Title

Rules are made to be broken: a “simple” model organism reveals the complexity of gene regulation

Journal

Current Genetics, 67(1)

ISSN

0172-8083

Authors

Higdon, Andrea L
Brar, Gloria A

Publication Date

2021-02-01

DOI

10.1007/s00294-020-01121-8

Peer reviewed

Published in final edited form as:

Curr Genet. 2021 February ; 67(1): 49–56. doi:10.1007/s00294-020-01121-8.

Rules are made to be broken: a “simple” model organism reveals the complexity of gene regulation

Andrea L. Higdon[1,2], **Gloria A. Brar**[1,2]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[2]Center for Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA

Abstract

Global methods for assaying translation have greatly improved our understanding of the protein coding capacity of the genome. In particular, it is now possible to perform genome-wide and condition-specific identification of translation initiation sites through modified ribosome profiling methods that selectively capture initiating ribosomes. Here we discuss our recent study applying such an approach to meiotic and mitotic timepoints in the simple eukaryote, budding yeast, as an example of the surprising diversity of protein products—many of which are non-canonical—that can be revealed by such methods. We also highlight several key challenges to studying non-canonical protein isoforms that have precluded their prior systematic discovery. A growing body of work supports expanded use of empirical protein coding region identification, which can help relieve some of the limitations and biases inherent to traditional genome annotation approaches. Our study also argues for the adoption of less static views of gene identity and a broader framework for considering the translational capacity of the genome.

Introduction:

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

Decades of fruitful research on the functions and regulation of proteins has benefited from gene annotations, which integrate DNA sequence information with predictions about the regions of these sequences that are eligible for decoding into proteins. Initial gene

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

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

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

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

the genome than can be readily inferred from sequence analysis alone, even in the well-studied and simple budding yeast.

Translation initiation site profiling detects non-canonical protein isoforms

Our straightforward definition of what constitutes a coding region remained more or less intact for decades in the absence of tools to empirically and systematically put it to the test. Ribosome profiling, a method for capturing and sequencing ribosome-protected fragments of mRNA, started to change our understanding by providing a global picture of the positions and levels of translation (Ingolia et al. 2009; reviewed in Ingolia 2014, Brar and Weissman 2015). A modified version of ribosome profiling—in which cells are pretreated with a specific type of translation inhibitor, such as harringtonine or lactimidomycin, which block the first elongation cycle of the ribosome—strongly favors the capture of ribosomes that have just completed translation initiation (Ingolia et al. 2011, Lee et al. 2012, Schneider-Poetsch et al. 2010). Relative to traditional ribosome profiling, this allows cleaner detection of translation initiation sites, unobscured by signal from elongating ribosomes within ORFs. Its application to mammalian cells revealed complexity in TIS usage, but these data are challenging to interpret (Ingolia et al. 2011, Fields et al. 2015). Identifying the coding region based on the start codon depends on transcript isoform definitions, which are incomplete and likely to be highly conditionally regulated in mammals, based on the small subset that have been studied in great depth (reviewed in Baralle and Giudice 2017). In addition, when multiple transcript isoforms are present, it is difficult to unambiguously assign TIS peaks to a specific transcript isoform (Figure 1A).

We recently adapted this approach, which we call translation initiation site profiling (TIS-profiling), for budding yeast and applied it to samples that spanned a meiotic time course (Eisenberg et al. 2020). Meiosis is a highly regulated developmental program that, in yeast, converts a diploid cell into four haploid spores. It requires dynamic and precisely regulated waves of gene expression changes to achieve dramatic morphological changes to cellular components (reviewed in Marston and Amon 2005, vanWerven and Amon 2011). Standard ribosome profiling revealed large temporally regulated changes in the quantities of proteins made from nearly all annotated genes, and also hinted at qualitative changes in the identity of the proteins being produced, including evidence for translation in 5’ leaders (traditionally defined as UTRs or “untranslated regions”) of approximately half of mRNAs expressed in meiosis (Brar et al. 2012). The translation in many of these 5’ leaders could be attributed to upstream open reading frames (uORFs), and transcripts showing such translation often appeared to have translation of several overlapping uORFs. Based on this, and the ensemble nature of ribosome profiling data, which reveals all translated positions for the pool of a given transcript in the samples collected, it was difficult to unambiguously assign reads to specific ORFs for these transcripts. TIS-profiling, in contrast, enables comprehensive identification of ORF start codons by enriching for ribosome footprint signal representing post-initiation ribosomes.

Using this method, we observed widespread translation initiation for non-canonical regions, including uORFs and both in-frame and out-of-frame ORFs that were internal to annotated ORFs (Eisenberg et al. 2020). We also found cases of translation initiation in 5’ leaders that

are in-frame with annotated ORFs and have no intervening stop codon. These cases would be expected to produce N-terminally extended alternate protein isoforms, which we investigated in depth. In our data, we identified 149 genes with such extended isoforms, representing a small but notable fraction of the ~6000 annotated yeast genes. These isoforms are, as a class, more abundantly produced during meiosis relative to vegetative growth, fitting with a general trend of increased translation from canonical and non-canonical start codons in upstream regions during meiosis. Our system was particularly well-suited for identifying these extended protein isoforms because the sparse alternative splicing in budding yeast allowed us to make more unambiguous coding region predictions than is possible in organisms with prevalent alternative splicing (Figure 1A).

Challenges of studying N-terminally extended, near-cognate initiated ORFs

A handful of N-terminally extended protein isoforms have been identified and characterized by single-gene studies, but prior to genome-wide ribosome profiling studies they were generally considered to be anomalies (Chang and Wang, 2004, Heublein et al. 2019, Kearse and Wilusz 2017, Kritsiligkou et al. 2017, Suomi et al. 2014, Tang et al. 2004, Touriol et al. 2003). Our work, along with recent work from other groups, indicates that extended protein isoforms are much more prevalent than previously thought, in yeast and other organisms (Fields et al. 2015, Fritsch et al. 2012, Ingolia et al. 2011, Ivanov et al. 2011, Monteuuis et al. 2019, Sapkota et al. 2019). How did the majority escape detection for so long? A number of features make them particularly challenging to identify or detect, making it understandable that they could be missed, even for well-studied genes.

The primary reason that these proteins were overlooked is that they do not conform to the typical rules of gene annotation. Extended isoforms, which by definition start upstream of the furthest upstream in-frame AUG of an ORF, cannot possibly initiate at an AUG. Instead, they initiate at near-cognate start codons, which differ from AUG by one base (reviewed in Kearse and Wilusz 2017). Since near-cognate start codons are not included in traditional ORF annotations, these isoforms would never be predicted to exist under that framework. Near-cognate initiation is also less efficient than AUG initiation, and therefore the resultant extended protein isoforms are often low-abundance and difficult to detect relative to their corresponding canonical isoform (Chen et al. 2008, Clements et al. 1988, Kolitz et al. 2009).

Most of the extensions that we identified are also relatively close in size to the AUG-initiated isoform, making them difficult to distinguish by the method that would seem most straightforward, a western blot of the protein. The majority (~80%) of the extensions in our set were less than 10% larger than their corresponding annotated isoform, which is typically too small of a difference to confidently distinguish on a standard western blot, especially if the protein is at low abundance, as extensions often are (Figure 2A; Eisenberg et al. 2020). If a researcher did not have external evidence of an extension (such as from TIS-profiling), they would have no reason to suspect a second isoform was being produced. If, on the other hand, there is reason to look harder for an extended isoform, it is possible to isolate only the extended isoform by mutating the AUG start codon of the annotated isoform. However, we observed that in many cases, this in turn leads to nonsense-mediated decay (NMD) of the transcript produced from this mutated locus, significantly decreasing protein production

from that gene. These transcripts are likely subject to NMD because, in the absence of the AUG start codon for the annotated isoform, initiation at the upstream near-cognate codon is not sufficient to prevent initiation at subsequent out-of-frame AUG codons, signaling for recruitment of NMD factors and decay of the transcript (Figure 2B; Celik et al. 2017). For many genes, we were only able to detect the extended protein isoform in a *upf1Δ* background, which abrogates the NMD pathway (reviewed in Hug et al. 2016).

We have established that extended isoforms can easily be missed by standard molecular biology approaches and are understandably absent from gene annotations. Conservation analysis, however, is another strategy for coding region detection that has complementary strengths to other annotation approaches. Could it have been useful for detecting these isoforms prior to TIS-profiling? In the set of extensions that we identified, we attempted to find signatures of conservation within the 5'-most regions that are unique to the extensions relative to the annotated isoform (Eisenberg et al. 2020). To our surprise, even previously characterized extensions, such as for the tRNA synthetase *ALA1*, showed little evidence of conservation across yeast species (Tang et al. 2004). We suspect that, rather than suggesting a lack of functional relevance, this instead may indicate that the selective pressures acting on some extensions may be difficult to detect by sequence conservation analysis alone. In cases in which the properties of the amino acids (for example, charge or hydrophobicity) are more important than the specific sequence, we would expect to see conservation of the ability to make the extension, but not necessarily strong conversation of the sequence itself. Signal sequences are a particularly salient example, as they are often present at N-termini of proteins and can tolerate a large amount of sequence degeneracy while still maintaining the same function; indeed, most of the currently characterized extended protein isoforms have been shown to have altered protein localization (Chang and Wang 2004, Kaiser and Botstein 1990, Kritsiligkou et al. 2017, Suomi et al. 2014, Tang et al. 2004). Further characterization of extended protein isoforms will be necessary to understand the types of functions they serve and whether those functions are indeed less subject to sequence-level constraints. Regardless of the reason for their general lack of sequence conservation, however, this feature again illustrates the difficulty in detecting extended protein isoforms without empirical data.

While extremely useful for empirically identifying translation initiation sites, it is important to note that TIS-profiling data can include both false-positives and false-negatives. Some translation elongation inhibitors, such as cycloheximide, have been shown to cause pre-initiation complex stacking and artificially enhance near-cognate initiation (Kearse et al. 2019). This is not thought to occur with post-initiation inhibitors like lactimidomycin and harringtonine, but it remains important to be vigilant to artifacts that could be introduced through drug treatment (Kearse et al. 2019). For example, our study showed that while TIS peak heights are generally roughly quantitative to translation levels, near-cognate codons in particular appear to have artificially high peak heights relative to measured protein output, for reasons that are not fully understood (Eisenberg et al. 2020). These data are nonetheless incredibly useful in identifying likely alternative protein isoforms, enabling both genome-wide analysis of regulatory trends and making careful individual validation and characterization much more tractable.

What makes an ORF? Working towards a more inclusive definition

The extended protein isoforms identified in our study may have been missed previously because they are only produced in meiosis or simply because they use near-cognate start codons rather than the typical AUG start, revealing two prevalent biases in existing gene annotations: bias towards laboratory mitotic growth conditions and bias towards certain rules of translation that were defined by individual studies and then broadly generalized despite known exceptions. Our concept of what defines an open reading frame is rigid, albeit for good reason. Even with its compact genome, *S. cerevisiae* still has thousands of genes, many of which have not yet been characterized in detail (Wood et al. 2019). To prioritize regions for study, it is useful to use certain rules to predict protein coding regions, namely that they start with an AUG, end with a stop codon, and are of a length capable of producing a stable peptide (reviewed in Dinger et al. 2008). These guidelines have served us well for many years, but with development of technologies for global empirical identification of coding regions, it may be time to revisit these rules to create a more inclusive definition of what constitutes an ORF.

It has become increasingly clear that translation initiation at non-AUG start codons is a biologically relevant way of making protein isoforms (Kearse and Wilusz 2017). *In vitro* reporter studies have shown that near-cognate initiation, while an order of magnitude less efficient than that at AUGs, can still produce protein (Kolitz et al. 2009). Prior to our work, only a handful of functional extended isoforms had been characterized in single-gene studies, but these include cases with clear and important biological function. The tRNA synthetase gene *ALA1*, for example, uses an upstream ACG codon in addition to an AUG start codon to produce two isoforms that localize to the mitochondria and cytoplasm, respectively, and are necessary for translation in both locations (Tang et al. 2004). In our study, we observed this and numerous other examples of near-cognate initiation, and similar studies in other mammalian systems have also revealed widespread near-cognate initiation (Fields et al. 2015, Ingolia et al. 2011). Although the vast majority of these near-cognate-initiated isoforms remain functionally uncharacterized, their prevalence and usage across very evolutionarily diverged organisms suggests that near-cognate codons should be considered as possible ORF starts when annotating genes in the future.

Since examples of near-cognate initiation have been known for many years, should near-cognate codons have been included in annotations all along? Unfortunately, in the absence of empirical TIS usage data, it is simply not feasible to do so. A notable pitfall of expanding ORF definitions to include near-cognate start codons is that it creates a much more difficult computational prediction problem, by making the number of potential ORFs unrealistically large. In fact, our TIS-profiling data revealed that very few of the available in-frame near-cognate start codons in 5’ leaders are actually used to initiate translation, and the factors contributing to this specificity are still largely unknown (Figure 2C). Our study supports a role for eIF5A in modulating near-cognate usage, and other studies have suggested additional factors, like RNA structure, that may facilitate near-cognate initiation (Eisenberg et al. 2020, Guenther et al. 2018, Kozak 1990). Careful integration of these different types of data, as well as experiments aimed at unraveling the interplay between multiple trans and cis factors, will be important for fully understanding why some start codons are chosen over

others. Until this point, we will need to rely on empirical data to know which TISs are used. In turn, these data will likely inform our understanding and ability to predict TIS selection.

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

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

Our annotation of the genome and assignment of function to gene products draws heavily from studies of domesticated yeast strains under standard lab conditions. This skews our perception of “functional relevance” or “essentiality” towards nutrient-rich mitotic growth, which differs greatly from the conditions in which the wild ancestors of our domesticated lab strains evolved (reviewed in Liti 2015, Engel et al. 2014). While truly understanding the evolutionary trajectory and life history of yeasts will require a population genetic approach and study of wild yeast species ecology, we can still glean tremendous insight into the diversity of their gene regulatory mechanisms from studying domesticated yeasts under a broad array of conditions. By collecting TIS-profiling data across a meiotic time course, for example, we were not only able to see dynamic regulation patterns but also detected many protein isoforms that are not produced in vegetative growth conditions (for example *MDH1*, Figure 1B).

The true capacity of gene expression regulation cannot be detected within the confines of standard laboratory growth conditions, and in fact, many regulatory strategies that appear illogical or inefficient only begin to make sense in the light of environmental pressures. An example of a seemingly wasteful phenomenon, first discovered in the context of yeast meiosis, is Long Undecoded Transcript Isoform (LUTI) production, which accounts for many of the cases where mRNA and protein levels are decoupled during meiosis (Chen et al. 2017, Cheng et al. 2018). Here, two transcript isoforms are produced from the same locus: a shorter transcript that produces functional protein and a longer (and often abundant) LUTI, whose coding sequence translation is repressed by uORF translation in the extended 5’ leader. This LUTI appears to serve no function beyond the co-transcriptional repression of the shorter transcript conferred by LUTI production (Chia et al. 2017). Making an extra transcript rather than just turning the other off seems wasteful, but in the context of the

highly coordinated process of meiotic differentiation, this could provide a handy mechanism for simultaneously activating and inactivating sets of genes with the same transcription factor in a precisely temporally coordinated manner (reviewed in Tresenrider and Unal 2018, Otto and Brar 2018). In another seemingly wasteful phenomenon, during vegetative growth, many transcripts are produced but spliced inefficiently, their intron-contained transcripts degraded, as a way to downregulate genes that are specific to meiosis or response to environmental stresses. This strategy, however, may allow them to remain primed to upregulate production of the spliced transcripts as soon as the necessary cues are in place (Juneau et al. 2007, Pleiss et al. 2007).

Studying stress conditions challenges our assumptions on the “normal” regulation or function of a gene. In our own work, we find ourselves relying on phrases such as “main isoform” or “annotated isoform” to distinguish between the previously known and newly identified isoforms. However, in many cases, we find that the new isoform is in fact more robustly produced, perhaps at more time points or with more dynamic regulation than the annotated isoform, calling into question an easy binary categorization between a “main” and “alternative” isoform. Indeed, across biology, we frequently categorize the functions of a protein into their “main” and “moonlighting” roles, but “main” often just means the function that was discovered first, is most abundant during “normal” conditions, or has the most conventional regulation. Our increasingly nuanced understanding of transcript and protein isoform production suggests that it may be time to develop a less hierarchical naming system, and perhaps one that incorporates transcript and protein isoforms that serve a regulatory function rather than only a direct protein-template function.

Conclusions

By studying the repertoire of proteins produced across the developmental process of meiosis in budding yeast, we have seen cells bend canonical rules of translation to produce an astounding diversity of protein products, especially during times of stress and upheaval. The apparent simplicity of budding yeast makes it an especially useful organism for exploring conserved complexities, and its strengths can complement those of similar efforts in other organisms (Ingolia et al. 2011, Fields et al. 2015, Fritsch et al. 2012, Lee et al. 2012, Sapkota et al. 2019, Stern-Ginossar et al. 2012). Decades of research have built off of certain rules of gene regulation, and even the things produced within that framework are mind-bogglingly complex and beautiful. Looking forward, however, we know that improvements in technology can allow us to go beyond those rules to observe yet more levels of complexity and seek to understand them. As the “simple” budding yeast has shown us time and time again, rules are made to be broken.

References

Aitken CE, Lorsch JR, 2012 A mechanistic overview of translation initiation in eukaryotes. Nat Struct Mol Biol 19, 568–576. 10.1038/nsmb.2303 [PubMed: 22664984]

Baralle FE, Giudice J, 2017 Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol 18, 437–451. 10.1038/nrm.2017.27 [PubMed: 28488700]

Brar GA, Weissman JS, 2015 Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol 16, 651–664. 10.1038/nrm4069 [PubMed: 26465719]

Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS, 2012 High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. Science 335, 552–557. 10.1126/science.1215110 [PubMed: 22194413]

Brockdorff N, McCabe M, Norris P, Cooper J, Swift S, Kay F, n.d. The Product of the Mouse Xist Gene Is a 15 kb Inactive X-Specific Transcript Containing No Conserved ORF and Located in the Nucleus 12.

Celik A, Baker R, He F, Jacobson A, 2017 High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA 23, 735–748. 10.1261/rna.060541.116 [PubMed: 28209632]

Chang K-J, Wang C-C, 2004 Translation Initiation from a Naturally Occurring Non-AUG Codon in Saccharomyces cerevisiae. J. Biol. Chem 279, 13778–13785. 10.1074/jbc.M311269200 [PubMed: 14734560]

Chen J, Tresenrider A, Chia M, McSwiggen DT, Spedale G, Jorgensen V, Liao H, van Werven FJ, Onal E, 2017 Kinetochore inactivation by expression of a repressive mRNA. eLife 6, e27417 10.7554/eLife.27417 [PubMed: 28906249]

Chen S-J, Lin G, Chang K-J, Yeh L-S, Wang C-C, 2008 Translational Efficiency of a Non-AUG Initiation Codon Is Significantly Affected by Its Sequence Context in Yeast. J. Biol. Chem 283, 3173–3180. 10.1074/jbc.M706968200 [PubMed: 18065417]

Cheng Z, Otto GM, Powers EN, Keskin A, Mertins P, Carr SA, Jovanovic M, Brar GA, 2018 Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell 172, 910–923.e16. 10.1016/j.cell.2018.01.035 [PubMed: 29474919]

Chia M, Tresenrider A, Chen J, Spedale G, Jorgensen V, Onal E, van Werven FJ, 2017 Transcription of a 5' extended mRNA isoform directs dynamic chromatin changes and interference of a downstream promoter. eLife 6, e27420 10.7554/eLife.27420 [PubMed: 28906248]

Clements JM, Laz TM, Sherman F, 1988 Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. Mol. Cell. Biol 8, 4533–4536. 10.1128/MCB.8.10.4533 [PubMed: 3141793]

Dinger ME, Pang KC, Mercer TR, Mattick JS, 2008 Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. PLoS Comput Biol 4, e1000176 10.1371/journal.pcbi.1000176 [PubMed: 19043537]

Eisenberg AR, Higdon AL, Hollerer I, Fields AP, Jungreis I, Diamond PD, Kellis M, Jovanovic M, Brar GA, 2020 Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-lnitiated Protein Isoforms in Yeast. Cell Systems S2405471220302404 10.1016/j.cels.2020.06.011

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, Weng S, Wong ED, Lloyd P, Skrzypek MS, Miyasato SR, Simison M, Cherry JM, 2014 The Reference Genome Sequence of Saccharomyces cerevisiae : Then and Now. G3 4, 389–398. 10.1534/g3.113.008995 [PubMed: 24374639]

Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Hacohen N, Carr SA, Ingolia NT, Regev A, Weissman JS, 2015 A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Molecular Cell 60, 816–827. 10.1016/j.molcel.2015.11.013 [PubMed: 26638175]

Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, Schreiber S, Platzer M, Krawczak M, Hampe J, Brosch M, 2012 Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Research 22, 2208–2218. 10.1101/gr.139568.112 [PubMed: 22879431]

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG, 1996 Life with 6000 Genes. Science 274, 546–567. 10.1126/science.274.5287.546 [PubMed: 8849441]

Guenther U-P, Weinberg DE, Zubradt MM, Tedeschi FA, Stawicki BN, Zagore LL, Brar GA, Licatalosi DD, Bartel DP, Weissman JS, Jankowsky E, 2018 The helicase Dedip controls use of near-cognate translation initiation codons in 5' UTRs. Nature 559, 130–134. 10.1038/S41586-018-0258-0 [PubMed: 29950728]

Heublein M, Ndi M, Vazquez-Calvo C, Vogtle F-N, Ott M, 2019 Alternative Translation Initiation at a UUG Codon Gives Rise to Two Functional Variants of the Mitochondrial Protein Kgd4. Journal of Molecular Biology 431, 1460–1467. 10.1016/j.jmb.2019.02.023 [PubMed: 30822412]

Hug N, Longman D, Cáceres JF, 2016 Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res 44, 1483–1495. 10.1093/nar/qkw010 [PubMed: 26773057]

Ingolia NT, 2014 Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15, 205–213. 10.1038/nrg3645 [PubMed: 24468696]

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS, 2009 Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324, 218–223. 10.1126/science.1168978 [PubMed: 19213877]

Ingolia NT, Lareau LF, Weissman JS, 2011 Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell 147, 789–802. 10.1016/j.cell.2011.10.002 [PubMed: 22056041]

Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV, 2011 Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Research 39, 4220–4234. 10.1093/nar/qkr007 [PubMed: 21266472]

Juneau K, Palm C, Miranda M, Davis RW, 2007 High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proceedings of the National Academy of Sciences 104, 1522–1527. 10.1073/pnas.06103541Q4

Kaiser CA, Botstein D, 1990 Efficiency and diversity of protein localization by random signal sequences. Mol. Cell. Biol 10, 3163–3173. 10.1128/MCB.10.6.3163 [PubMed: 2160595]

Kearse MG, Goldman DH, Choi J, Nwaezeapu C, Liang D, Green KM, Goldstrohm AC, Todd PK, Green R, Wilusz JE, 2019 Ribosome queuing enables non-AUG translation to be resistant to multiple protein synthesis inhibitors. Genes Dev. 33, 871–885. 10.1101/qad.324715.119 [PubMed: 31171704]

Kearse MG, Wilusz JE, 2017 Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 31, 1717–1731. 10.1101/gad.305250.117 [PubMed: 28982758]

Kolitz SE, Takacs JE, Lorsch JR, 2008 Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. RNA 15, 138–152. 10.1261/rna.1318509 [PubMed: 19029312]

Kozak M, 1990 Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proceedings of the National Academy of Sciences 87, 8301–8305. 10.1073/pnas.87.21.8301

Kritsiligkou P, Chatzi A, Charalampous G, Mironov A, Grant CM, Tokatlidis K, 2017 Unconventional Targeting of a Thiol Peroxidase to the Mitochondrial Intermembrane Space Facilitates Oxidative Protein Folding. Cell Reports 18, 2729–2741. 10.1016/j.celrep.2017.02.053 [PubMed: 28297675]

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. Proceedings of the National Academy of Sciences 109, E2424–E2432. 10.1073/pnas.1207846109

Liti G, 2015 The fascinating and secret wild life of the budding yeast S. cerevisiae. eLife 4, e05835 10.7554/eLife.05835

Marston AL, Amon A, 2004 Meiosis: cell-cycle controls shuffle and deal. Nat Rev Mol Cell Biol 5, 983–997. 10.1038/nrm1526 [PubMed: 15573136]

Monteuuis G, Miścicka A, Świrski M, Zenad L, Niemitalo O, Wrobel L, Alarm J, Chacinska A, Kastaniotis AJ, Kufel J, 2019 Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. Nucleic Acids Research 47, 5777–5791. 10.1093/nar/gkz301 [PubMed: 31216041]

Morris DR, Geballe AP, 2000 Upstream Open Reading Frames as Regulators of mRNA Translation. Mol. Cell. Biol 20, 8635–8642. 10.1128/MCB.20.23.8635-8642.200Q [PubMed: 11073965]

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M, 2008 The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 320, 1344–1349. 10.1126/science.1158441 [PubMed: 18451266]

Otto GM, Brar GA, 2018 Seq-ing answers: uncovering the unexpected in global gene regulation. Curr Genet 64, 1183–1188. 10.1007/sQ0294-018-0839-3 [PubMed: 29675618]

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C, 2007 Transcript Specificity in Yeast Pre-mRNA Splicing Revealed by Mutations in Core Spliceosomal Components. PLoS Biol 5, e90 10.1371/journal.pbio.005009Q [PubMed: 17388687]

Renz PF, Valdivia-Francia F, Sendoel A, 2020 Some like it translated: small ORFs in the 5'UTR. Experimental Cell Research 396, 112229 10.1016/j.vexcr.2020.112229 [PubMed: 32818479]

Sapkota D, Lake AM, Yang W, Yang C, Wesseling H, Guise A, Uncu C, Dalai JS, Kraft AW, Lee J-M, Sands MS, Steen JA, Dougherty JD, 2019 Cell-Type-Specific Profiling of Alternative Translation Identifies Regulated Protein Isoform Variation in the Mouse Brain. Cell Reports 26, 594–607.e7. 10.1016/j.celrep.2018.12.077 [PubMed: 30650354]

Schmitt AM, Chang HY, 2017 Long Noncoding RNAs: At the Intersection of Cancer and Chromatin Biology. Cold Spring Harb Perspect Med 7, a026492 10.1101/cshperspect.a026492 [PubMed: 28193769]

Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, Green R, Shen B, Liu JO, 2010 Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat Chem Biol 6, 209–217. 10.1038/nchembio.304 [PubMed: 20118940]

Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, Mann M, Ingolia NT, Weissman JS, 2012 Decoding Human Cytomegalovirus. Science 338, 1088–1093. 10.1126/science.1227919 [PubMed: 23180859]

Suomi F, Menger KE, Monteuuis G, Naumann U, Kursu VAS, Shvetsova A, Kastaniotis AJ, 2014 Expression and Evolution of the Non-Canonically Translated Yeast Mitochondrial Acetyl-CoA Carboxylase Hfalp. PLoS ONE 9, e114738 10.1371/journal.pone.0114738 [PubMed: 25503745]

Suzuki K, Hashimoto T, Otaka E, n.d. Yeast ribosomal proteins: XI. Molecular analysis of two genes encoding YL41, an extremely small and basic ribosomal protein, from Saecharomyces cerevisiae 6.

Tang H-L, Yeh L-S, Chen N-K, Ripmaster T, Schimmel P, Wang C-C, 2004 Translation of a Yeast Mitochondrial tRNA Synthetase Initiated at Redundant non-AUG Codons. J. Biol. Chem 279, 49656–49663. 10.1074/jbc.M40808120Q [PubMed: 15358761]

Touriol C, Bornes S, Bonnal S, Audigier S, Prats H, Prats A-C, Vagner S, 2003 Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biology of the Cell 95, 169–178. 10.1016/S0248-4900(03100033-9 [PubMed: 12867081]

Tresenrider A, Ünal E, 2018 One-two punch mechanism of gene repression: a fresh perspective on gene regulation. Curr Genet 64, 581–588. 10.1007/s00294-017-0793-5 [PubMed: 29218463]

van Werven FJ, Amon A, 2011 Regulation of entry into gametogenesis. Phil. Trans. R. Soc. B 366, 3521–3531. 10.1098/rstb.2011.0081 [PubMed: 22084379]

Wang Z, Gerstein M, Snyder M, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. 10.1038/nrg2484 [PubMed: 19015660]

Wood V, Lock A, Harris MA, Rutherford K, Bahler J, Oliver SG, n.d. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? 8.

Yu X, Warner JR, 2001 Expression of a Micro-protein. J. Biol. Chem 276, 33821–33825. 10.1074/jbc.M103772200 [PubMed: 11451953]

Zhang H, Wang Y, Lu J, 2019 Function and Evolution of Upstream ORFs in Eukaryotes. Trends in Biochemical Sciences 44, 782–794. 10.1016/j.tibs.2019.03.002 [PubMed: 31003826]

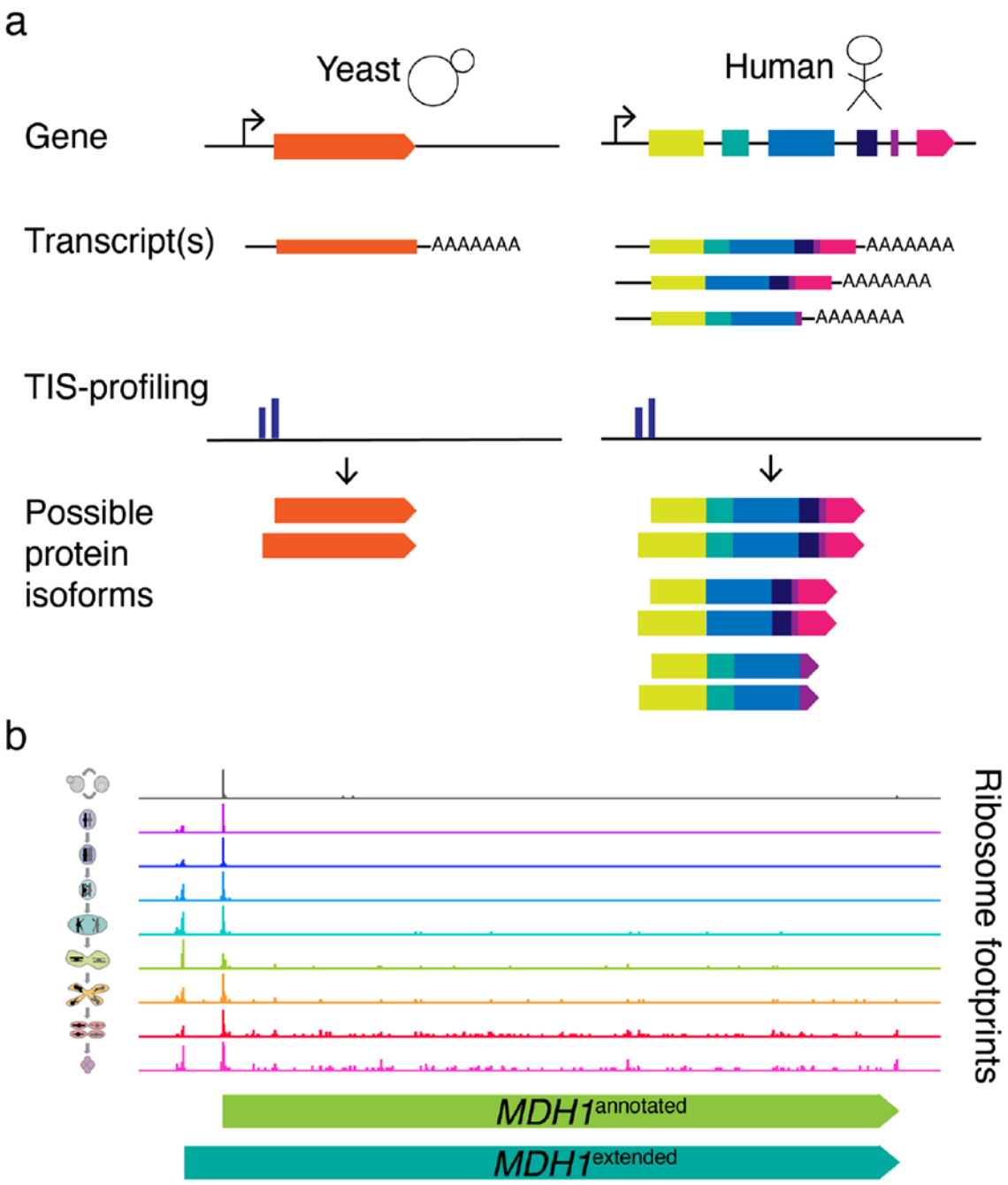

Figure 1.

(a) Comparison of protein isoform prediction from TIS-profiling data in yeast (left panel) and humans (right panel). Alternative transcript isoforms present in humans contribute to ambiguity in protein isoform identity. (b) Example of TIS-profiling data for an N-terminally extended protein isoform at the *MDH1* locus. Timepoints from top to bottom, illustrated by the cartoon on the left: vegetative exponential, 0, 1.5, 3, 4.5, 6, 8, 10, and 22 hours after addition to sporulation media.

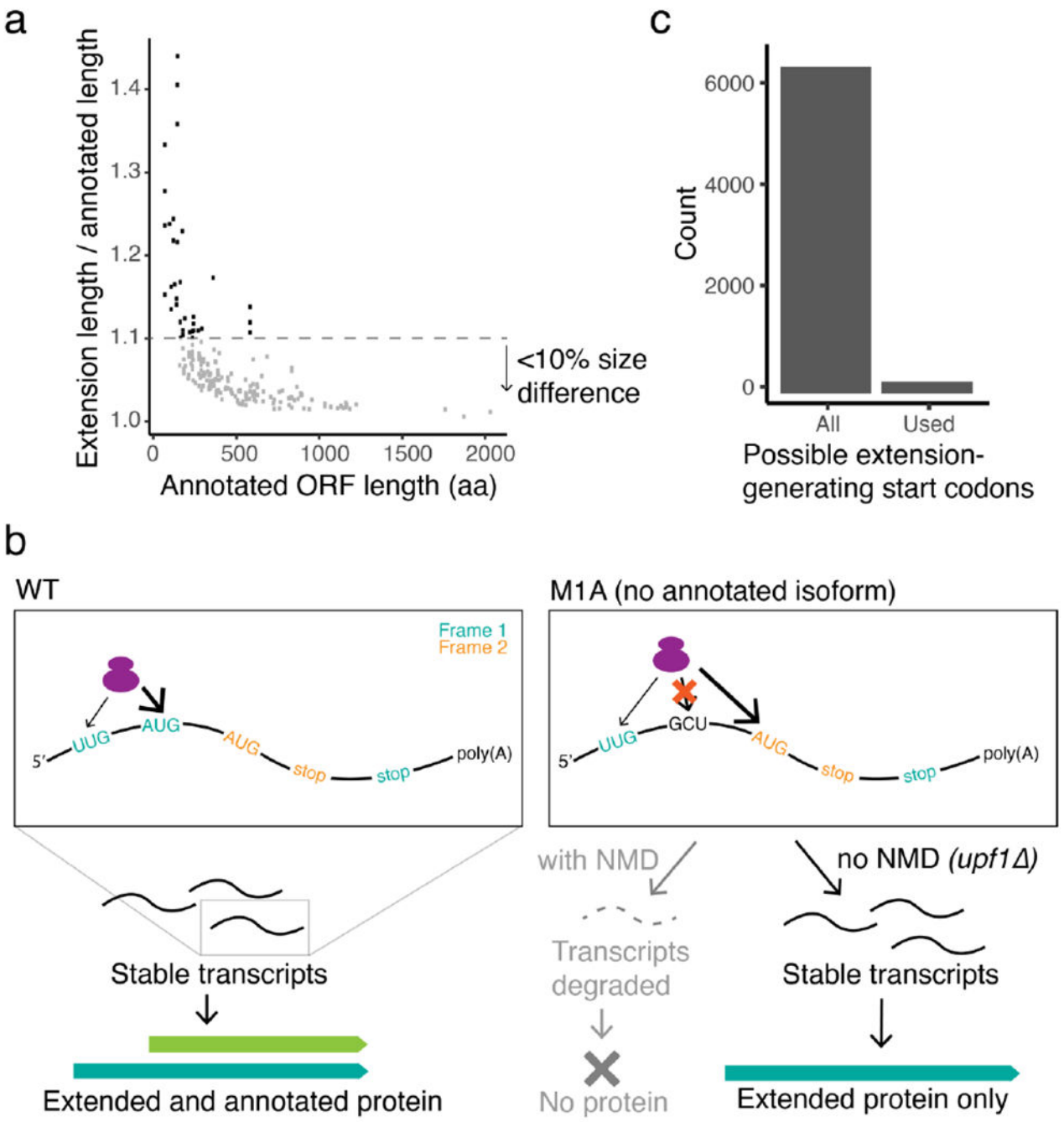

Figure 2.

(a) Length of annotated protein isoform compared to ratio of extended isoform to annotated isoform. Gray dots represent genes with an extension differing by less that 10% in length from the annotated isoform. (b) Schematic of NMD effects resulting from AUG start codon mutation. WT transcript in WT background (left panel) results in stable transcripts that produce both protein isoforms. M1A (mutation of annotated start codon to alanine) transcript in WT background (right panel, left arrow) results in translation from out-of-frame AUG-initiated short ORFs and triggers NMD. M1A transcript in *upf1Δ* background cannot

be degraded via NMD and can be translated to produce the extended isoform alone. (c) Bar chart of all possible extension-generating (in-frame with no in-frame stop codon) near-cognate start codons in 5’ leaders of yeast genes compared to extension-generating start codons with evidence of translation in TIS-profiling data.