to differences in the ability of LTM to inhibit the two different types of post-initiation ribosome complexes

or in their timespan of initiation. We also considered the possibility that near-cognate-initiated proteins might be subject to proteasome-mediated degradation, but at least for Ymr31, we did not observe an increase in the alternate isoform in cells in which proteasome activity was inhibited by MG132 (Figure 2.S6B, 2.S6C).

We further investigated whether the discrepancy between protein levels and TIS peak height indicated that TIS-profiling peaks were not quantitatively predictive of translation levels. This was not generally true, at least for AUG-initiated ORFs, as the height of 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 2.S6D, 2.S6E). This is seen by comparisons of individual time points (Figure 2.S6E), as well as by calculating correlation scores for each gene across all time points (Figure 2.S6D). 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 2.S6D). 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.

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.

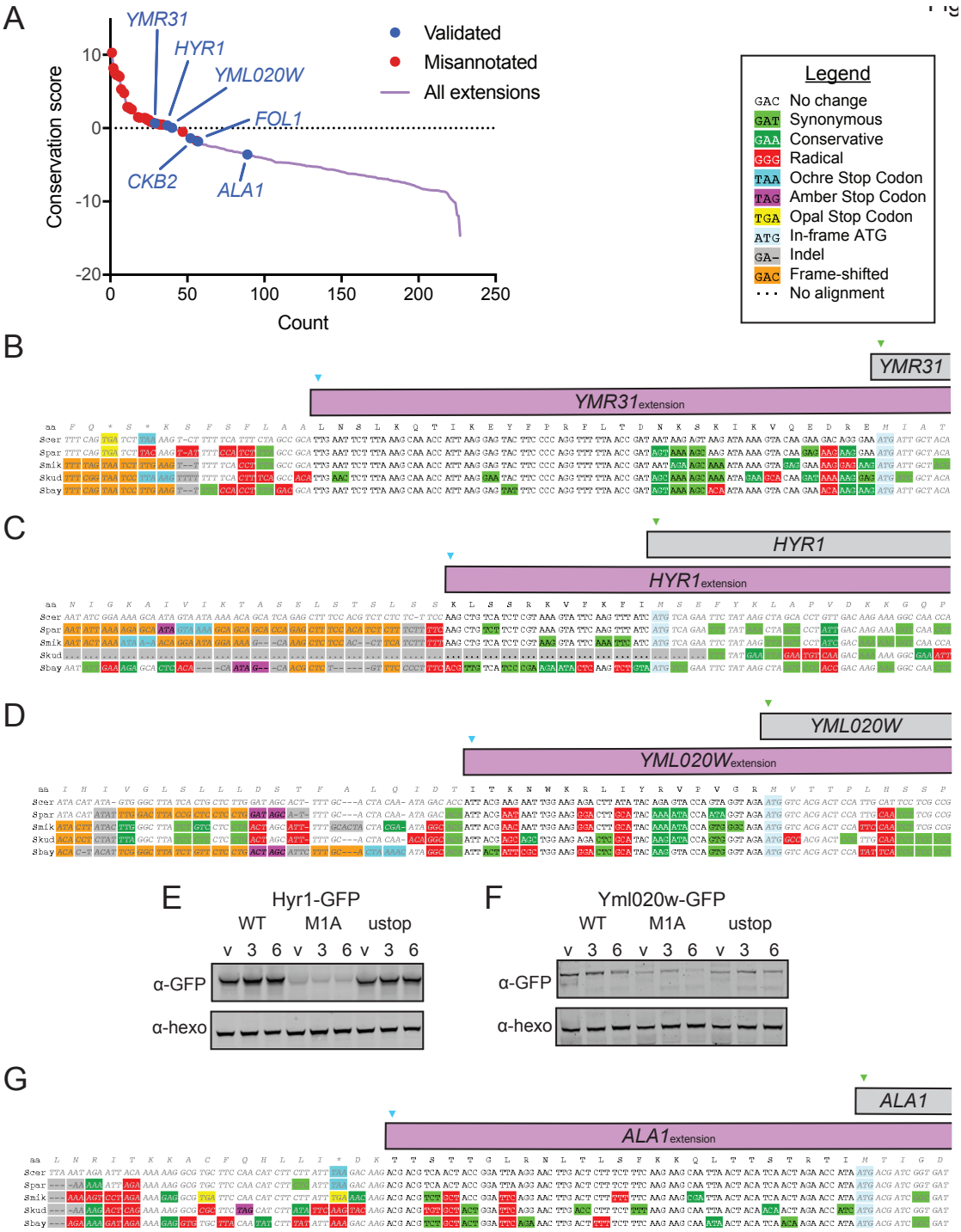

Figure 2.5: Most ORF extensions are poorly conserved
(A) Plot of PhyloCSF conservation scores for extension ORFs. Misannotated extensions are shown with red dots, and validated extensions are shown with blue dots, including three previously validated extensions (*YMR31*, *HYR1* and *ALA1)*. The additional validated extensions (*YML020W*, *CKB2* and *FOL1)* were validated in this study.
(B-D) Alignments showing level of conservation for *YMR31* (B)*, HYR1* (C) *and YML020W* (D), all of which have positive conservation scores.
(E) Western blot of Hyr1-GFP including *WT*, *M1A* and *ustop* constructs. Samples were taken in vegetative exponentially growing cells (v), and at 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.

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, 2.S5B). Similarly, we detect an N-terminally extended isoform of Yml020w in cells carrying the *YML020W-M1A* construct (Figure 2.5F, 2.S5C).

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.

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, 2.S5D). 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., 2009). We further found that the mRNA levels of GFP from the *ALA1^{GFP}-M1A* construct were dramatically decreased relative to the *ALA1^{GFP}-WT* construct (Figure 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 2.S7A). Consistent with this hypothesis, we observed that both mRNA and protein levels of the *ALA1^{GFP}-M1A* reporter construct increased in an NMD-deficient mutant background (*upf1*Δ*)*, although not to the level of the extended isoform in the *ALA1^{GFP}-WT* reporter construct (Figure 2.6B-D).

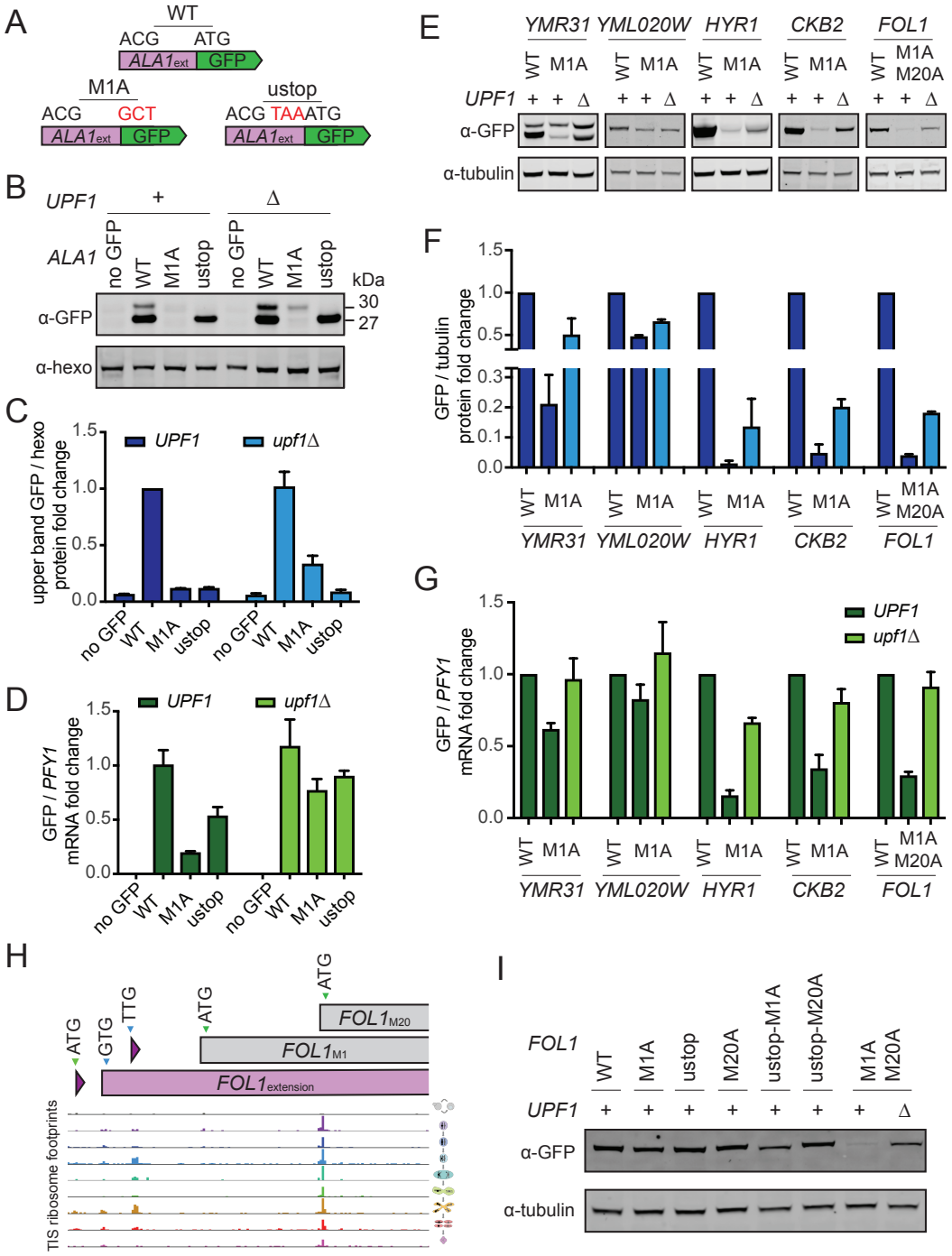

Figure 2.6: Extended ORF transcripts with no in-frame ATG are degraded by NMD
(A) Schematic for *ALA1* tagging strategy, using a reporter including the region upstream of the ATG, and either including (*WT*) or not including (*M1A*) the in-frame ATG in front of the GFP, and a mutant with a stop codon upstream of the in-frame ATG (*ustop*).
(B) Western blot for Ala1^{GFP} reporters in WT and *upf1*Δ vegetative cells. The band corresponding to the extension (30 kDa), can be seen in the *WT* construct, but is not seen in the *M1A* construct in WT cells. In a *upf1*Δ 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 *upf1*Δ background.
(E) Western blot analysis of Ymr31-GFP, Hyr1-GFP, Fol1-GFP, Ckb2-GFP and YML020W-GFP for the *WT* and *M1A* constructs in *UPF1* cells and the *M1A* construct in *upf1*Δ cells at 4.5 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.

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 (Figure 2.5E; 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, 2.S5E), 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 2.S7A-C), although this distance is thought to be a key factor in specifying yeast NMD substrates (Reviewed in Hug et al., 2016). We did, however, observe a moderately positive association between the distance of the 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 2.S7B).

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 *upf1*Δ cells, confirming translation from the upstream GUG identified by TIS-profiling (Figure 2.6H, 2.6I, 2.S5F). 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.

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 2.S8). eIF5A has recently been shown to influence translation elongation and termination (Gregio et al., 2009; Henderson and Hershey, 2011; Saini et al., 2009; Schuller et al., 2017), but was initially identified for activity in promoting a late stage of translation initiation *in vitro* (Benne and Hershey, 1978; Kemper et al., 1976; Lopo et al., 1986; Schreier et al., 1977). A CRISPRi screen in human cell lines identified eIF5A as a factor that enhanced translation of the CUG-initiated 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).

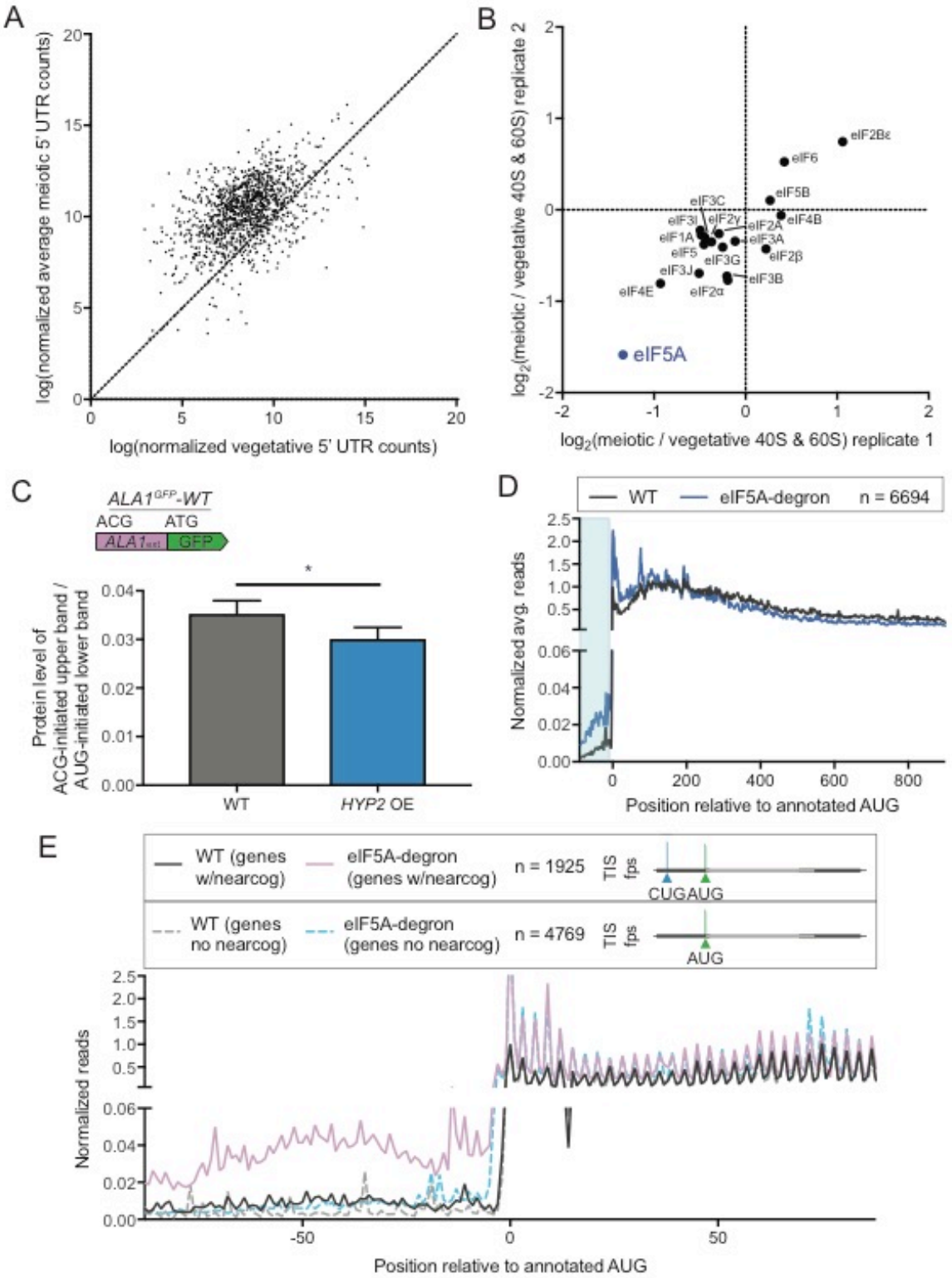

Figure 2.7: eIF5A levels regulate pervasive non-AUG-initiated translation
(A) Comparison of vegetative and average meiotic 5' read density measurements.
(B) Enrichment of translation factors comparing meiotic and vegetative samples for two replicates, determined by quantitative mass spectrometry of 40S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

(C) Western blot quantification of *ALA1*^{GFP}*-WT* reporter in meiosis with copper induction in strains containing or lacking a copper-inducible overexpression (OE) *HYP2* allele. Non-AUG-initiated GFP upper band 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.

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, 2.S5G, 2.S5H), suggesting that lower eIF5A is at least partly responsible for the increased translation from near-cognate codons seen in meiotic cells. The small effect seen here is not surprising, as simply overexpressing eIF5A may not increase the relevant functional pool of this factor, which not only has multiple characterized roles as noted above, but is also regulated by hypusine modification (Hershey et al., 1990). Indeed, mass spectrometry data show that Lia1, one of the enzymes responsible for Hyp2 hypusination, is dramatically decreased in meiotic cells, which would lead to lower Hyp2 activity (Figure 2.S8). 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.

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 2.4D). 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 2.S6C). We favor a model in which near-cognate-initiating ribosomes pause longer at TISs and are thus captured there more efficiently by ribosome profiling. It is also possible that ribosomes initiating at near-cognate and AUG TISs differ in their susceptibility to LTM-based inhibition, leading to preferential capture of reads at near-cognate sites by TIS-profiling.

Although previous studies have identified individual cases of extensions or predicted potential extensions computationally, it has not been possible to experimentally determine the pervasiveness of alternate protein isoforms beginning at non-AUG codons. This has become a recent area of interest, with three of the six established

cases in yeast identified in just the last 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., 2009), 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 2.S9A, 2.S9B). 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., 2009). The basis for the additional specificity beyond near-cognate codon identity cannot be explained by optimal context cues used to define the set of AUG start 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 2.S9C). Such constraint typically indicates function beyond protein coding, such as RNA structure. Consistently, a conserved, stable RNA

structure is predicted downstream of the AUU by RNAz analysis, (Figure 2.S9C), which may contribute to the high initiation efficiency at this site (Kozak, 1990).

We found that eIF5A is a trans-factor that contributes to translation of near-cognate-initiated protein isoforms in meiotic cells. eIF5A is known to associate with 60S ribosomal subunits and has been reported to affect multiple aspects of translation (Gregio et al., 2009; Melnikov et al., 2016; Schuller et al., 2017). We found low eIF5A association with ribosomal subunits in meiosis, leading us to investigate 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., 2018; Manjunath et al., 2019; Schuller et al., 2017). The recent finding that low eIF5A enhances CUG-initiated MYC translation in mammals, as well, suggests a conserved mechanism in the regulation of near-cognate-initiated protein isoforms (Manjunath et al., 2019).

An especially intriguing outstanding question raised by this study is the potential function of the many new protein extensions that were identified. Their generally low conservation suggests that they could expand the function of conserved proteins in a species-specific manner. All six known cases of near-cognate initiated alternate protein isoforms result in mitochondrial targeting of the extended protein and dual mitochondrial/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 2.S9D), 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 Data and Code Availability

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

2.4.2 Yeast strain construction

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

BrÜn 13 MATa wild-type
BrÜn 14 MATalpha wild-type
BrÜn 15 MATa/alpha wild-type
BrÜn 1362 MATa/alpha wild-type
BrÜn 5805 MATa/a wild-type
BrÜn 12507 MATa/alpha ymr31::KanMX trp1::YMR31-WT-yEGFP::TRP1
BrÜn 12508 MATa/alpha ymr31::KanMX trp1::YMR31-M1A-yEGFP::TRP1
BrÜn 12509 MATa/alpha ymr31::KanMX trp1::YMR31-ustop-yEGFP::TRP1
BrÜn 12510 MATa/alpha hyr1::KanMX trp1::HYR1-WT-yEGFP::TRP1
BrÜn 12511 MATa/alpha hyr1::KanMX trp1::HYR1-ustop-yEGFP::TRP1
BrÜn 12880 MATa/alpha hyr1::KanMX trp1::HYR1-M1A-yEGFP::TRP1
BrÜn 16920 MATalpha trp1::ALA1-yEGFP_WT::TRP1
BrÜn 16922 MATalpha trp1::ALA1-yEGFP_M1A::TRP1
BrÜn 18006 MATa/alpha ymr31::KanMX trp1::YMR31-upM-M1A-yEGFP::TRP1
BrÜn 18039 MATa/alpha ymr31::KanMX trp1::YMR31-upM-yEGFP::TRP1
BrÜn 18547 MATalpha trp1::ALA1-yEGFP_ustop::TRP1
BrÜn 18766 MATa upf1::NatMX
BrÜn 19023 MATa trp1::ALA1-yEGFP_WT::TRP1 upf1::NatMX
BrÜn 19025 MATa trp1::ALA1-yEGFP_M1A::TRP1 upf1::NatMX
BrÜn 19033 MATa trp1::ALA1-yEGFP_ustop::TRP1 upf1::NatMX
BrÜn 19302 MATa/alpha ymr31::KanMX trp1::YMR31-WT-yEGFP::TRP1 pdr5::HygMX
BrÜn 19303 MATa/alpha ymr31::KanMX trp1::YMR31-M1A-yEGFP::TRP1
BrÜn 19430 MATa/alpha hyr1::KanMX trp1::HYR1-M1A-yEGFP::TRP1 upf1::NatMX
BrÜn 20203 MATa/alpha ymr31::KanMX trp1::YMR31-M1A-yEGFP::TRP1 upf1::NatMX
BrÜn 20858 MATa/alpha ymr31::KanMX trp1::YMR31-M1A-yEGFP::TRP1 upf1::NatMX pdr5::HygMX
BrÜn 21423 MATa/alpha trp1::FOL1-WT-yEGFP::TRP1
BrÜn 21426 MATa/alpha trp1::CKB2-WT-yEGFP::TRP1
BrÜn 21640 MATa/alpha trp1::YML020W-WT-yEGFP::TRP1
BrÜn 21716 MATa/alpha trp1::FOL1-M20A-yEGFP::TRP1
BrÜn 21719 MATa/alpha trp1::CKB2-M1A-yEGFP::TRP1
BrÜn 21723 MATa/alpha trp1::YML020W-ustop-yEGFP::TRP1
BrÜn 21816 MATa/alpha trp1::CKB2-M1A-yEGFP::TRP1 upf1::NatMX
BrÜn 22159 MATa/alpha trp1::FOL1-M1A-yEGFP::TRP1
BrÜn 22353 MATa/alpha trp1::FOL1-M1A-M20A-yEGFP::TRP1

BrÜn 22434 MATa/alpha trp1::FOL1-M1A-M20A-yEGFP::TRP1 upf1::NatMX
BrÜn 22526 MATa/alpha trp1::YML020W-M1A-M31A-yEGFP::TRP1
BrÜn 22529 MATa/alpha trp1::YML020W-M1A-M31A-yEGFP::TRP1 upf1::NatMX
BrÜn 23156 MATa/alpha trp1::FOL1-ustop-yEGFP::TRP1
BrÜn 23157 MATa/alpha trp1::FOL1-ustop-M20A-yEGFP::TRP1
BrÜn 23955 MATa/alpha trp1::FOL1-ustop-M1A-yEGFP::TRP1
BrÜn 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*.

2.4.3 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.

2.4.4 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.

2.4.5 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 4^{o}C. 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.

2.4.6 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.

2.4.7 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).

2.4.8 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: *GFP* (oGAB-2736/oGAB-2737), PFY1 (oGAB-3301/oGAB-3302), and *HYP2* (oGAB-7864/oGAB-7865).

2.4.9 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 2.S2A-B.

2.4.10 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.

2.4.11 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 2.S4B, 2.S4C).

2.4.12 Context analysis

Maximum motif score analysis was performed using Mochiview (Homann and Johnson, 2010) 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.

2.4.13 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.

2.4.14 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.

2.5 Supplemental Figures

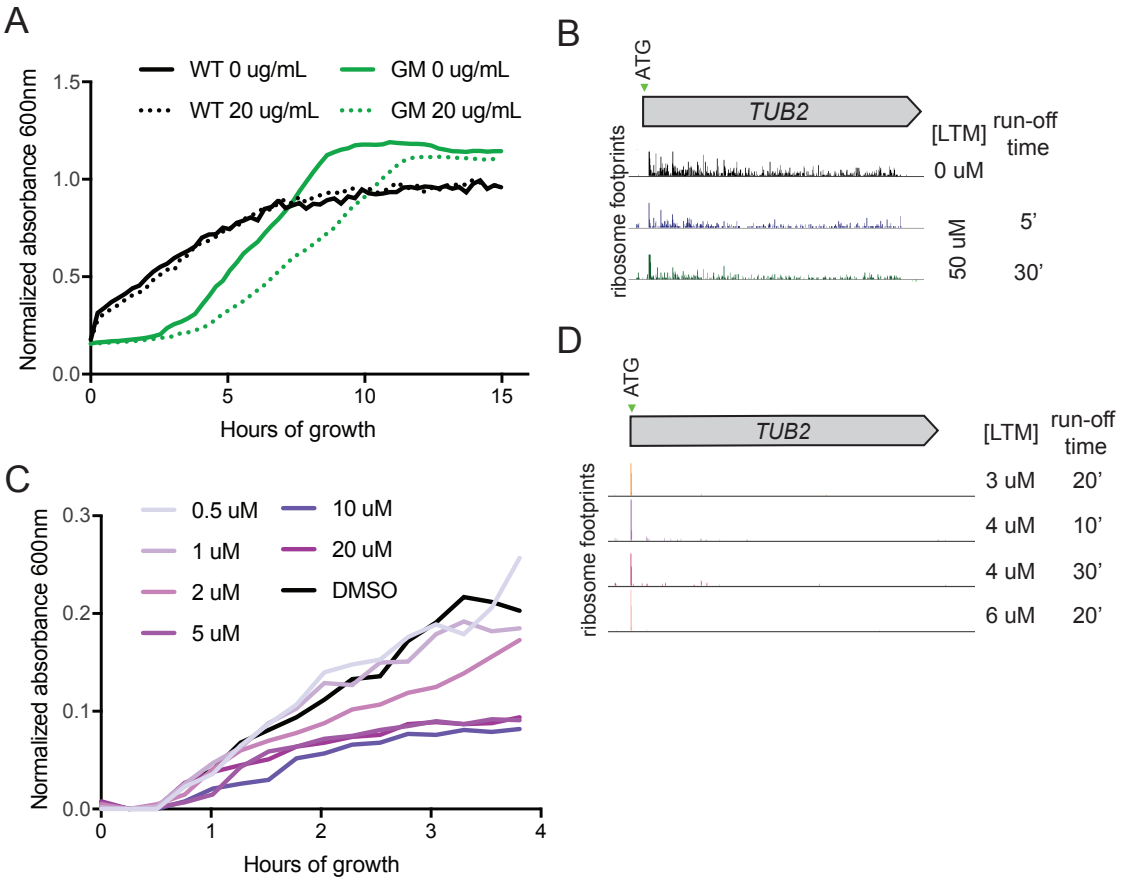

Figure 2.S1: Optimization of TIS-profiling conditions for yeast
(A) Growth curve of WT cells or Green Monster (GM) mutant cells treated with harringtonine. The GM strain lacks 16 ABC transporter drug efflux genes. Solid lines indicate no treatment and dotted lines indicate 20 ug/mL of harringtonine. 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*.

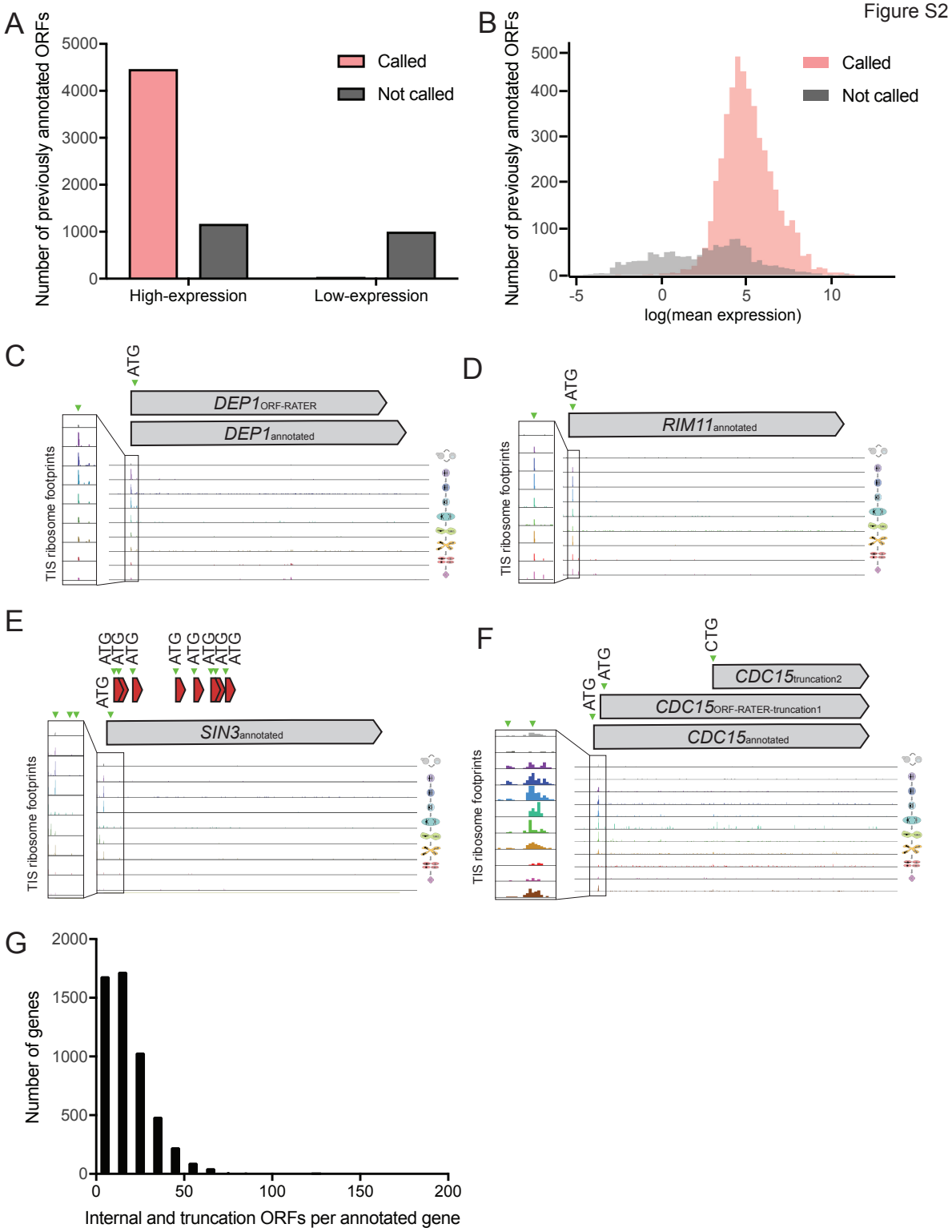

Figure 2.S2: Categories of false positive and false negative ORF-RATER calls
(A) Previously annotated ORFs that are called (pink) or not called (gray), at expression values greater (high-expression) or less than (low-expression) 5 mean RPKM. Approximately half of annotated ORFs that were not called have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

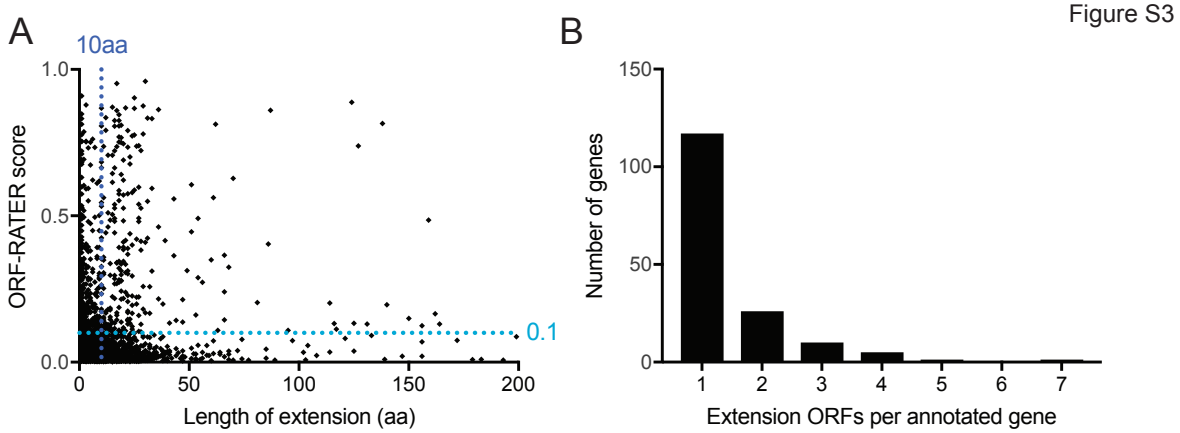

Figure 2.S3: Properties of extension ORFs used for setting cutoffs
(A) Length versus score for all extension ORFs, with a line showing the length cutoff at 10 amino acids and the score cutoff of 0.1.
(B) Number of extension ORFs called per annotated gene.

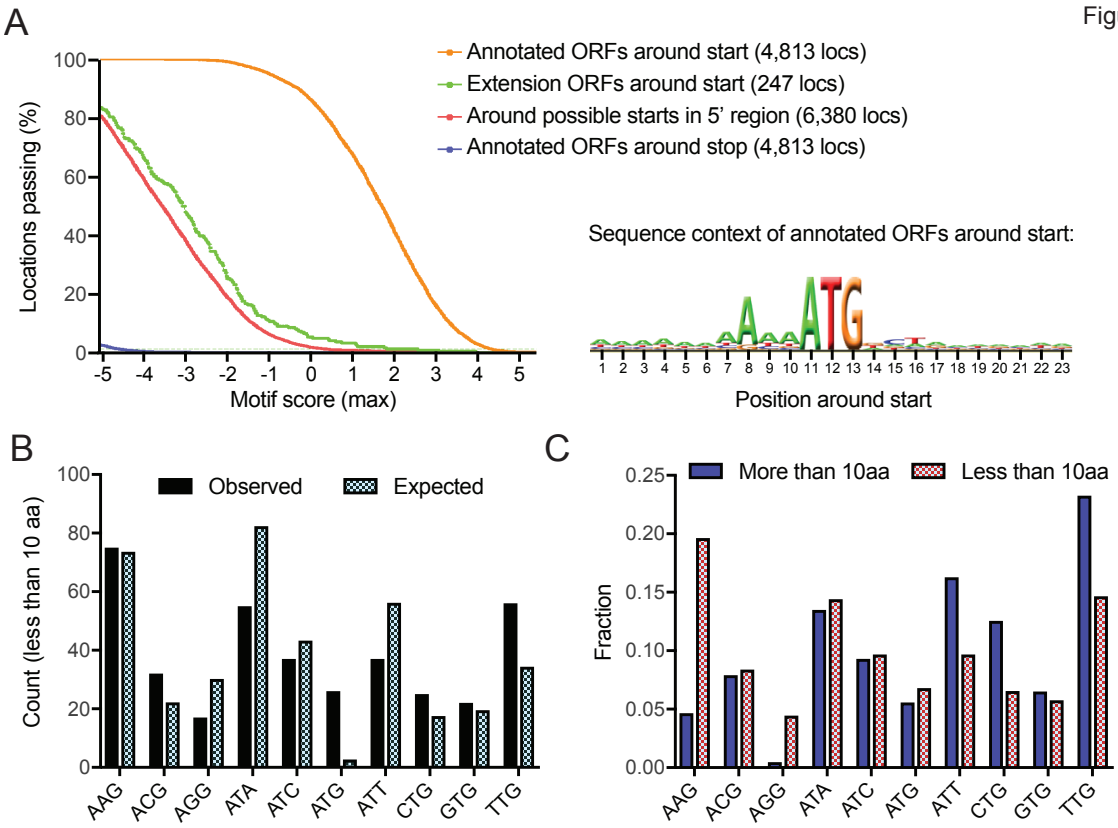

Figure 2.S4: Translated near-cognate-initiated ORFs do not show Kozak sequence context enrichment

(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.

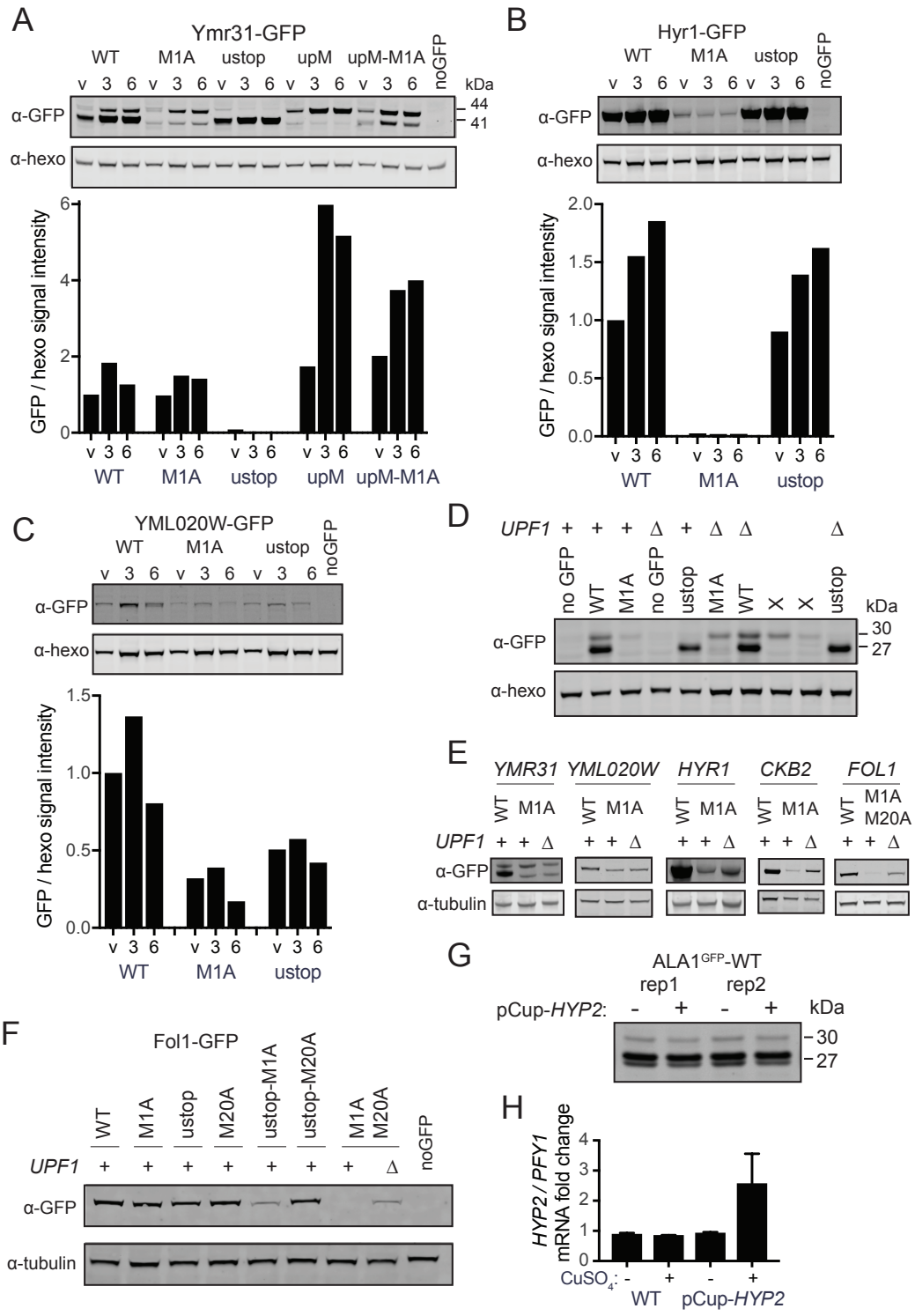

Figure 2.S5: Western blot replicates and quantification for alternate isoforms
(A) Replicate western blot of *YMR31-GFP* constructs, as in Figure 2.4C (top) and quantification of upper GFP band relative to hexokinase loading control for three replicates (bottom).
(B) Replicate western blot of *HYR1-GFP* 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 *ALA1*^{GFP} 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 *upf1*Δ, as in Figure 2.6E.
(F) Replicate western blot of *FOL1-GFP* constructs, as in Figure 2.6I.
(G) Western blot of *ALA1*^{GFP}*-WT* reporter for cells with and without the *pCup-HYP2* construct with copper ($CuSO_4$) addition leading to overexpression of eIF5A for two replicates, which is quantified in Figure 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 ($CuSO_4$) addition for three replicates.

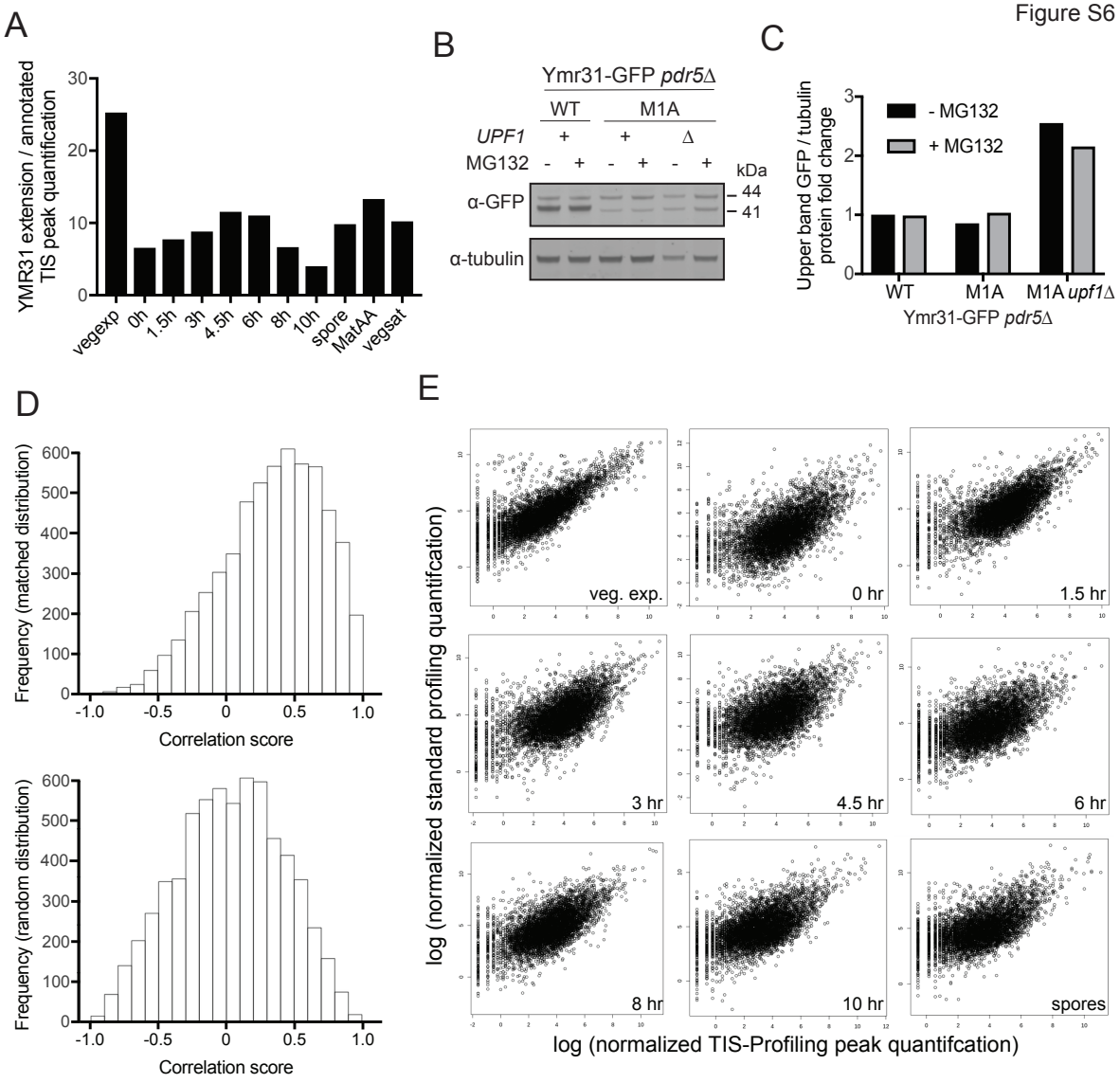

Figure 2.S6: Positive correlation of TIS peaks with gene expression for annotated AUG sites but not near-cognate sites

(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 $<2.2 \times 10^{-16}$.

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

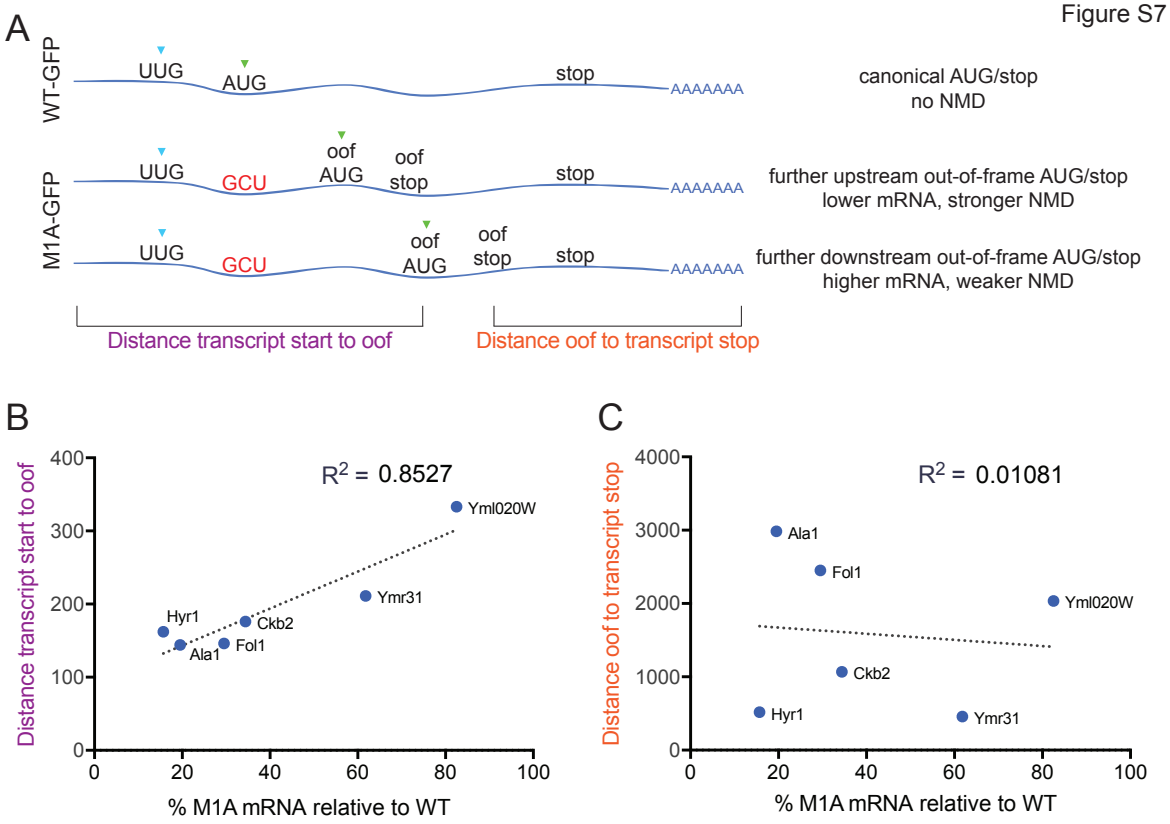

Figure 2.S7: Effect of NMD for M1A transcripts does not correlate with distance from premature stop to transcript end

(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 R^2 value of 0.8527 is seen, indicating that a shorter distance from the transcript start to the oof ORF correlates positively with less *M1A* mRNA relative to *WT* and therefore stronger NMD.

(C) Correlation between the distance from the end of the newly created oof ORF to the end of the transcript relative to the percent of *M1A / WT* mRNA level from Figure 2.6G. A correlation with an R^2 value of 0.01081 is seen, indicating esentially no association between the distance from the oof ORF to transcript stop and the strength of NMD.

Figure S8

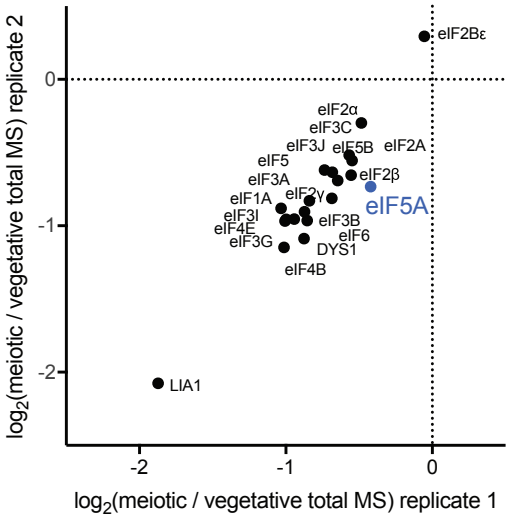

Figure 2.S8: Total protein abundance of initiation and hypusination factors
Enrichment of translation factors (as in Figure 2.7B) and hypusination factors Lia1 and Dys1 comparing meiotic and vegetative samples for two replicates, determined by quantitative (TMT10) mass spectrometry of whole cell extract from meiotic and vegetative cells.

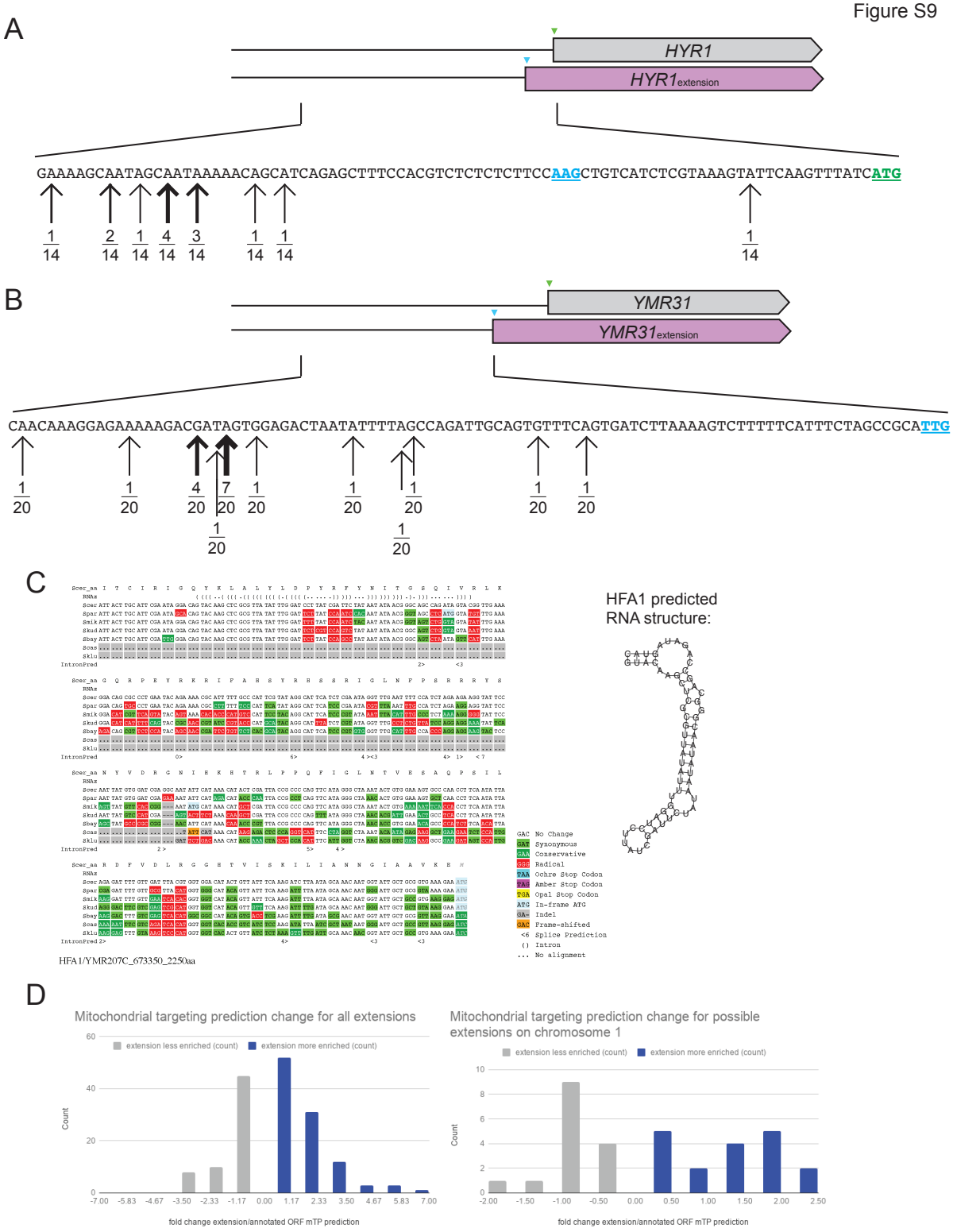

Figure 2.S9: HFA1 RNA structure and mitochondrial targeting sequence prediction
(A) 5'RACE analysis of *HYR1*. Locations of transcription start sites are indicated with arrows, with the number of sequencing reads at that site indicated. A total of 14 transcription start sites were sequenced. (B) 5'RACE 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

Ribosomal subunits are degraded and re-synthesized in yeast meiosis

Part of the work presented in this chapter has previously been published in the following manuscript: Eisenberg, A.R.,* Higdon, A.,* Keskin, A., Hodapp, S., Jovanovic, M., and Brar, G.A., 2018. Precise Post-translational Tuning Occurs for Most Protein Complex Components during Meiosis. *Cell Reports*, 25, 3603-3617.e2. (* indicates co-authorship).

https://doi.org/10.1016/j.celrep.2018.12.008

3.1 Introduction

The protein complement of a cell defines its structure and function and is determined by the relative rates of synthesis and degradation for each protein present. The mechanisms and specificity determinants of synthetic processes in gene expression, especially transcription, have been well studied. In addition, the basic classes of mechanisms by which proteins are degraded within cells, including through regulated ubiquitin-based protein turnover, have been defined. However, it remains difficult to systematically determine the impact of regulated protein degradation in an unperturbed system, even at steady state. Comprehensively assessing the timing and specificity of protein degradation mechanisms in the context of cellular differentiation is an even greater challenge, but also particularly critical, as the transitions between sequential cellular stages require waves of both synthesis of new proteins and removal of pre-existing proteins.

Early examples of regulated degradation were identified by single-gene analyses, such as the case of Cyclin during the cell cycle and meiosis (Evans et al., 1983). Cyclin protein synthesis was observed to be constitutive, but the protein level fluctuated, leaving regulated protein degradation as the remaining explanation for the protein expression pattern observed. These observations ultimately led to the discovery of the conserved anaphase-promoting complex/cyclosome (APC/C)-based specificity mechanism, which is responsible for a key event in cell division (Irniger et al., 1995; King et al., 1995; Sudakin et al., 1995). Here we apply a similar approach, based on examining genome-wide protein patterns during a natural process to identify cases of protein degradation more globally. We recently generated a complex dataset that enabled these analyses, and that includes deep and matched mRNA-, translation-, and protein-level measurements during the natural process of meiotic differentiation in budding yeast (Cheng et al., 2018). These analyses revealed trends in degradation in the cell, and we focus here on the degradation and re-synthesis of ribosomal proteins following meiosis in maturing gametes.

3.2 Results

3.2.1 Protein Degradation Inferred from Comparison of Protein and Translation Data

Although our comprehensive dataset of gene expression through meiotic differentiation (Figure 3.1A) was previously analyzed only for transcriptional and translational regulation (Cheng et al., 2018), we found that comparison of translation and protein patterns could also be used to infer post-translational regulation of gene expression that accurately captures expected regulation. For example, the protein for cyclin-dependent kinase (CDK) Cdc28 is required through much of meiosis and was present for a period of at least 3 hours after translation ceased, suggesting protein stability during this time frame (Figure 3.1B). In contrast, protein levels of the synaptonemal complex (SC) component Zip1 declined in concert with a decrease in translation, which is consistent with the known active degradation of this protein in late prophase (Figure 3.1C; Sourirajan and Lichten, 2008). In an extreme case of such regulation, the CDK inhibitor Sic1 showed decreased protein levels, despite high ongoing translation during meiotic S phase and prophase, which is consistent with the critical known regulation of this protein (Figure 3.1D; Dirick et al., 1998).

While it is established that protein-level adjustment through degradation can occur in diverse mutant and perturbed conditions, its prevalence in wild-type cells under natural conditions has been difficult to assess. Coherently reconciling results from individual gene studies and large-scale studies, as well as between regulation in wild-type and perturbed cellular conditions, has been challenging. Quantitatively comparing the levels of mRNA, translation, and protein in parallel in wild-type cells should address this problem, but this requires comparing sequencing- and mass spectrometry-based measurements. These two fundamentally different types of measurements show different dynamic ranges of detection, and by traditional mass spectrometry, precise direct comparison of measurements between proteins with different physical properties is difficult.

We reasoned that our dataset provided a unique opportunity to assess the degree of post-translational adjustment of levels of stable protein complex members in wild-type cells because of several advantageous properties of the data. First, we measured mRNA, translation, and protein from matched extracts, allowing their direct comparison. Second, we measured nearly 80% of the annotated yeast proteome. Third, we made measurements for several sequential time points, which allowed analysis of trends and correction for differences in dynamic ranges of detection for mass spectrometry and sequencing data.

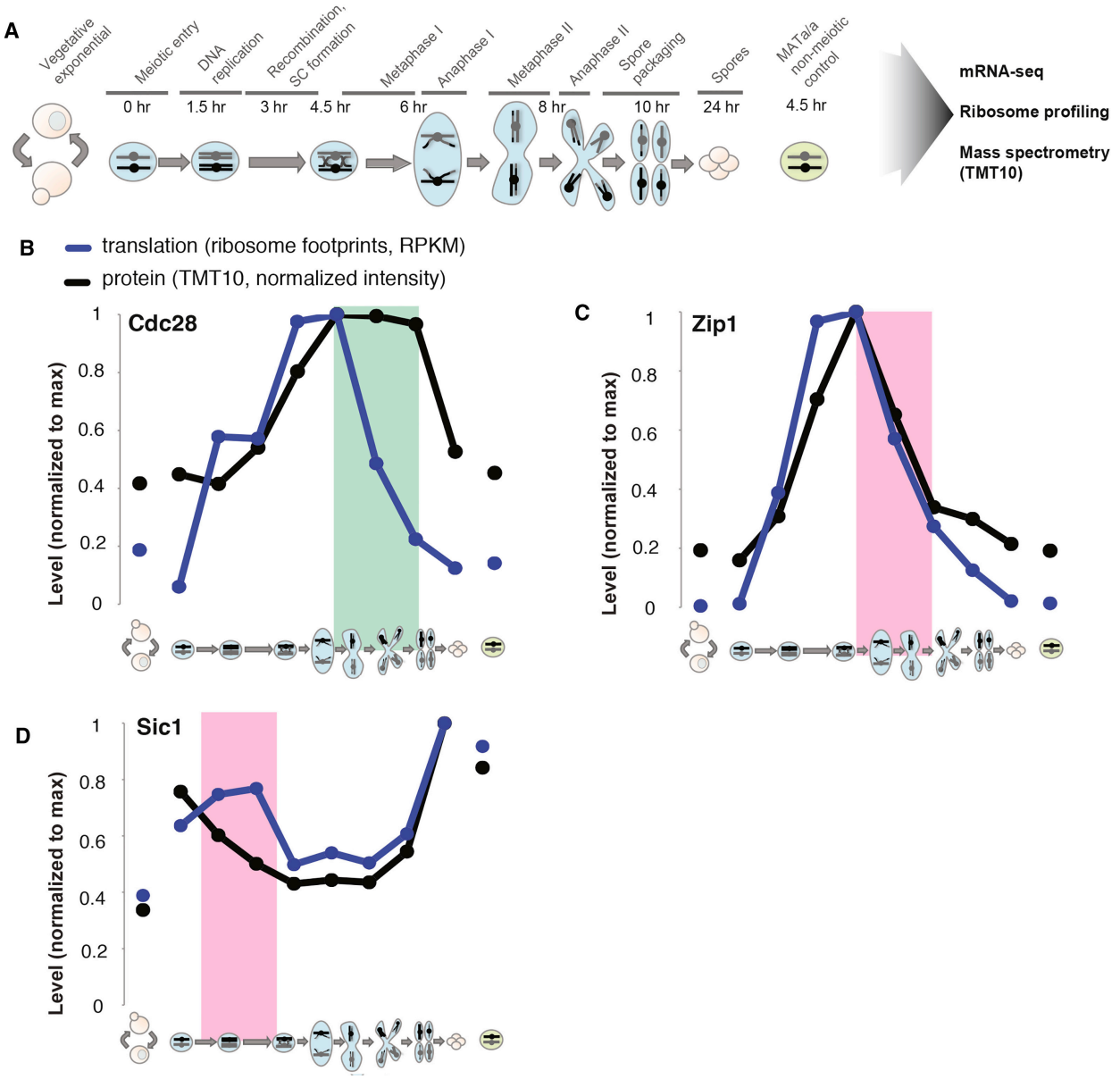

Figure 3.1. Regulated Protein Degradation Can Be Detected by Analysis of Protein Levels during Meiosis

(A) Schematic of meiotic gene expression experiment. Illustrations representing vegetative growth or meiotic stage are used to depict sample identity throughout figures. Left-hand vegetative cells are exponentially growing, and far-right cells are in nutrient-poor sporulation medium. Meiotic stages are noted above central portion of illustration and time in sporulation medium is noted directly below.

(B–D) Comparison of translation, assayed by ribosome footprint density (blue) and protein, assayed by quantitative mass spectrometry (black) are shown over time points for:

(B) Cdc28 (green box highlights a period of inferred protein stability; RPKM, reads per kilobase million);

(C) Zip1 (pink box highlights a period of inferred protein instability that matches known regulation); and

(D) Sic1 (pink box highlights a period of inferred protein instability that matches known regulation).

3.2.2 Ribosomal Proteins Are Highly Co-regulated at All Levels and Degraded Late in the Meiotic Program

The ribosome is a large multiprotein complex with known co-regulation of component synthesis. Clustering of ribosomal protein (RP) levels did not reveal subclustering based on any reported physical feature of the ribosome or large versus small subunit identity (Figure 3.2A). Rather, 72 of the 98 RP-encoding genes that we quantified at the protein level showed extremely similar patterns over the meiotic program at every level of gene expression measured. The other 26 showed slightly different protein-level patterns, the basis of which we do not yet understand (Figure 3.2A). We also noted that protein-level patterns for all RP genes examined indicated protein degradation late in MII and re-synthesis in spores (Figure 3.2A). Comparison of the mRNA, translation, and protein abundance measurements for this group of genes revealed a signature of protein degradation similar to what we observed for the known degradation target Sic1 (Figure 3.1D), with translation actually increased in late meiosis while protein levels decrease (Figure 3.2B). In spores, subsequently, protein levels increased to a level similar to early meiotic cells, while translation (and mRNA levels) remained high (Figure 3.2B). While the transcriptional uptick in RP genes had previously been seen during sporulation (Chu et al., 1998), its association with protein degradation was not evident.

To confirm this regulation independently, we GFP-tagged the RP gene *RPL26B* in a strain carrying mCherry-tagged histone H2B (encoded by *HTB1*). Both tags were heterozygous in diploid cells. Thus, during meiotic stages in which the cytosol was continuous, before spore packaging, we expect to see homogeneous green cytosolic and red nuclear signals. Following spore individualization, we would expect to see red signal remaining in all four spore nuclei if histones are stable, suggesting that they were synthesized before spore packaging (Figure 3.2C). In contrast, if RPs were degraded and re-synthesized in spores, then only the two spores carrying the *RPL26B-GFP* allele should be green and the other two spores should lose GFP signal relative to earlier stages (Figure 3.2C).

Indeed, this was what we observed. We saw evidence that Htb1 continues to be synthesized in spores, resulting in an increase in signal in the two spores that presumably carry the *HTB-mCherry* allele, but in the case of *RPL26B,* we observed an increase in signal in two spores and a decrease in the other two (Figures 3.2C and 3.2D). This was observed and quantified for individual cases and was a general trend among cells of this genotype (Figures 3.2D and 3.2E). The loss of GFP signal in two spores was not due to photobleaching resulting from time-lapse imaging, as cells on the same microfluidic plate that were not previously imaged showed similar relative levels of GFP in a 2:2 bright:dim ratio as those that were imaged over a period of time (Figure 3.3). A similar effect could be seen in cells carrying a heterozygous *RPL29-GFP* allele (Figures 3.2E and 3.3). We concluded that ribosomes are actively degraded and re-synthesized at the end of the yeast meiotic program.

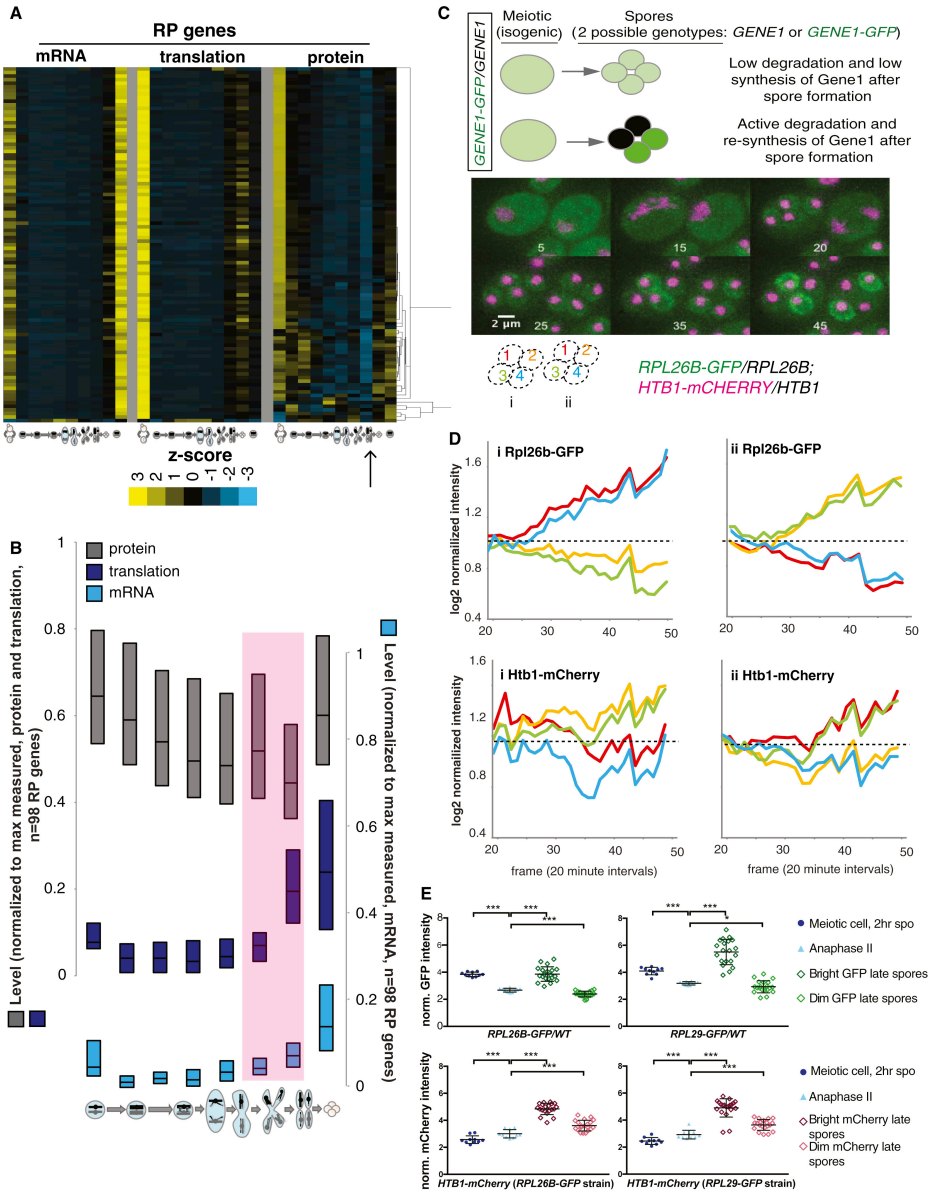

Figure 3.2. RPs Are Actively Degraded Late in the Meiotic Program

(A) Hierarchical clustering of protein levels was performed for all RP genes quantified (right), and is compared to matched translation (middle) and mRNA (left). Values shown are z-score normalized.

(B) Quartile analysis of all RPs at all levels of expression. Pink shading represents period late in meiosis when transcription and translation increase but protein decreases, a hallmark of active degradation.

(C) A strategy to identify active protein degradation and re-synthesis after spore wall formation. This approach uses heterozygous GFP tagging of the protein of interest, in this case Rpl26b, in diploid cells. Before spore formation, protein from both alleles is in the cytosol. After spore formation, if a protein is degraded and re-synthesized, then the fluorescent signal should decrease in spores that inherited the untagged allele and should increase in spores that inherited the tagged allele. This is observed for Rpl26b, but not histone protein Htb1. Inset numbers represent frame numbers for 20-min intervals; scale bar represents 2 µM. (D) Quantification of the fluorescence over time for the two cells in (C), starting when spore individualization begins. Note the decrease in GFP signal in two spores and the increase in the other two. (E) Quantification of additional cells (n = 10 tetrads) from the experiment in (C) and (D) and a similar experiment using heterozygous *RPL29-GFP*. Error bars represent SD. p values determined by paired t test: *p = 0.033, ***p < 0.001.

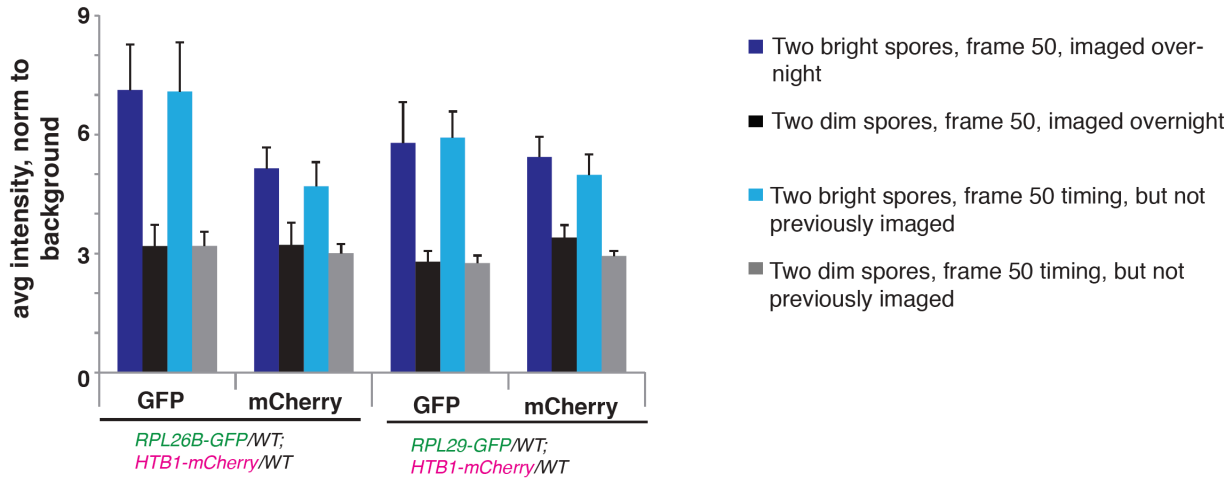

Figure 3.3. The low signal in two spores without fluorescently-tagged RP genes is not due to photo-bleaching.
Average integrated intensity of the two brightest or two dimmest spores from tetrads are compared for strains carrying heterozygously tagged Rpl26b and Htb1 (left) or Rpl29 and Htb1 (right). Plotted values are either for cells imaged after 50 frames or for not previously imaged cells at the same time point. 10 tetrads were counted per strain.

3.2.3 Deletions of proposed ribophagy genes *DOA1* and *UBP3* do not show a strong effect on ribosome degradation in meiosis

One possible mechanism of ribosome degradation during meiosis is through an autophagy-like mechanism, where the cell engulfs certain components and degrades them non-specifically. Other cellular components have been shown to be degraded through autophagy pathways, including mitochondria and ER, through mitophagy and ER-phagy, respectively (Otto et al., in preparation; Sawyer et al., 2019). There are several proteins that have been proposed to contribute to ribophagy – the autophagic degradation of ribosomes. Doa1 is a ubiquitin-binding cofactor that works in conjuction with Ubp3, which is part of a deubiquitination complex (Ossareh-Nazari et al., 2010). Both proteins have been shown to be involved in ribophagy in nutrient poor conditions, but have not been studied in the context of meiosis.

To assess whether these genes play a role in the degradation of ribosomes at the end of meiosis, deletions of *DOA1* and *UBP3* were generated in strains containing the heterozygous GFP tagged ribosomal subunit genes Rpl26b and Rpl29. These strains were imaged during meiosis and compared to strains without the deletions. No significant difference was observed in the amount of degradation occurring for the spores lacking the GFP tagged copy of the ribosomal subunit gene (Figure 3.4A). By western blot, there is also no change in the relative amount of Rpl-GFP to free GFP, which is a proxy for autophagy (Figure 3.4B). If Doa1 or Ubp3 were leading to the degradation of ribosomes in this context, we would expect that deleting them would show less free GFP being made. Together, this indicates that these two genes, at least acting alone, are not entirely responsible for the degradation of ribosomes during meiosis. Further analysis of the factors that affect ribosome degradation is needed to understand the mechanisms and the extent of their role in this process.

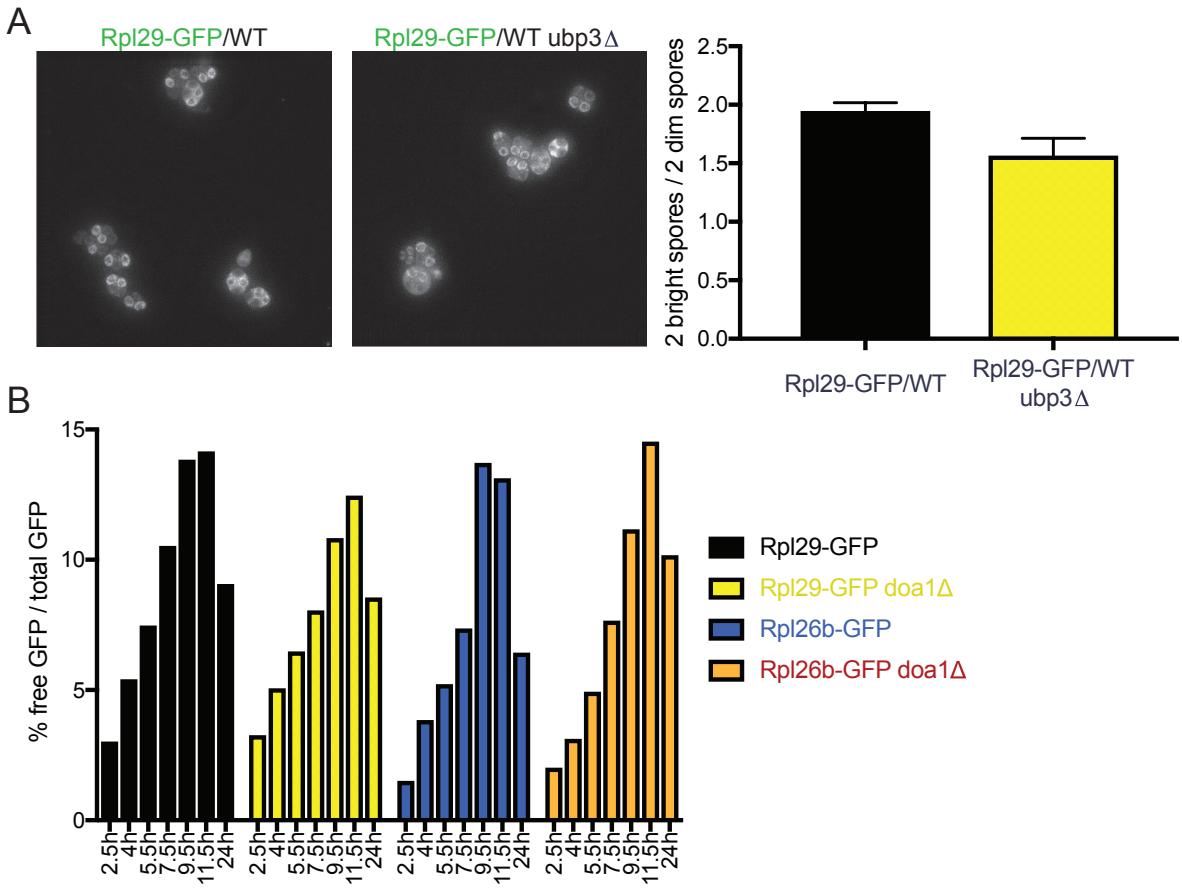

Figure 3.4. Doa1 and Ubp3 do not significantly affect ribosomal subunit degradation in meiosis.
(A) Microscopy of Rpl29-GFP/WT cells at 24 hours after meiosis, with UBP3 (left) and with ubp3Δ (center). Quantification of fluorescence of 2 bright spores relative to the 2 dim spores for 3 tetrads (right) shows a small but insignificant change in the relative brightness of the spores that inherit the GFP-tagged form of Rpl29 versus ones that do not.
(B) Western blot quantification of Rpl29-GFP and Rpl26b-GFP bands, with and without doa1Δ.

3.3 Discussion

Important individual examples of regulated protein degradation during meiotic differentiation are well characterized, but it has been challenging to determine the pervasiveness of this mode of regulation in meiosis. We report signatures in matched global quantitative mass spectrometry and ribosome profiling data that allowed us to identify specific, natural, and previously unidentified cases of regulated protein degradation during the yeast meiotic program. These signatures include periods of stable or even increased translation of a given gene, while protein abundance decreases, as well as periods of particularly rapid decline in protein levels for groups of genes in concert. These trends are sensitively detected in our dataset because cells undergoing meiotic differentiation do not display the type of dilution due to cell growth and division that is a major contributor to protein-level decline during mitotic growth (Christiano et al., 2014).

Analyses of protein data revealed previously unrecognized, coordinated degradation and re-synthesis of RP subunits following gamete (spore) formation (Figure 3.2). Why would cells expend energy to degrade RPs and concomitantly re-synthesize them? We propose two explanations. First, this could be a mechanism of cellular quality control. It has been shown that the abnormal nucleolar morphology observed in aged yeast cells is reset in all four gametes by the process of meiotic differentiation (Ünal et al., 2011). It is possible that this nucleolar morphology reflects defective rRNA synthesis or processing, and thus resultant ribosomes may be of poor quality. Because gamete quality is important for an organism's genome stability on an evolutionary scale and the proteins synthesized in a gamete provide critical functions, including mediating gene expression, the destruction and re-synthesis of a gamete's ribosomes may be a mechanism of ensuring gamete integrity. Second, it is possible that ribosome composition or modification is altered in meiosis relative to mitotic growth and that these alterations must be reset after meiosis. The translation of upstream ORFs (uORFs) within 5′ leaders of thousands of mRNAs is dramatically upregulated during meiosis, even in cases in which an apparently identical transcript is present under mitotic and meiotic conditions (Brar et al., 2012). If meiotic modification to the core translation machinery contributes to this effect, then destruction of this machinery could enable cells to return to mitotic translation patterns.

The findings reported here regarding protein degradation are informative in considering the cellular balance between perfection and efficiency. Yeast cells undergo meiosis in the absence of glucose or amino acids, and this is therefore a context in which cellular economy of resources is extremely important. The prevalence of degradation in these cells suggests that this strategy is advantageous to the cell, and understanding the basis for this advantage could provide additional insights into cellular resource management in gene expression regulation.

3.4 Materials and Methods

3.4.1 Yeast growth and sporulation

All yeast strains used were diploid *Saccharomyces cerevisiae* of the SK1 background. Strains used in this study:
BrÜn7085 (*MATa/α rpl29::RPL29-HTA-GFP:KanMX/RPL29 htb1::HTB1-mCherry-HISMX6/HTB1)*
BrÜn7087 (*MATa/α rpl26b::RPL26B-HTA-GFP:KanMX/RPL26B htb1::HTB1-mCherry-HISMX6/HTB1)*
For meiotic time courses, yeast were inoculated into YEPD overnight, then diluted to OD_{600} 0.2 into buffered YTA and grown for 12 hours. Cells were washed in water and resuspended in sporulation media supplemented with 0.02% raffinose. Time points were taken at indicated times.

3.4.2 Western blotting

Western blotting was performed using a standard TCA-based protocol. Briefly, 2.5 OD units of culture were treated with 5% TCA at 4C for at least 10 min. Samples were then washed with 1 mL acetone. Acetone was aspirated and pellets were dried overnight at RT. Lysates were made by adding 100 µL protein lysis buffer [50 mM TE, 3 mM DTT, 1.1 mM PMSF (Sigma), 1 µM pepstatin A, 1X protease inhibitor cocktail (Roche)] and 1 volume acid-washed glass beads (Sigma), and bead-beating for 5 min at RT. 3X SDS loading buffer was added and samples were boiled for 5 min. Beads were pelleted by centrifugation and 5 µL supernatant was loaded onto 4–12% Bis-Tris polyacrylamide gels. Following electrophoresis, proteins were transferred using a semi-dry transfer apparatus (Trans-Blot Turbo, BioRad). The following antibodies were used: mouse anti-V5 (Invitrogen, 1:2,000), rabbit anti-hexokinase (Stratech, 1:10,000), anti-mouse and anti-rabbit secondaries (Li-Cor, 1:15,000). Primary antibody incubation was overnight, secondary for 1-2 hr. Blots were visualized and quantified using a Li-Cor system.

3.4.3 Meiotic cell staging

The meiotic stage of a cell was determined based on its DAPI morphology by fluorescent microscopy. 200 cells were counted per strain per time point.

3.4.4 Heterozygous RP-GFP imaging

After two hours in SPO media, 100 uL of cells were placed in a CellASIC ONIX Microfluidic Plate (Y04D) and maintained with fresh SPO media at 2psi using the CellASIC ONIX Microfluidic Perfusion System (CellASIC Corp., Hayward, CA, USA). The cells were held at 30 C using a thermostatic system for the microscope stage.
Cells were imaged using a DeltaVision microscope with a 60x/1.42 oil-immersion objective (DeltaVision, GE Healthcare, Sunnyvale, CA) and filters: FITC (EX475/28, EM525/48) and mCherry (EX575/25, EM625/45). Images were acquired using the softWoRx software (softWoRx, GE Healthcare) with z stacks of 8 slices with 0.5 µm spacing. For time lapse imaging, images were taken every 20 minutes for 12 hours. After 24 hours, images were taken from the same points that had been imaged during the time-lapse portion, as well as points from the same wells that had not been imaged previously.

Images and information can be found in the ‘20160311’ imaging folder. Strain 7085 (Rpl29) was image #s 1-3 and 6-15; strain 7087 (Rpl26b) was image #s 4, 5, and 16-26. Files that are named as 20160308-ae-rplgfp01_01_R3D_D3D are the movie files, the ones that have been converted to .avi files are the most representative movies. Files that are named as 20160308-ae-rplgfp-24h_01 are the 24 hour time points of the points that were imaged in the movie. Files with 24h_c are Rpl26b points that were not imaged during the movie and 24h_d are Rpl29 points that were not imaged during the movie.

3.4.5 Heterozygous RP-GFP image quantification

All images were deconvolved using softWoRx software accompanying the DeltaVision microscope, and maximum-intensity projections were generated using ImageJ/FIJI image processing software (RRID: SCR_002285; (Schneider et al., 2012). Mean intensity of signal from the cells was measured using the “measure” tool in FIJI, and was divided by the background signal from the same image.

3.4.6 Data and Software Availability

Genome-wide data analyzed here were generated previously (Cheng et al., 2018). In short, mRNA levels were assayed by mRNA-seq, translation measurements were assayed by ribosome profiling, and protein levels were assayed by quantitative mass spectrometry (TMT10) for 8 time points during the meiotic program and two vegetative time points (one in rich media and one in sporulation media matched to meiotic samples). All measurements showed high reproducibility, with R values ranging from 0.935 to 0.992. All original data can be found at NCBI GEO: GSE108778 and MassIVE: MSV000081874.

CHAPTER 4

Conclusions and Future Directions

4.1 Gene regulation in yeast is complicated

While budding yeast has been one of the most widely studied organisms, it is clear that there are still aspects of its gene regulation and genome structure that are not fully understood. Simply knowing the full genome sequence is not enough for us to interpret all of the changes that can take place in the cell. With more knowledge we may eventually be able to understand all aspects, as all the information that defines cellular function must be encoded in the DNA genome. Through small steps in identifying gene locations, protein interactions, and cellular pathways, we can begin to construct the entire web of regulation within the cell. One layer that was further uncovered and confirmed through the research presented here is the existence of non-AUG initiated extended protein isoforms. Through identifying these cases, several interesting aspects of gene regulation in meiosis were illuminated, leading to potential areas of further research relating to these findings.

From performing TIS-profiling and annotating initiation sites based on this dataset, it became clear that there are many genes and annotated TISs that are likely incorrect. There are often explanations for why these misannotations exist, such as strain differences and sequencing errors. Some, however, were based on prior assumptions that the TIS is the first AUG on a transcript that is in frame with the stop codon that would produce the longest protein. Many genes were correctly annotated based on these computational predictions, but additional evidence, such as from the TIS-profiling dataset and individual validation could be used to reannotate many other cases. This type of analysis would also be aided by more accurate and complete transcript annotations, which have been performed using TL-seq in several meiotic time points in yeast (unpublished data, Van Werven lab). Combining the two datasets – transcription start site and translation initiation site – could greatly increase our ability to systematically and accurately annotate ORFs. Having more accurate ORF annotations is important for all types of future studies, as knowing where a protein starts can influence experimental design and interpretation.

Another category of interest for follow up is cases of internal initiation – either in-frame, leading to a truncation, or out-of-frame, leading to a short internally translated ORF. Truncated ORFs are an especially promising group, due to the potential to synthesize portions of already known, functional proteins. We have identified several genes that are likely to have a truncated ORF and are following up on several individual cases, as well as looking into the properties of the truncations as a class. Potential functions for truncations may be easily predicted, especially if the truncated portion isolates certain functional domains of the protein while excluding others. The potential for function for internal ORFs that are out-of-frame is less clear, but looking at individual cases to observe their expression, localization and binding partners could be done to understand

what their purpose in the cell is. These cases were not evaluated in depth here mainly due to the difficulty in systematically identifying true positives while excluding both false negatives and false positives. We believe that many false positives exist in the ORF-RATER calls due to signal from elongating ribosomes that were inhibited by LTM, but that these may be discernable from true positives using additional analysis of the TIS-profiling dataset.

The last main category that was further emphasized as being of interest is ORFs that initiate from non-AUG TISs. This includes the extended ORFs studied here, but more broadly includes upstream ORFs (uORFs), as well as other types of ORFs, such as downstream ORFs that initiate at non-AUG sites. The observation that extended ORFs are enriched for the start codons that have been shown previously to be used most efficiently initiated indicate that they are not simply background or random initiation. This type of analysis could be performed on other categories of ORFs such as uORFs, which could also help prioritize cases likely to represent true and functional translation.

The relatively simple genome of yeast makes it a great starting place for understanding cellular mechanisms for modulating which proteins are made and when. Principles and modes of regulation learned from budding yeast can then be applied to other organisms and systems, either to understand function or to modulate function in a way that is useful. Making alternative protein isoforms through non-AUG translation initiation sites has not been thought to be common, and therefore has not been explored thoroughly. The potential, even if minor, role these isoforms could have on modulating the proteome is important to keep in mind, especially in the context of changing cellular conditions. This mode of regulation could be leveraged synthetically to generate low levels of alternatively targeted proteins. For instance, if a protein is normally localized to the cytoplasm, adding an in-frame non-AUG TIS prior to a mitochondrial targeting sequence could generate a lowly expressed mitochondrial-localized form of that protein. To be able to use non-AUG extensions in this way, however, more needs to be learned about the mechanism and specificity of their usage.

4.2 Gene regulation in yeast meiosis is even more complicated

Under steady state conditions, the cell needs to maintain cellular processes and ensure genes are expressed in their correct amounts. When any perturbation occurs, then the cell uses pathways to respond to the change and allow for adjustments to be made. There are many types of stresses that can trigger this, including limited nutrients or altered temperature, chemical, and mechanical conditions. The way the cell responds is critical for the cell's survival, and so both the amount and type of response need to be executed precisely.

Meiosis is one of the most complex processes that a yeast cell undergoes, and it is necessary to generate genetic diversity. This developmental program accurately and efficiently creates haploid gametes from a diploid organism, which can then go on to combine with a different haploid to form a new diploid. To complete meiosis, a yeast cell must coordinate a variety of morphological and functional changes with high accuracy.

Based on the limited nutrients available to the cell during meiosis, there are lower overall translation levels as the cell only makes proteins that are necessary. By saving translation for the pathways and genes that are necessary, the cell is very effective in using the resources it has to drive the complex series of changes needed to create viable gametes. Any gene that is being expressed is then likely to be critical for the cell, as energy is not being used unnecessarily.

To have the cell undergo the stages of meiosis, cellular events such as spore wall formation and symmetrical partitioning of organelles must be properly executed and coordinated to ensure the production of fit gametes. To accomplish this, gene regulation is very precisely regulated. Because yeast have such a simple genome structure, there are various ways the cell can change to accommodate the necessary regulatory mechanisms needed to alter the proteome to undergo meiosis. In addition to what has been shown for modulating the transcription start sites and producing longer transcript isoforms in meiosis, using non-canonical translation initiation sites is an added layer to create more protein diversity in the cell. While there may be additional modes of regulation that the cell can employ, it is likely that some of the most complicated and extensive regulation takes place during meiosis, making it a useful context to identify possible regulation mechanisms that may be broadly used during changing cellular conditions.

The regulatory mechanisms that yeast use during meiosis can then be further applied to understand the possible modes of regulation other organisms have, but may only use during certain conditions. Since meiosis is a highly conserved process, it is likely that at least some of the phenomena that occur in yeast meiosis also occur in higher eukaryotes, such as humans. However, the more complex gene structures, especially with alternative splicing, make it more difficult to both detect and understand the effects of these modes of regulation. By understanding what the cell is capable of through studying gene regulation in yeast meiosis, we can then identify cases in other organisms or conditions where these mechanisms are used, giving a more in depth and complete picture of how gene regulation takes place.

4.3 Near-cognate codon usage could provide more modes of regulation, especially under changing conditions

While many AUG-initiated proteins are critical for cells to survive and grow, the additional layer of non-AUG initiation leading to translated proteins provides even more options for the cell to modulate what is being made and when. It is still not clear what the purpose of making non-AUG ORFs is, but the cell is producing them, which is unlikely to happen if they have no function at all. Additionally, they are produced in a regulated manner, indicating a need at specific times and conditions. When the ribosome initiates at a non-AUG codon upstream of a canonical AUG, there is no obvious distinction between making an extension or a uORF. For this reason, it is possible that only one of these categories is being made intentionally, while the other is more of a side effect. There are non-AUG ORFs that are being translated, based on ribosome profiling studies, but the effect of individual non-AUG uORFs has not been

established. Still, either or both has the potential to greatly increase the possible ways the cell can modulate what it is making, as there are many non-AUG TISs in the genome that have not been previously thought to be used in a meaningful way.

There are many factors that have been shown to influence non-AUG initiation, including the initiation factor eIF5A, which was shown here to lead to higher levels of non-AUG usage in meiosis due to its decreased association with ribosomes in this condition. It is thought that this is occurring because eIF5A facilitates elongation through difficult-to-translate codons, and so when it is absent, there is ribosome queuing leading to a higher propensity to initiate at a less optimal initiation site (Gutierrez et al., 2013; Ivanov et al., 2018; Manjunath et al., 2019). Other proteins proposed to influence near-cognate initiation have also been thought to increase the amount of time that the ribosome is around the non-AUG TIS (Kearse et al., 2019; Martin-Marcos et al., 2011; Nielsen et al., 2004; Takacs et al., 2011). The increased resident time at non-AUG start codons is a likely mechanism of increasing initiation at these sites.

Identifying the factors that influence near-cognate usage, especially in the context of meiosis, is important for fully understanding how and why these non-canonical initiation sites are used. Looking for proteins that change between vegetative and meiotic conditions can be especially helpful, since we know that there is increased non-AUG usage in meiosis. Similar to eIF5A, which was disenriched in mass spectrometry of proteins that associate with initiating ribosomes in meiotic versus vegetative cells, there were other proteins that changed, including Gcd6, which is enriched in meiotic samples. Initial testing of Gcd6 alone did not show large changes to non-AUG-initiated uORF reporters, but testing with additional reporters or in combination with other factors could elucidate its potential role in near-cognate usage. There are also other factors that have been shown in the literature to affect near-cognate usage, such as the DEAD-box helicase Ded1, which in combination with its paralog, Dbp1, may be playing a role in near-cognate usage in meiosis (Guenther et al., 2018).

In addition to *trans* factors, there are likely also *cis* factors that influence near-cognate usage, which could work in conjunction or separately from the *trans* factors. The codon itself can influence how likely it is for a ribosome to initiate, and certain codons, including CUG and UUG have been shown to lead to the highest initiation rates (Kolitz et al., 2009). However, these rates are still much lower than that for AUG TISs, indicating that there are likely more factors than simply the codon contributing to the efficiency of initiation. Sequences surrounding the TIS, also known as the codon context, have also been shown to play a role in codon selection by the ribosome. We did not observe a strong positive context around non-AUG extension TISs, but there could still be specific sites that are in especially good context that are more likely to be used. Lastly, the structure of the surrounding RNA can also play a role in initiation. For example, if there is a stable RNA structure directly after the non-AUG TIS, such that the ribosome is stalled by the structure, it may linger over the non-AUG TIS, and be more likely to initiate. Little is known about how RNA structure changes in meiosis, and performing analysis of RNA structure in meiosis could provide valuable insights into whether RNA structure contributes to generally increased or specific non-AUG initiation.

4.4 Extension ORFs initiated from non-AUG codons may diversify the meiotic proteome

We have now identified a number of extension ORFs in yeast, and their production is enriched in meiosis. However, their functions have not been established. Individual extensions have been shown to have specific functions in a non-meiotic context, either by changing localization or function (Heublein et al., 2019; Tang et al., 2004). Although these types of changes were not detected for the other extensions investigated here, it does not exclude the possibility that there is a function associated with the extended isoform. One possibility is that there is redundancy of the function of the extension isoforms, so there would not be an effect of getting rid of just one. There may also be effects that are only observed under certain conditions, such as stresses, including aging. Looking at the function of extensions under different sensitizing conditions could help elucidate their role in the cell. One important consideration in understanding the function of the extensions as a class is that there may only be a subset of extensions that are made that are important for the cell, especially dependent on the condition. Determining the function as a class but also the function of individual extensions could help in understanding whether there is a general benefit to making extended isoforms or if there are only a few that are beneficial, and the remainder are made as a byproduct of the overall increase in non-AUG initiation observed.

To specifically determine the effect extensions may have on meiosis, a screen could be performed to identify extensions that, when manipulated, result in a phenotype, indicative of their function normally. One straightforward phenotype to look at would be the effect on meiotic timing and efficiency. Here, cells that are not able to make the extended protein could be assayed for their ability to complete meiosis, with strains that were not able to sporulate indicating a role for the extension during meiosis. Other phenotypes that would be interesting would be relating to the function of the individual extension being studied, as well as the normal localization. For instance, identifying extensions that have an effect on mitochondrial function in meiosis if the canonical form is not normally localized there could be a way to assess the extensions that are functioning in the mitochondria. Other gene-specific assays could be developed to assess whether the described function of the gene is different between the extension and canonical isoforms. Understanding the function of the extensions would greatly enhance our understanding of why the cell is synthesizing these non-canonical isoforms, and is an important area of further exploration.

Observing the localization of the extension protein relative to the annotated protein is also a way to infer function of the extension, especially with regard to its differences with the annotated protein. Here, a GFP tag was added to the C-terminus of the protein, which allowed for microscopy of the *WT, M1A* and *ustop* mutants to inform changes in localization. However, no obvious or interpretable changes were observed for the genes that were looked at in this way. One major issue in looking at the localization of the extension protein in *M1A* cells without deleting *UPF1* was that the extension levels were incredibly low for some of the cases. With the *upf1*Δ, the levels may be high enough to observe the localization, which has not been tested yet. Co-localization with known

organelle markers (such as *CIT1* for mitochondria) could also be helpful in identification of the localization pattern for the extension proteins. Additionally, a split-GFP approach similar to (Monteuuis et al., 2019) could be used to specifically detect whether there is protein localized to a location of interest.

Another possible function of adding an extended region to a protein could be to alter the protein-protein interactions. This function is more difficult to predict than localization changes, but could greatly alter the protein's function, allowing a different set of interactions to occur when the extension is made. To test changes in interactions, immunoprecipitations and mass spectrometry could be performed on cells containing *WT, M1A* and *ustop* constructs. Comparisons between what proteins associate with which samples would be indicative of what interactions are being altered.

4.5 Meiotic ribosome degradation could indicate specialized functions

We know that gene regulation changes during meiosis, and from this work we also know that ribosomes are degraded and resynthesized at the end of meiosis. Based on this, an appealing idea is that ribosomes are specialized for the functions the cell requires during meiosis that are not needed in other cellular states. Because synthesizing ribosomes is very energy intensive for the cell, it is unlikely that this turnover would happen if there were not important reasons for the cell to do it. One possible reason the cell might want to get rid of its old ribosomes is in case they have acquired any damage or mutations. It has been shown that after meiosis, haploid gametes have the same lifespan regardless of the age of their diploid progenitor cell (Ünal et al., 2011). Through meiosis, any age-associated damage is cleared and not inherited into the gametes, which is a way to ensure that the reproductive products are as healthy as possible, so they have the best chance to continue passing on their genetic material. If ribosomes accumulated damage in the same ways that mitochondria or other cellular components can, it is still possible the cell would want to resynthesize fresh and functioning ribosomes, again, to make sure the gametes are as happy as possible.

Another possible reason for ribosome replacement following meiosis is if there is some difference in the ribosomes during meiosis that are not optimal for translation under non-meiotic conditions. The non-AUG usage that occurs in meiosis could also be connected to this phenomenon, as this is a mostly unexplained feature of meiotic cells that the ribosome plays a role in. If the ribosomes were different in some way such that they were more likely to initiate at non-AUG start sites, this would help in understanding both the ribosome degradation as well as the non-AUG start site usage observed in meiosis. This connection is very interesting, and would raise further questions regarding the mechanisms of how the ribosome is altered in a way that would lead to its initiation at near-cognates more frequently. The "specialized ribosome" hypothesis has been controversial, but if this were found to be occurring in meiosis, would provide additional support for the idea that ribosomes can be different in different conditions (Genuth and Barna, 2018; Mills and Green, 2017).

It is still unclear whether the entire ribosome or just certain subunits are degraded and resynthesized during meiosis. We observed that at least three ribosomal proteins, from both the large and small subunit, were degraded and resynthesized. However, these are all proteins that tolerate GFP tags, which not all subunits do, making it difficult to look at each individual subunit. Initial work looking at known ribophagy factors did not show that this was playing a major role in the degradation of ribosomes in meiosis. Further investigation into the pathways that are controlling degradation could be useful to understand how the ribosome's degradation is being handled by the cell and whether it would require the whole ribosome to be degraded at once or not.

REFERENCES

Andrews, S.J., Rothnagel, J.A., 2014. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet 15, 193–204. doi:10.1038/nrg3520

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. doi: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. doi:10.1126/science.1215110

Brent, M.R., 2005. Genome annotation past, present, and future: how to define an ORF at each locus. 15, 1777–1786. doi:10.1101/gr.3866105

Chen, Jin, Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., Weissman, J.S., 2020. Pervasive functional translation of noncanonical human open reading frames. Science 367, 1140–1146. doi:10.1126/science.aay0262

Chen, Jingxun, 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, 823. doi:10.7554/eLife.27417

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. doi:10.1016/j.cell.2018.01.035

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, 331. doi:10.7554/eLife.27420

Christiano, R., Nagaraj, N., Fröhlich, F., Walther, T.C., 2014. Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. CellReports 9, 1959–1965. doi:10.1016/j.celrep.2014.10.065

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., Herskowitz, I., 1998. The transcriptional program of sporulation in budding yeast. Science 282, 699–705. doi:10.1126/science.282.5389.699

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. doi:10.1128/mcb.8.10.4533

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., Menschaert, G., 2013. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics 14, 648. doi:10.1186/1471-2164-14-648

Davis, C.A., Grate, L., Spingola, M., Ares, M., 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. 28, 1700–1706. doi:10.1093/nar/28.8.1700

Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., Roucou, X., 2017. Small proteins

encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. Proteomics 1700058–37. doi:10.1002/pmic.201700058
Dever, T.E., Green, R., 2012. The Elongation, Termination, and Recycling Phases of Translation in Eukaryotes. Cold Spring Harbor Perspectives in Biology 4, a013706–a013706. doi:10.1101/cshperspect.a013706
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. doi:10.1371/journal.pcbi.1000176
Dirick, L., Goetsch, L., Ammerer, G., Byers, B., 1998. Regulation of meiotic S phase by Ime2 and a Clb5,6-associated kinase in Saccharomyces cerevisiae. Science 281, 1854–1857. doi:10.1126/science.281.5384.1854
Duina, A.A., Miller, M.E., Keeney, J.B., 2014. Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics 197, 33–48. doi:10.1534/genetics.114.163188
Evans, T., Rosenthal, E.T., Youngblom, J., Distel, D., Hunt, T., 1983. Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. Cell 33, 389–396. doi:10.1016/0092-8674(83)90420-8
Fresno, M., Jiménez, A., Vázquez, D., 1977. Inhibition of translation in eukaryotic systems by harringtonine. Eur J Biochem 72, 323–330. doi:10.1111/j.1432-1033.1977.tb11256.x
Garreau de Loubresse, N., Prokhorova, I., Holtkamp, W., Rodnina, M.V., Yusupova, G., Yusupov, M., 2014. Structural basis for the inhibition of the eukaryotic ribosome. Nature 513, 517–522. doi:10.1038/nature13737
Genuth, N.R., Barna, M., 2018. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Molecular Cell 71, 364–374. doi:10.1016/j.molcel.2018.07.018
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. doi:10.1038/s41586-018-0258-0
Gupta, R., Sadhale, P.P., Vijayraghavan, U., 2015. SUB1 Plays a Negative Role during Starvation Induced Sporulation Program in Saccharomyces cerevisiae. PLoS ONE 10, e0132350–20. doi:10.1371/journal.pone.0132350
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. Molecular Cell 51, 35–45. doi:10.1016/j.molcel.2013.04.021
Heublein, M., Burguillos, M.A., Vo¨gtle, F.N., Teixeira, P.F., Imhof, A., Meisinger, C., Ott, M., 2014. The novel component Kgd4 recruits the E3 subunit to the mitochondrial α-ketoglutarate dehydrogenase. Mol. Biol. Cell 25, 3342–3349. doi:10.1091/mbc.E14-07-1178
Hinnebusch, A.G., 2014. The scanning mechanism of eukaryotic translation initiation. Annu. Rev. Biochem. 83, 779–812. doi:10.1146/annurev-biochem-060713-035802
Hinnebusch, A.G., 2011. Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol. Mol. Biol. Rev. 75, 434–67– first page of table of contents. doi:10.1128/MMBR.00008-11

Hollerer, I., Barker, J.C., Jorgensen, V., Tresenrider, A., Dugast-Darzacq, C., Chan, L.Y., Darzacq, X., Tjian, R., Ünal, E., Brar, G.A., 2019. Evidence for an Integrated Gene Repression Mechanism Based on mRNA Isoform Toggling in Human Cells. G3 (Bethesda) g3.200802.2018–9. doi:10.1534/g3.118.200802
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. doi: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. doi:10.1016/j.cell.2011.10.002
Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y.W., Siddell, S.G., Brierley, I., 2016. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. PLoS Pathog 12, e1005473. doi:10.1371/journal.ppat.1005473
Irniger, S., Piatti, S., Michaelis, C., Nasmyth, K., 1995. Genes involved in sister chromatid separation are needed for B-type cyclin proteolysis in budding yeast. Cell 81, 269–278. doi:10.1016/0092-8674(95)90337-2
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. Molecular Cell 70, 1–18. doi:10.1016/j.molcel.2018.03.015
Iwasaki, S., Ingolia, N.T., 2017. The Growing Toolbox for Protein Synthesis Studies. Trends in Biochemical Sciences 42, 612–624. doi:10.1016/j.tibs.2017.05.004
Jackson, R., Standart, N., 2015. The awesome power of ribosome profiling. RNA 21, 652–654. doi:10.1261/rna.049908.115
Jackson, R.J., Hellen, C.U.T., Pestova, T.V., 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. Nature Publishing Group 11, 113–127. doi:10.1038/nrm2838
Jan, C.H., Williams, C.C., Weissman, J.S., 2014. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science 346, 716–716. doi:10.1126/science.1257521
Jorgensen, V., Chen, J., Vander Wende, H., Harris, D., McCarthy, A., Breznak, S., Wong-Deyrup, S.W., Chen, Y., Rangan, P., Brar, G.A., Sawyer, E.M., Chan, L.Y., Ünal, E., 2020. Tunable Transcriptional Interference at the Endogenous Alcohol Dehydrogenase Gene Locus in Drosophila melanogaster. G3 (Bethesda). doi:10.1534/g3.119.400937
Kachroo, A.H., Laurent, J.M., Yellman, C.M., Meyer, A.G., Wilke, C.O., Marcotte, E.M., 2015. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348, 921–925. doi:10.1126/science.aaa0769
Kearse, M.G., Wilusz, J.E., 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes & Development 31, 1717–1731. doi:10.1101/gad.305250.117
Kearse, M.G., Goldman, D.H., Choi, J., Nwaezeapu, C., Liang, D., Green, K.M., Goldstrohm, A.C., Todd, P.K., Green, R., Wilusz, J.E., 2019. Ribosome queuing enables non-AUG translation to be resistant to multiple protein synthesis inhibitors. Genes & Development 33, 871–885. doi:10.1101/gad.324715.119
Kim Guisbert, K.S., 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. doi:10.1261/rna.030510.111
King, G.A., Goodman, J.S., Schick, J.G., Chetlapalli, K., Jorgens, D.M., McDonald, K.L., Ünal, E., 2019. Meiotic cellular rejuvenation is coupled to nuclear remodeling in budding yeast. Elife 8, 72. doi:10.7554/eLife.47156
King, R.W., Peters, J.M., Tugendreich, S., Rolfe, M., Hieter, P., Kirschner, M.W., 1995. A 20S complex containing CDC27 and CDC16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin B. Cell 81, 279–288. doi:10.1016/0092-8674(95)90338-0
Kochetov, A.V., 2008. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. BioEssays 30, 683–691. doi:10.1002/bies.20771
Kolitz, S.E., Takacs, J.E., Lorsch, J.R., 2009. Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. RNA 15, 138–152. doi:10.1261/rna.1318509
Kozak, M., 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299, 1–34. doi:10.1016/j.cell.2011.04.006
Kozak, M., 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. 12, 857–872. doi:10.1093/nar/12.2.857
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. CellReports 18, 2729–2741. doi:10.1016/j.celrep.2017.02.053
Lamb, N.E., Yu, K., Shaffer, J., Feingold, E., Sherman, S.L., 2005. Association between maternal age and meiotic recombination for trisomy 21. Am. J. Hum. Genet. 76, 91–99. doi:10.1086/427266
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. U.S.A. 109, E2424–32. doi:10.1073/pnas.1207846109
Leskinen, P., Virta, M., Karp, M., 2003. One-step measurement of firefly luciferase activity in yeast. Yeast 20, 1109–1113. doi:10.1002/yea.1024
Li, J., Liang, Q., Song, W., Marchisio, M.A., 2017. Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast S. cerevisiae. J Biol Eng 11, 25–14. doi:10.1186/s13036-017-0068-1
Lin, J., Zhou, D., Steitz, T.A., Polikanov, Y.S., Gagnon, M.G., 2018. Ribosome-Targeting Antibiotics: Modes of Action, Mechanisms of Resistance, and Implications for Drug Design. Annu. Rev. Biochem. 87, 451–478. doi:10.1146/annurev-biochem-062917-011942
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. CellReports 29, 3134–3146.e6. doi:10.1016/j.celrep.2019.10.129
Martin-Marcos, P., Cheung, Y.-N., Hinnebusch, A.G., 2011. Functional elements in initiation factors 1, 1A, and 2β discriminate against poor AUG context and non-AUG start codons. Mol. Cell. Biol. 31, 4814–4831. doi:10.1128/MCB.05819-11
Mills, E.W., Green, R., 2017. Ribosomopathies: There's strength in numbers. Science

358. doi:10.1126/science.aan2755
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. doi:10.1093/nar/gkz301
Mouilleron, H., Delcourt, V., Roucou, X., 2015. Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic Acids Research. doi:10.1093/nar/gkv1218
Nasif, S., Contu, L., Mühlemann, O., 2018. Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. Semin. Cell Dev. Biol. 75, 78–87. doi:10.1016/j.semcdb.2017.08.053
Nielsen, K.H., Szamecz, B., Valásek, L., Jivotovskaya, A., Shin, B.-S., Hinnebusch, A.G., 2004. Functions of eIF3 downstream of 48S assembly impact AUG recognition and GCN4 translational control. EMBO J. 23, 1166–1177. doi:10.1038/sj.emboj.7600116
Ossareh-Nazari, B., Bonizec, M., Cohen, M., Dokudovskaya, S., Delalande, F., Schaeffer, C., Van Dorsselaer, A., Dargemont, C., 2010. Cdc48 and Ufd3, new partners of the ubiquitin protease Ubp3, are required for ribophagy. EMBO Rep. 11, 548–554. doi:10.1038/embor.2010.74
Sawyer, E.M., Joshi, P.R., Jorgensen, V., Yunus, J., Berchowitz, L.E., Ünal, E., 2019. Developmental regulation of an organelle tether coordinates mitochondrial remodeling in meiosis. The Journal of Cell Biology 218, 559–579. doi:10.1083/jcb.201807097
Schmeing, T.M., Ramakrishnan, V., 2009. What recent ribosome structures have revealed about the mechanism of translation. Nature 461, 1234–1242. doi:10.1038/nature08403
Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. Nature Chemical Biology 9, 671–675. doi:10.1038/nmeth.2089
Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., Saghatelian, A., 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nature Chemical Biology 9, 59–64. doi:10.1038/nchembio.1120
Sourirajan, A., Lichten, M., 2008. Polo-like kinase Cdc5 drives exit from pachytene during budding yeast meiosis. Genes & Development 22, 2627–2632. doi:10.1101/gad.1711408
Stern-Ginossar, N., Ingolia, N.T., 2015. Ribosome Profiling as a Tool to Decipher Viral Complexity. Annu Rev Virol 2, 335–349. doi:10.1146/annurev-virology-100114-054854
Sudakin, V., Ganoth, D., Dahan, A., Heller, H., Hershko, J., Luca, F.C., Ruderman, J.V., Hershko, A., 1995. The cyclosome, a large complex containing cyclin-selective ubiquitin ligase activity, targets cyclins for destruction at the end of mitosis. Mol. Biol. Cell 6, 185–197. doi:10.1091/mbc.6.2.185
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. 45, 1433–1441. doi:10.7164/antibiotics.45.1433

Suomi, F., Suomi, F., Menger, K.E., Menger, K.E., Monteuuis, G., Monteuuis, G., Naumann, U., Naumann, U., Kursu, V.A.S., Kursu, V.A.S., Shvetsova, A., Shvetsova, A., Kastaniotis, A.J., Kastaniotis, A.J., 2014. Expression and evolution of the non-canonically translated yeast mitochondrial acetyl-CoA carboxylase Hfa1p. PLoS ONE 9, e114738. doi:10.1371/journal.pone.0114738

Susor, A., Jansova, D., Anger, M., Kubelka, M., 2015. Translation in the mammalian oocyte in space and time. Cell Tissue Res 363, 69–84. doi:10.1007/s00441-015-2269-6

Takacs, J.E., Neary, T.B., Ingolia, N.T., Saini, A.K., Martin-Marcos, P., Pelletier, J., Hinnebusch, A.G., Lorsch, J.R., 2011. Identification of compounds that decrease the fidelity of start codon recognition by the eukaryotic translational machinery. 17, 439–452. doi:10.1261/rna.2475211

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. Journal of Biological Chemistry 279, 49656–49663. doi:10.1074/jbc.M408081200

Tzani, I., Ivanov, I.P., Andreev, D.E., Dmitriev, R.I., Dean, K.A., Baranov, P.V., Atkins, J.F., Loughran, G., 2016. Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. Open Biol. 6, 150203–13. doi:10.1098/rsob.150203

Ünal, E., Kinde, B., Amon, A., 2011. Gametogenesis eliminates age-induced cellular damage and resets life span in yeast. Science 332, 1554–1557. doi:10.1126/science.1204349

Vazquez-Laslop, N., Thum, C., Mankin, A.S., 2008. Molecular mechanism of drug-dependent ribosome stalling. Molecular Cell 30, 190–202. doi:10.1016/j.molcel.2008.02.026

Wilson, D.N., Doudna Cate, J.H., 2012. The structure and function of the eukaryotic ribosome. Cold Spring Harbor Perspectives in Biology 4, a011536–a011536. doi:10.1101/cshperspect.a011536

Wilson, J.E., Pestova, T.V., Hellen, C.U.T., Sarnow, P., 2000. Initiation of Protein Synthesis from the A Site of the Ribosome. Cell 102, 1–10. doi:10.1016/s0092-8674(00)00055-6

Zitomer, R.S., Walthall, D.A., Rymond, B.C., Hollenberg, C.P., 1984. Saccharomyces cerevisiae ribosomes recognize non-AUG initiation codons. Mol. Cell. Biol. 4, 1191–1197. doi:10.1128/mcb.4.7.1191

APPENDIX

A.1 Properties of extension ORFs

A.1.1 Localization of N-terminal extension isoforms in meiosis

A leading hypothesis for the function of the extension isoforms would be to add a targeting sequence within the extended region, changing the localization of the extension protein relative to the canonical protein. To examine this possibility, genes with extension were tagged with GFP and mutated to only make the extension form (*M1A*) or only make the canonical form (*ustop*). Microscopy of the *WT*, *M1A*, and *ustop* strains would then help determine changes in the cellular localization of the different forms, especially by comparing the *M1A* and *ustop* mutants. This was performed for a number of genes with extensions, and no obvious changes in the localization were observed. For Ymr31-GFP, both the *WT* and *M1A* tagged strains showed mitochondrial localization (Figure A.1). Co-localization with Cit1-mCardinal, a mitochondrial protein tagged with a red fluorescent protein, led to confirmation that both forms are localized to the mitochondria. The lower levels of the Ymr31-M1A extension protein that were observed by western blot (Figure 2.4C) were also observed by microscopy.

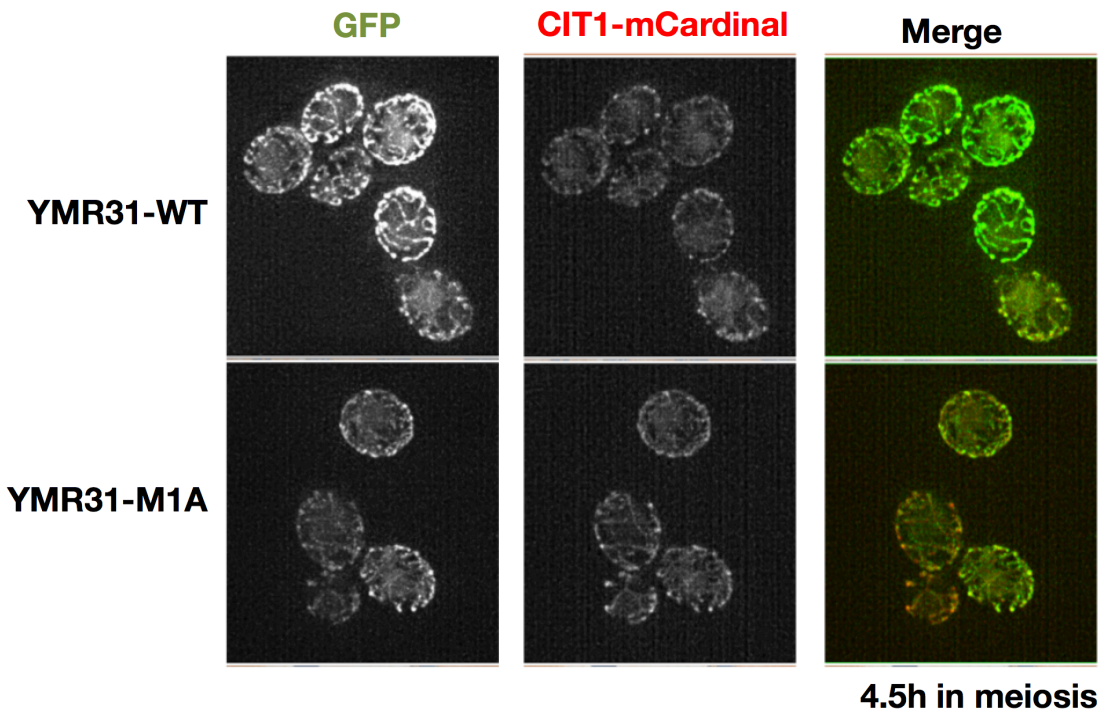

Figure A.1: Microscopy of Ymr31-WT-GFP and Ymr31-M1A-GFP at 4.5 hours in meiosis. Co-localization with Cit1-mCardinal indicates both *WT* and *M1A* strains show mitochondrial localization patterns.

While the extension for *YMR31* was detectable in the *M1A* background, there were other extensions that this was not the case. One way to increase the level of the

extension protein to be able to see it by microscopy is to mutate the non-AUG start codon to AUG. This was done for *HYR1*, and then microscopy was performed on this strain to examine whether there is a difference in localization between the two forms (Figure A.2). No clear difference was seen in this case, when comparing the *WT*, *ustop*, and mutation to mutate the extension to an ATG with the *M1A* mutation (*upM-M1A*).

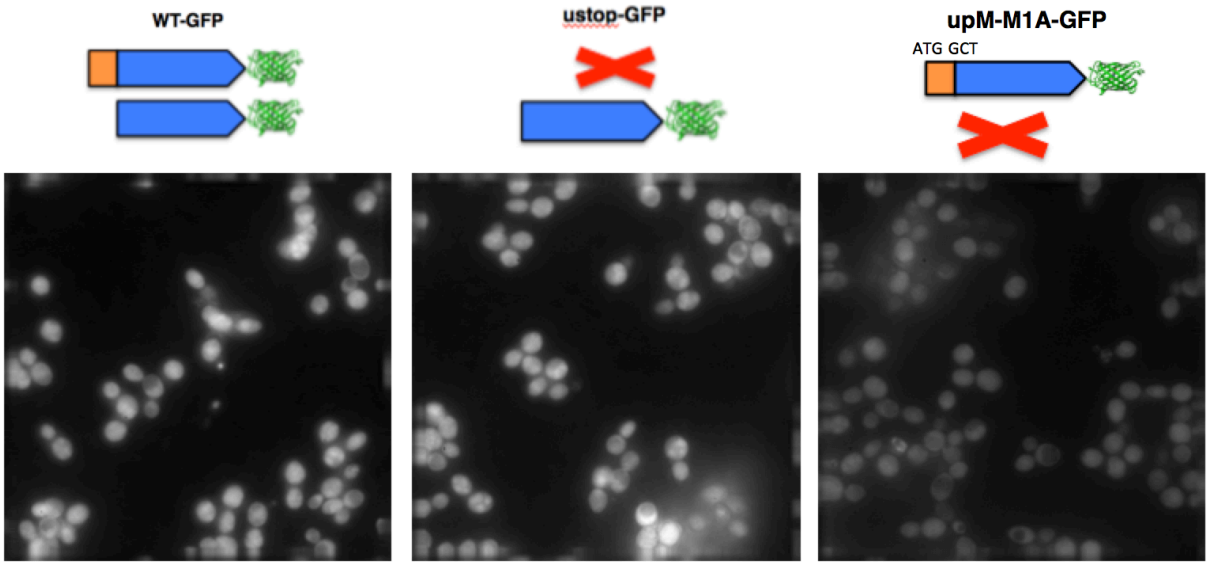

Figure A.2: Microscopy of Hyr1-WT-GFP, Hyr1-ustop-GFP and Hyr1-M1A-GFP at 4.5 hours in meiosis. No detectable change in localization is observed between the different forms.

While no obvious localization change was detected for Ymr31 or Hyr1, there may be other extension isoforms, similar to Ala1, that do get targeted differentially due to the inclusion of a targeting sequence within the extended region (Tang et al., 2004). Predictions of targeting sequences, especially mitochondrial, can be done to identify genes that are likely to have this type of altered localization (Monteuuis et al., 2019). However, predictions may not capture all possible targeting sequence changes, so looking at localization for any extension-containing genes is helpful. The low level of the *M1A* construct is a challenge in comparing localization, and performing microscopy on cells that have more similar levels of protein from the *M1A* construct are more likely to yield helpful results. Still, comparison to a strain that lacks a GFP tag completely can help determine the level of background fluorescence in the conditions being imaged.

A.1.2 Detection of additional extensions in artificial contexts

The strategy of mutating the upstream start codon to an AUG was also performed for *SUB1*, a gene with an extension that was not validated with the *M1A* mutant alone or with the *upf1Δ* to block NMD. The extension protein for *SUB1* is detected in the *upM* and *upM-M1A* mutants, indicating that it is possible for the cell to synthesize an extension from this TIS (Figure A.3). Validating that the extension can be made from its native non-AUG initiation site would help provide support for this extension isoform being made in cells normally.

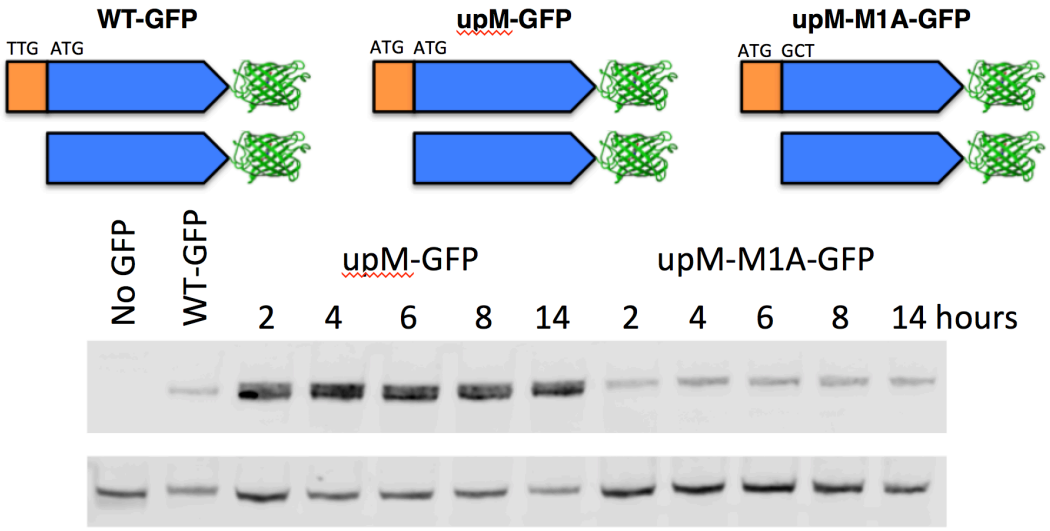

Figure A.3: Western blot analysis of Sub1-GFP with mutations. Mutations change the extension TIS to an ATG (*upM*) and in combination with mutating the canonical ATG to an alanine (*upM-M1A*). When both forms have an AUG start codon, two isoforms are detected by western blot. When only the extension isoform is expected to be made (*upM-M1A*), only the upper band is detected, indicating that the extension can be made isolated from the canonical form.

Another way to increase the level of the extension protein is to put a strong promoter in front of the extension, such that the mRNA levels are increased. This can also be done with an inducible system, so that the increase can be controlled in timing and amount. In the case of *NRG1*, a beta-estradiol (b-Est) inducible system containing a 8xLexO inducible operon and the *CYC1* promoter (pCyc1) was introduced upstream of *NRG1-M1A-GFP*. When induced with b-Est, the mRNA levels for the *M1A* construct were increased about 20-fold higher than *WT*, and the extension protein for *NRG1* was detected (Figure A.4).

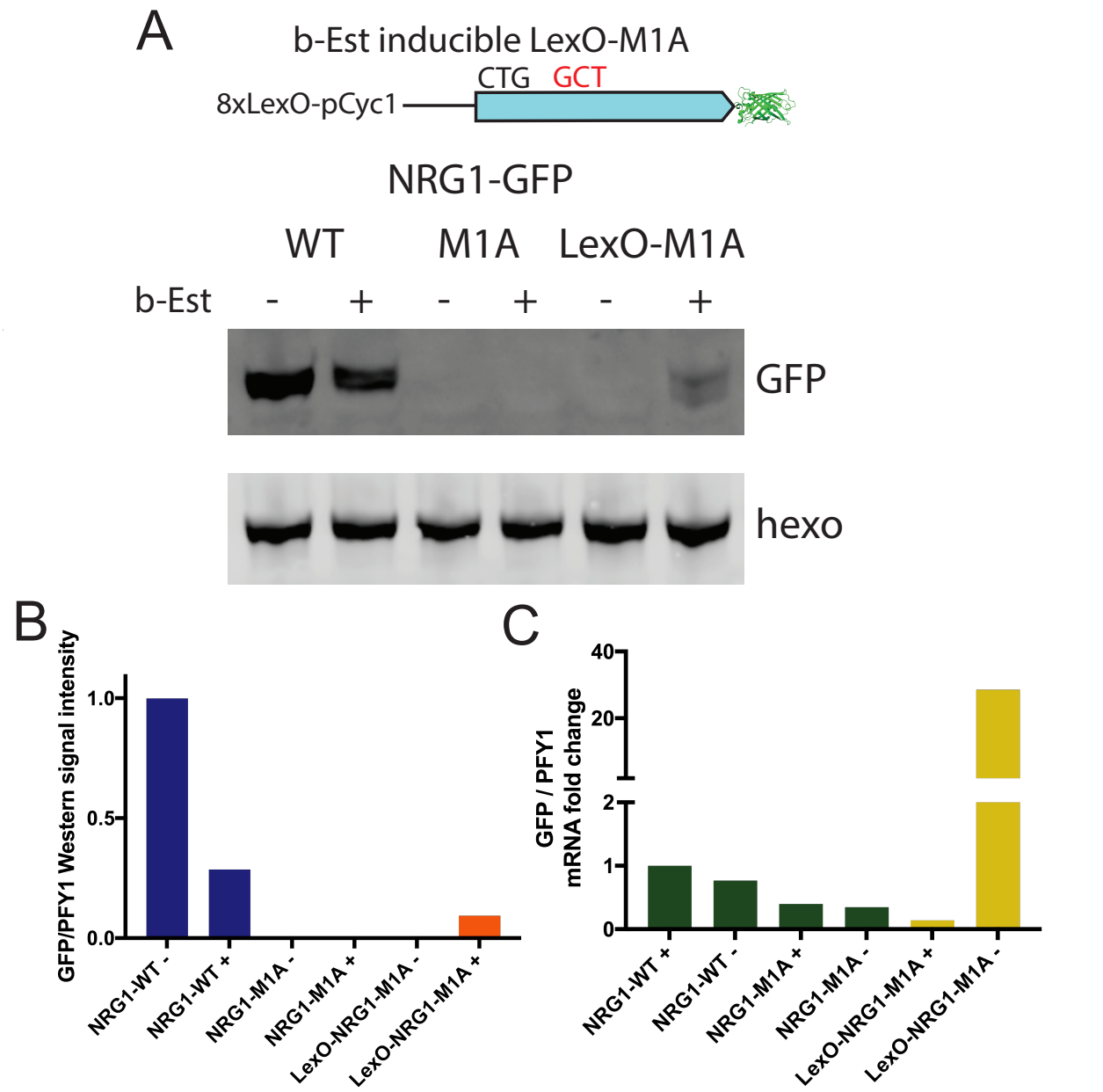

Figure A.4: The extension isoform for NRG1 can be detected when using a strong inducible promoter. (A) Western blot analysis of Nrg1-GFP in strains containing an inducible strong promoter (8xLexO-pCyc1). Upon induction with b-estradiol (b-Est), the extension protein in the 8xLexO-pCyc1-Nrg1-M1A-GFP (LexO-M1A) construct is detectable. (B) Quantification of the western blot showing a low but detectable level of extension protein for the LexO-Nrg1-M1A construct. (C) qPCR of Nrg1-GFP relative to *PFY1*, showing an over 20-fold induction of LexO-M1A mRNA upon b-Est addition (+).

It is clear that for *NRG1*, the extension can be made by cells from the non-AUG initiation site, however, the protein level is extremely low. When the mRNA is increased over 20-fold, the extension protein detected is still less than 10-fold that of the *WT* protein. This explains why, in this case, increasing the level of the mRNA of the *M1A* construct to similar levels as the *WT* by blocking NMD with the *upf1Δ* still did not lead to detection of the extension protein.

A.1.3 Effect of extensions on timing of meiotic progression

Extensions and near-cognate initiation generally are increased in meiotic cells. A prediction would be that if they are important for the cell in meiosis, getting rid of them would lead to effects on meiosis. One effect could be the timing of meiotic progression, which is normally synchronized through different stages and completed within about 12 hours. Extensions with the *ustop* mutation to get rid of the extension were put through meiosis and samples were taken throughout, and then the stage of meiosis was compared to that of *WT* cells. For *YMR31*, no difference was observed in the timing of progression through meiosis when comparing the strains (Figure A.5).

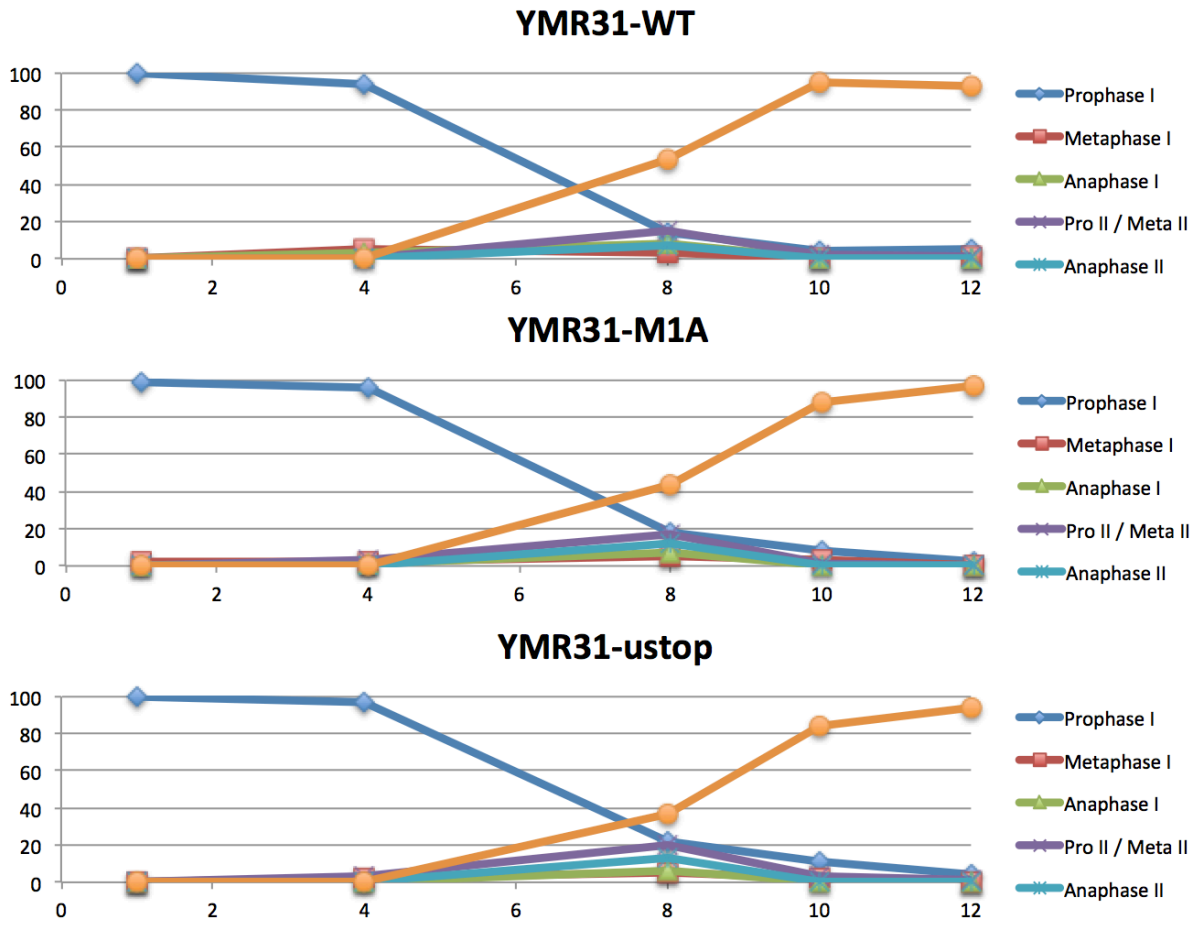

Figure A.5: Timing of meiotic progression for YMR31 strains. Comparison of spindles for WT, M1A and ustop constructs. Cells were fixed and immunofluorescence was performed to visualize tubulin and DAPI, which are markers of spindles and DNA, respectively. Different cell stages were counted at each time point, and comparison of the progression between strains showed no major difference in the timing of cells undergoing meiosis.

Meiotic progression analysis was performed for other extension containing genes (at least *HYR1*) as well, with no significant changes in timing detected. Other extension-containing genes could be analyzed using this method to determine the effect of the extension on meiotic timing.

Sporulation efficiency can also be assessed to determine a mutant's effect on meiosis. To do this, the cells are harvested at the end of meiosis (24 hours), and counted to determine how many underwent meiosis normally and formed tetranucleates. This is in comparison to cells that formed binucleates and mononucleates, which are often cells that did not fully or at all undergo meiosis. For *SUB1*, mutants that get rid of the extension by placing a stop codon before the canonical AUG (*ustop*), mutation of the extension non-AUG TIS (*umut*), or deletion of the entire gene (delete) led to increased sporulation efficiency seen by more tetranucleates (Figure A.6).

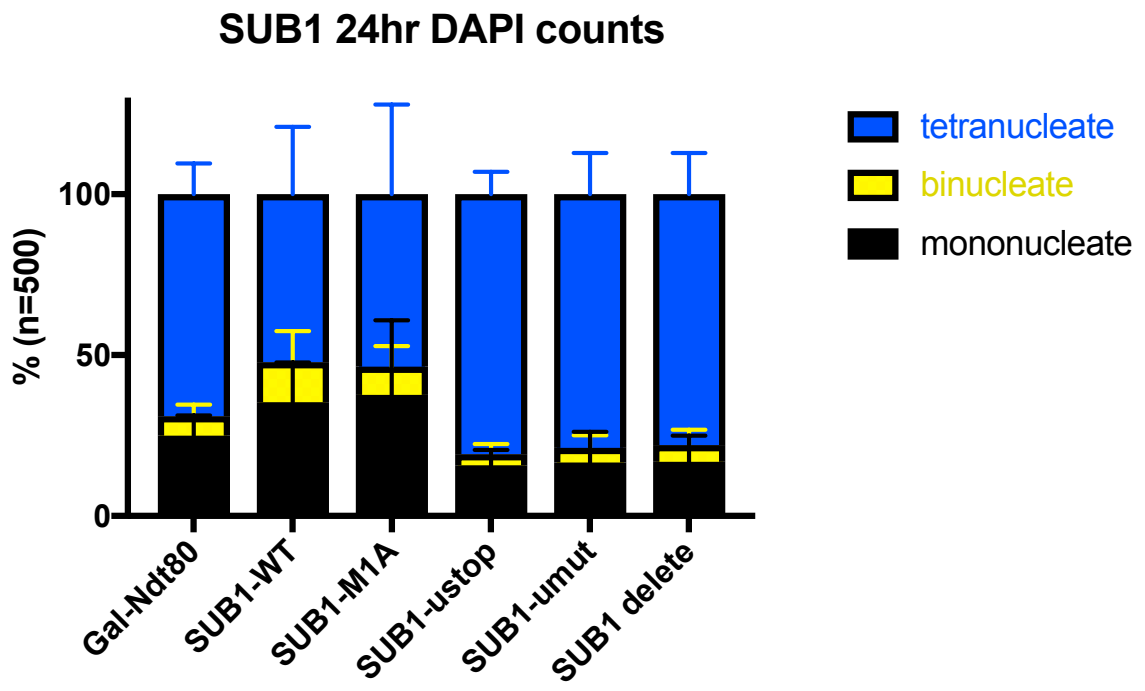

Figure A.6: Sporulation efficiency of *SUB1* strains. At the end of meiosis (24 hours), cells (n=500) were counted to see if they were tetra-, bi-, or mono-nucleates.

The increase in tetranucleates observed for *SUB1 ustop*, *umut* and deletion strains is intriguing, as it would indicate that getting rid of the extension can increase sporulation efficiency. *SUB1* has been proposed to have a negative regulatory role in meiosis, which would be consistent with this finding (Gupta et al., 2015). Further replicates and confirmation of this observation would be necessary to determine the effect of *SUB1* on meiotic efficiency.

A.1.4 Pull downs of extensions to identify changes in binding partners

Another possible change that the extension region could confer to a protein would be its interactions with binding partners. To assess changes in binding partners, pull downs can allow for comparison of different mutants and mass spectrometry of those pull downs can then identify the proteins that are changing. Conditions for the pull down should be optimized based on the specific protein, with buffer and lysis conditions being important, especially to prevent degradation of the protein.

Pull downs of Ymr31-GFP and Hyr1-GFP were optimized, but did not show differences and were difficult due to the low levels of protein from the *M1A* mutants. Between bead beating, mixermilling and using the Fast-prep for lysis, there seemed to be little difference, with the Fast-prep leading to the least degradation. Following the pull downs, coomassie staining of the input and eluate helped identify if there were other proteins being pulled down, and if so, if there were differences between the strains. If there are known interactors, checking that there are bands at the expected sizes would be a good indication that the pull down is working well. Then, mass spectrometry can be performed on the entire eluate sample to look for changes between the strains.

A.1.5 Other assays to determine function of extensions

To determine if an extension has an effect on the known function of a protein, assays can be developed based on the known function. If the extension is affecting this function, differences would be expected between the *WT*, *M1A* and/or *ustop* strains. For *YMR31*, which is a mitochondrial protein, growth in ethanol or at high temperatures could alter the function. No differences were seen in the growth rate, either by plate or liquid growth assay, between the different mutants (Figure A.7).

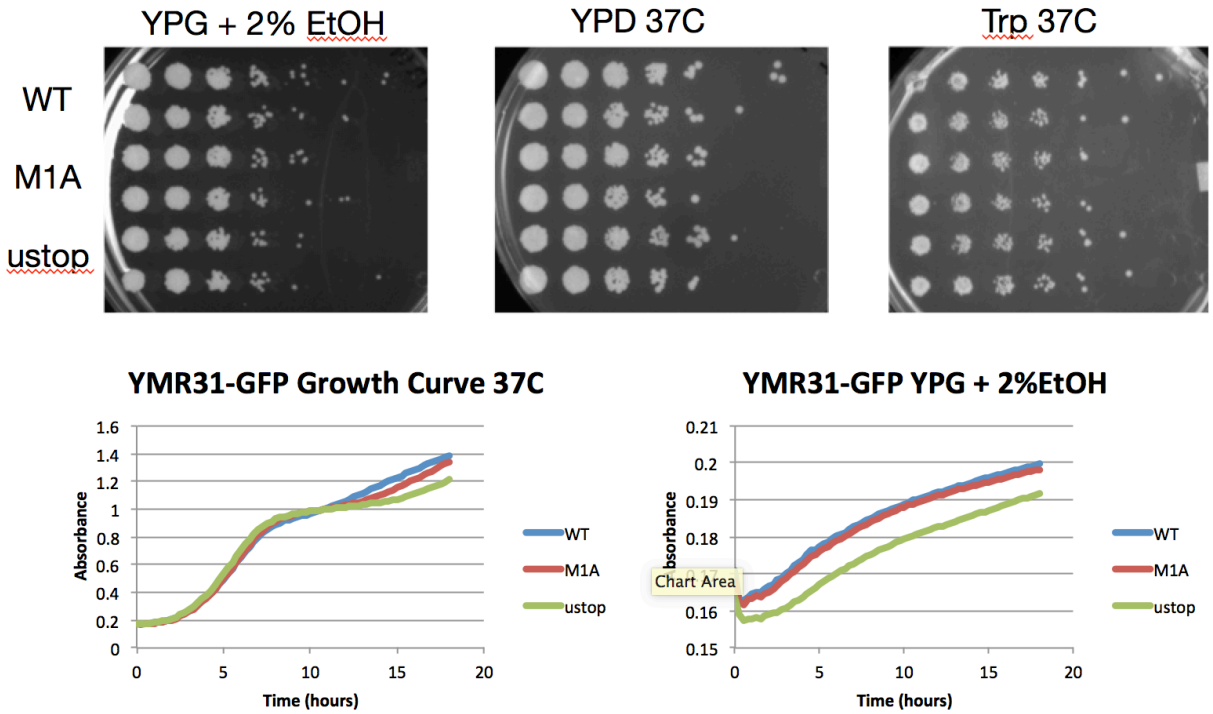

Figure A.7: Plate dilution assays for YMR31 strains. Comparisons of YMR31-WT, YMR31-M1A and YMR31-ustop strains on YPG + 2% EtOH, YPD at 37C, and Trp at 37C (top) and liquid growth assays in similar conditions to determine if the different constructs conferred different growth effects on the strains.

Other targeted assays based on the known function of the extensions could be useful to determine the effect of not having the extension isoform in cells. Developing the right assay for each gene is challenging but could yield important revelations about the function of the extensions individually and more broadly as a class.

A.1.6 Factors affecting the stability of extensions

One possibility for the lower level of the extended protein isoforms is that the stability of the protein is altered. The effect of the proteasome was tested using MG132 on cells with the Ymr31-GFP-M1A construct in the *upf1Δ* background (Figure 2.S6B). Other pathways could also affect the stability of the extension isoform specifically, and blocking these pathways and testing the effect on the extension protein could elucidate the pathway's contribution to the lower levels.

One pathway is the N-end rule pathway, of which *UBR1* is an E3 ubiquitin ligase involved in recognizing and ubiquitinating pathway substrates. Testing with a ubr1::KanMX deletion with some of the M1A mutants did not show any effect in increasing the level of protein, but testing the M1A mutants in the *upf1Δ* background has not been done and is more likely to show an effect since the protein is already detectable in these conditions. In addition to using MG132 to block the proteasome, a temperature sensitive mutant of *RPN6*, which is a regulatory subunit of the proteasome lid, could also be used to assess the effect of proteasome degradation on the extensions. Further testing with both MG132 and the rpn6-ts allele with additional extension mutants in the *upf1Δ* background would further confirm that the proteasome is not playing a role in degrading the extensions. Lastly, *UBC4* is the E2 ubiquitin-conjugating enzyme for the anaphase-promoting complex, which also plays a role in degradation of proteins. Testing with the ubc4::KanMX strain would help determine if this pathway is involved.

A.2 Factors that influence near-cognate usage in meiosis

The increased near-cognate usage that is observed in meiosis likely has a combination of *cis* and *trans* factors contributing to the effect. Identification of eIF5A (encoded by *HYP2* in yeast) was demonstrated to play a role in increasing near-cognate usage in meiosis (Figure 2.7). However, this factor alone did not account for all of the increase, and other trans factors have been investigated that could be working in addition or in combination with eIF5A.

A.2.1 *GCD6* is a trans factor that could affect near-cognate usage

From the mass spectrometry of 40S and 60S ribosomal subunits, *GCD6* was also identified as being significantly enriched in meiotic cells relative to vegetative (Figure 2.7B). *GCD6* is the catalytic epsilon subunit of eIF2B, and has been shown to be regulated by phosphorylated eIF2. To test whether this factor plays a role in near-cognate initiation, overexpression constructs were tested in vegetative cells and degron constructs were tested in meiotic cells. Much of the work done on this project was completed by Sarah Guo, and can be found in her senior honors thesis from May 2017. An inducible overexpression construct, pGAL10-3V5-GCD6, overexpresses *GCD6* in vegetative cells (Figure A.8). An auxin-inducible degron allele (GCD6-IAA7), shows depletion in vegetative cells (data not shown). However, this strain does not undergo

meiosis, meaning a different tag or system would be better to deplete this protein in a meiotic context.

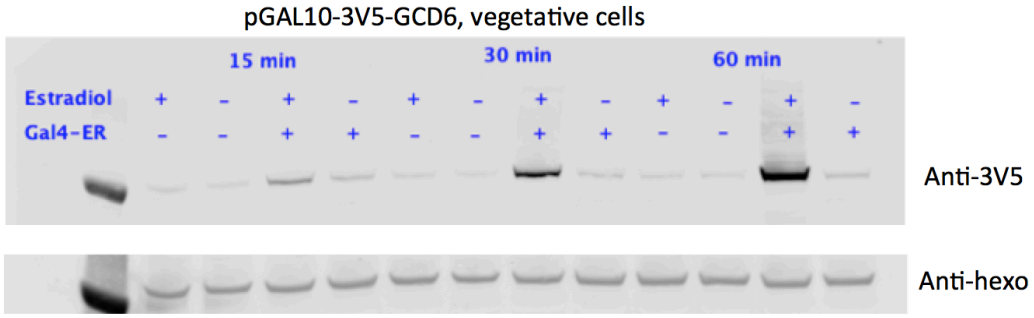

Figure A.8: *GCD6* overexpression in vegetative cells. Samples were taken 15, 30, or 60 minutes after addition of estradiol to induce the Gal4-ER system to induce the pGAL10 promoter driving expression of a 3V5-tagged *GCD6*. Protein is detected for the samples with estradiol added and the Gal4-ER most efficiently, indicating that the overexpression is working.

A.2.2 Other trans factors that could influence near-cognate usage in meiosis,

DED1, an ATP-dependent RNA helicase that is lower in abundance in meiosis, has been shown to play a role in near-cognate usage (Guenther et al., 2018). It also has a paralog, *DBP1*, which is higher in meiosis, although ribosome profiling of a *dbp1Δ* did not show obvious changes in near-cognate usage (unpublished data). Further work with this set of paralogs is being done by Emily Powers in the lab, and the constructs and mechanisms being worked out by her could aid in understanding these proteins' role in near-cognate usage in meiosis. eIF2A has also been thought to increase initiation from near-cognate-initiated internal intron TISs (unpublished data, Johnson lab), and could be examined at to evaluate its role in meiotic near-cognate usage as well. To test the effect of several trans factors, constructs that alter the level were combined in strains with the *HYR1-M1A* construct. If any affected near-cognate translation, the level of the protein would be expected to change. For all of the constructs tested, no change was observed when tested in vegetative cells (Figure A.9). The experimental set-up and measurements here were slightly different than that of the HYP2 experiments (Figure 2.7C), so testing these additional trans factors in the same system as was done for HYP2 would be very worthwhile to see if there is an effect on near-cognate initiation.

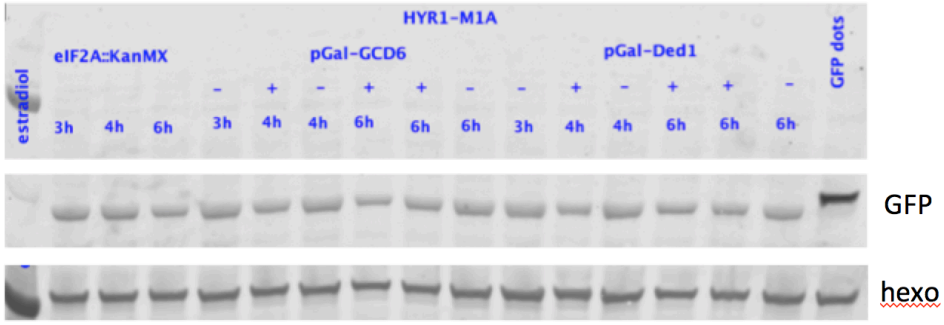

Figure A.9: Western blot of different trans factors and their affect on translation of the *HYR1-M1A* construct. No major changes were observed when eIF2A was depleted, or *GCD6* or *DED1* were overexpressed, all in vegetative cells.

A.2.3 Luciferase uORF reporter to detect changes in near-cognate usage

In addition to using the *M1A-GFP* constructs as a readout for near-cognate usage, several near-cognate-initiated uORFs have been made. The advantage of using a luciferase reporter is that the detection can be done in live cells, and the quantification can be very accurate (Leskinen et al., 2003). Initial testing using this reporter system was not promising, as there were very low reads for all near-cognate uORFs tested, even in meiosis conditions when the levels should be higher than background (Figure A.10).

Vegetative cells – luciferase positive controls

pH	Luciferin	SUP35-luc	YPT1-luc	WT	SC media	blank
3	+	11336	9056	24	26	21
	-	32	23	28	17	14
5	+	1260	1278	18	12	13
	-	17	14	13	15	15

Meiotic cells – luciferase uORF reporters

pH	Luciferin	YPT1-luc	uORF9833-luc	uORF590-luc	uORF3150-luc	uORF3150-out-luc	WT	SPO media
3	+	32	67	21	18	25	20	25
	-	20	17	21	19	22	21	17
5	+	117	1238	15	15	18	16	20
	-	16	20	19	15	14	16	21

Figure A.10: Luciferase reporter testing using vegetative and meiotic cells. Different pHs were used as suggested in the protocol, and measurements were taken using a plate reader.

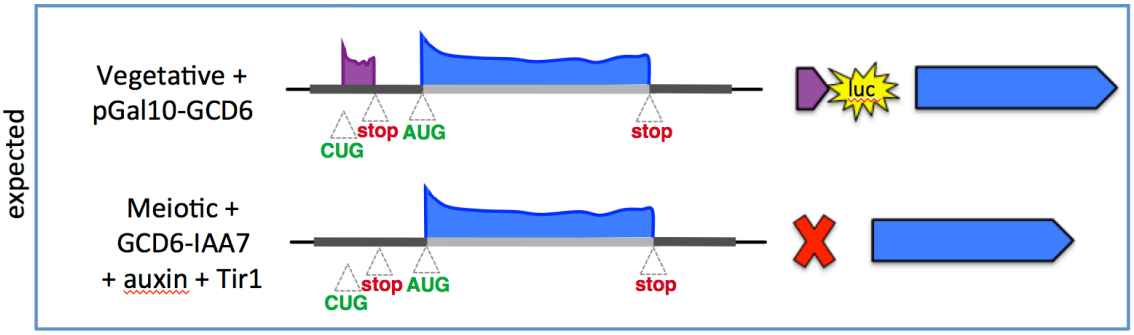

Figure A.11: Schematic for GCD6 modulation and the effect on a luciferase-uORF based reporter. By overexpressing GCD6 in vegetative, the expected result would be detection of a uORF-luciferase reporter that is otherwise not seen in vegetative cells. Conversely, translation of a uORF-luciferase reporter would be decreased in meiotic cells when a GCD6-IAA7 degron strain was added.

A.3 Proximity ribosome profiling

Performing proximity ribosome profiling could also aid in identifying extension ORFs that are alternatively targeted. To perform proximity profiling, based on (Jan et al., 2014), biotin needs to be added to cells to allow for rapid biotinylation of target proteins as they are being translated. Biotin gets into cells very well around pH 7, but in meiosis, the

media shifts to be around pH 8. Additionally, the media to start with needs to be biotin-free, so that when additional biotin is added it is incorporated specifically. This leads to several challenges in performing this experiment in meiotic cells. First, cell growth and meiosis need to occur in biotin-free conditions, which are not ideal. Second, the pH of the media needs to be adjusted to around 7 when the biotin is added. Optimization of several of these conditions was performed, but the combination of biotin-free media and low pH led to a lower sporulation efficiency than would be ideal for performing this experiment (Figure A.12). Further work on these conditions is being done by Jay Goodman in the lab, and it looks like improvements to the system could increase sporulation to normal.

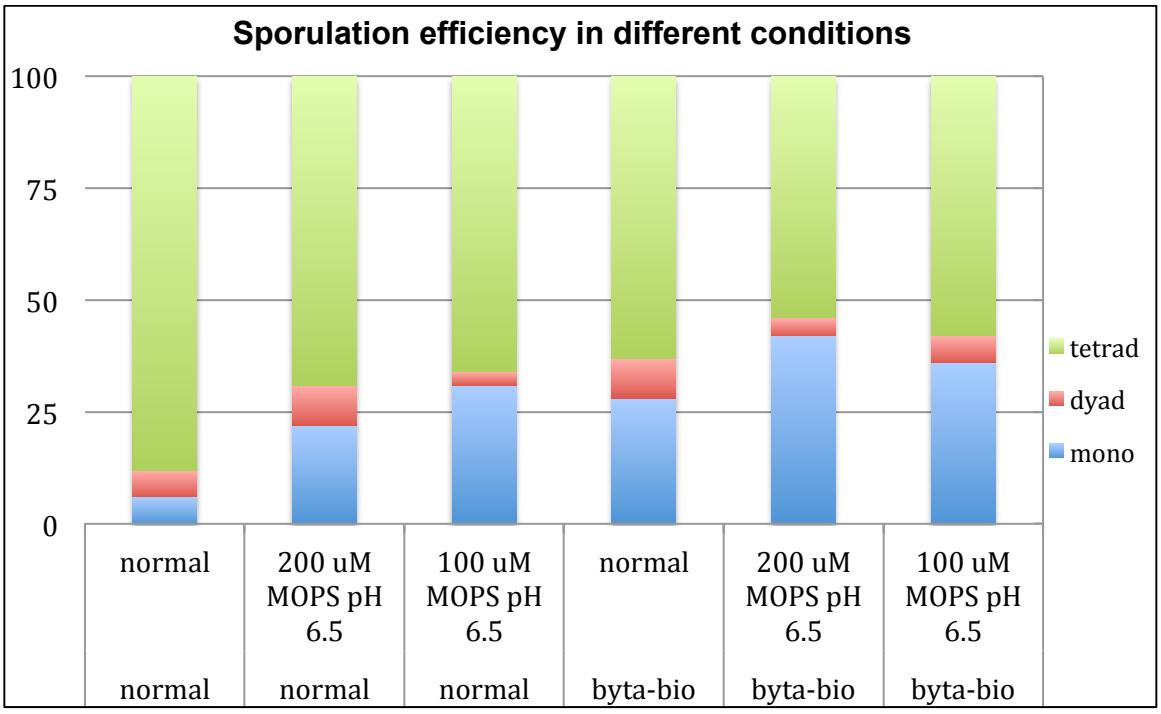

Figure A.12: Sporulation efficiency in different biotin-free growth conditions. Sporulation media (SPO) was adjusted using MOPS buffer at pH 6.5, which led to a low enough pH of the media to expect biotin to get in. Biotin-free BYTA was used as well, in comparison to normal, and showed lower sporulation efficiency.

To test whether the biotin labeling worked, streptavidin can be added to cell lysates prior to running on an SDS-PAGE gel, which binds to the biotinylated proteins and shifts them to a larger molecular weight. This was done for an initial testing of the biotin labeling for a Sec63-BirA tagged strain, with the estimated percent labeling of the donor protein (Rpl16-HTA), being around 10% (Figure A.13).

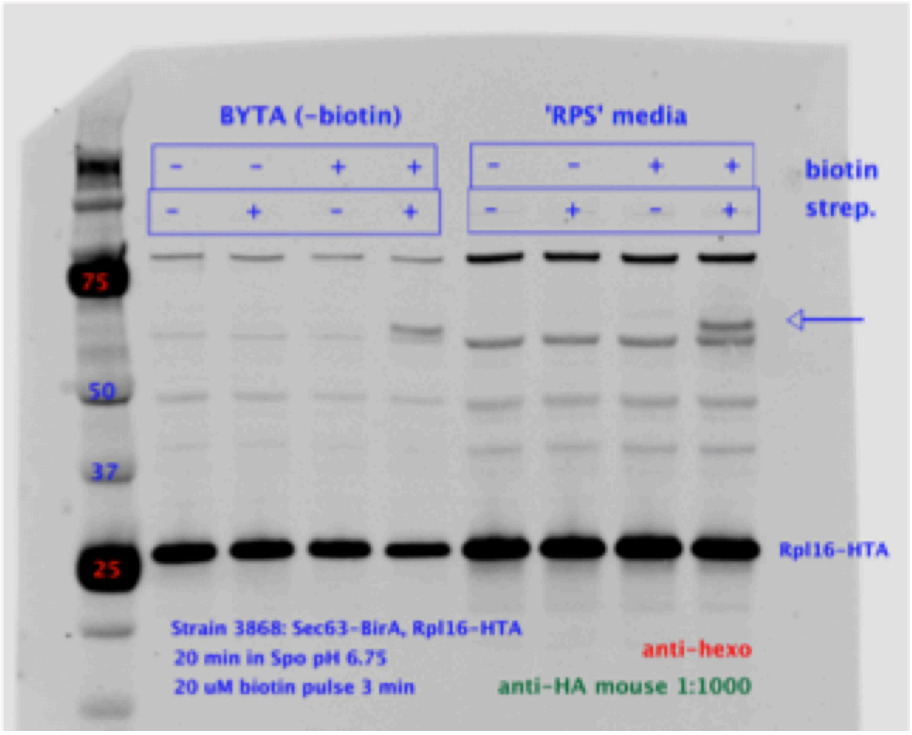

Figure A.13: Estimate of biotinylation of Rpl16-HTA by addition of streptavidin. The band indicated with an arrow is the biotinylated portion, which is about 10%.

Performing proximity ribosome profiling in meiotic cells has the potential to identify lots of interesting features of organelle targeting, and especially dual-targeting during meiosis. Further optimization of the conditions would need to be done to increase the sporulation efficiency as well as biotin labeling.

A.4 Conditions for growth of *upf1Δ* strains in meiosis

Strains lacking *UPF1* have variable growth defects, which can make them incapable of entering and undergoing meiosis. If the cells are able to grow well on non-fermentable media (YPG), then they will likely be able to undergo meiosis in sporulation media. After crossing two haploid *upf1Δ* strains, approximately 90% of individual diploids are unable to grow well on YPG. Therefore, at least ten diploids should be taken from the alpha-factor selection onto YPD 4% plates, and then each diploid should be patched to YPG to test growth. Only diploids that are able to grow well on YPG should be continued with. The basis for this high percentage of clones that do not grow on YPG is not understood, and could be followed up on.

A.5 Testing sporulation efficiency with different flask sizes and locations

There is often some variability in sporulation efficiency, between experiments and even between days. Different growth conditions had anecdotally given different effects on meiosis. To systematically test several aspects to determine if they had an effect, and if

so, which was best, two strains were tested using different sized flasks, with or without addition of raffinose, with or without rinsing the flask before adding cells (Figure A.14). Different shakers were used, including two large shakers on ground level and one smaller, sticky shaker on a benchtop. Multiple individuals counted the sporulation percentage, and averages between different combinations of people were used to compare spread of counting values.

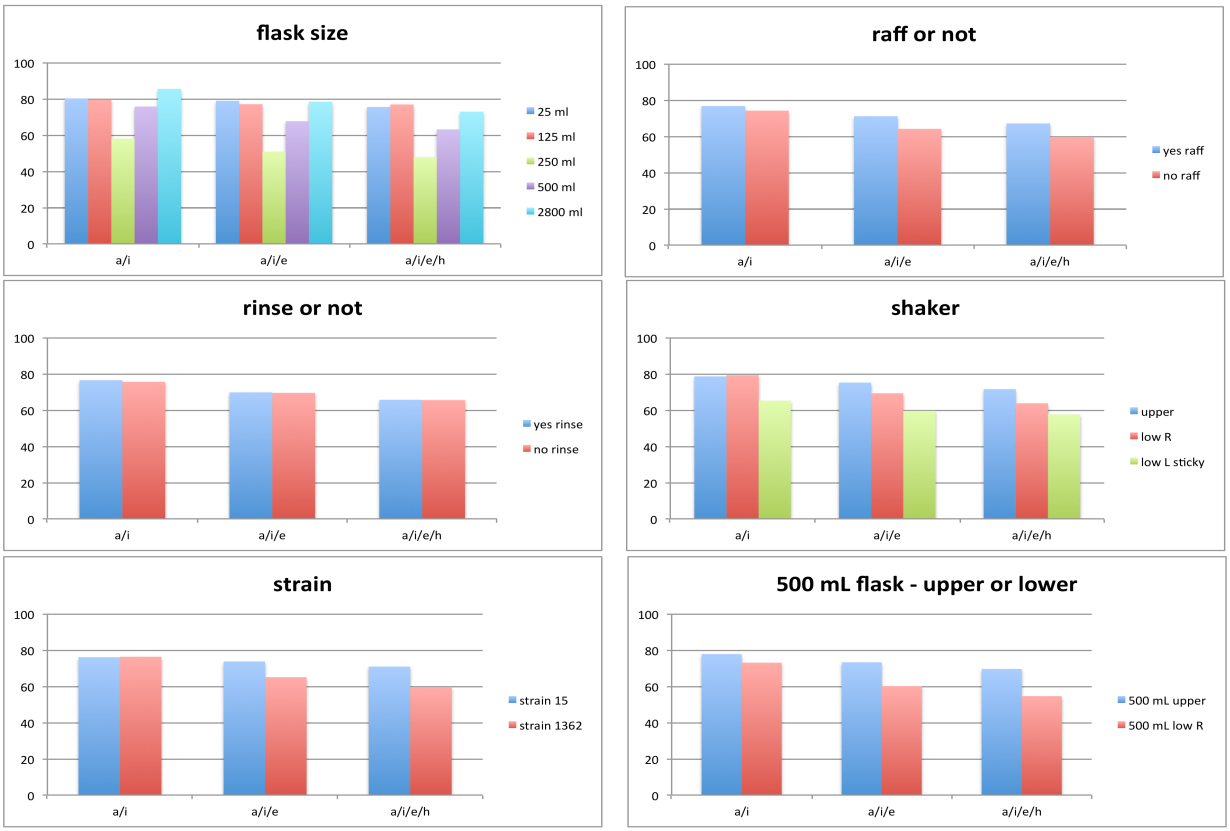

Figure A.14: Sporulation efficiency of cells grown in different conditions. Testing included different flask size, shaker location, raffinose addition and flask rinse. Counts were performed by multiple individuals, which are averaged in different combinations (a/I, a/i/e, a/i/e/h).

Flask size had an effect, with 250 mL flasks having the lowest sporulation efficiency. Adding raffinose helped slightly, and using the upper shaker was better than the lower shaker, which was better than the lower sticky shaker. Strain 15 was slightly more efficient than strain 1362. Overall, 500 mL flasks on the upper shaker yielded the best sporulation percentages, at just under 80%. This analysis demonstrated that there is variability in sporulation, and careful testing of different factors can help identify the combination that works best for a given strain and experimental setup. In the case of the LTM ribosome profiling, 500 mL flasks on the upper sticky shaker were determined to be best, and were used for this experiment as well as most others here.

A.6 Phalloidin staining of meiotic cells

For co-localization of actin, cells can be stained to look at phalloidin. Below is a protocol for staining meiotic cells, and representative images show the localization through several time points (Figure A.15). A detailed protocol can be found with lab protocols.

1. Fix 15-30 min (less time is better to preserve GFP signal, can either do EtOH or formaldehyde fixation)
2. Resuspend in KPO4-sorbitol + 1% Triton for 5 min
3. Wash once then resuspend in 100 uL in KPO4-sorbitol
4. Stain 1:200 rhodamine phalloidin in 30 uL, 1 h at RT
5. Wash 1x, add Dapi, image

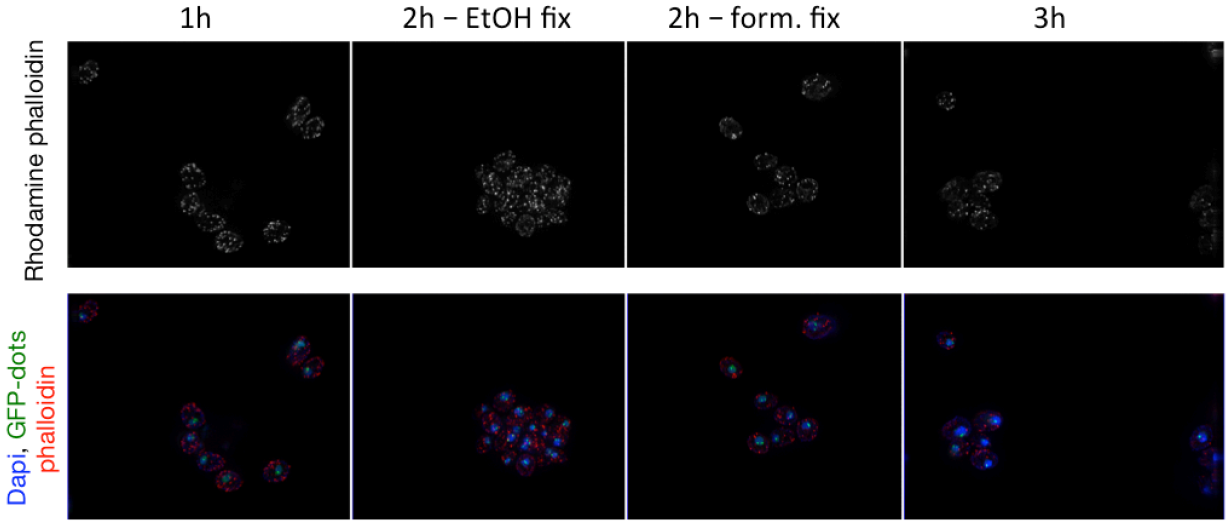

Figure A.15: Phalloidin staining of wild type meiotic cells at different time points in meiosis.