Rare detection of noncanonical proteins in yeast mass spectrometry studies

Aaron Wacholder[12] and Anne-Ruxandra Carvunis[123]

1. Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
2. Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
3. Corresponding author

*Correspondence: **anc201@pitt.edu**

Abstract

Ribosome profiling experiments indicate pervasive translation of short open reading frames (ORFs) outside of annotated protein-coding genes. However, shotgun mass spectrometry experiments typically detect only a small fraction of the predicted protein products of this noncanonical translation. The rarity of detection could indicate that most predicted noncanonical proteins are rapidly degraded and not present in the cell; alternatively, it could reflect technical limitations. Here we leveraged recent advances in ribosome profiling and mass spectrometry to investigate the factors limiting detection of noncanonical proteins in yeast. We show that the low detection rate of noncanonical ORF products can be explained by small size and low translation levels and does not indicate that they are unstable or biologically insignificant. In particular, no proteins encoded by evolutionarily young genes were detected, not even those with well-characterized biological roles. Additionally, we find that decoy biases can give misleading estimates of noncanonical protein false discovery rates, potentially leading to false detections. After accounting for these issues, we found strong evidence for four noncanonical proteins in mass spectrometry data, which were also supported by evolution and translation data. These results illustrate the power of mass spectrometry to validate unannotated genes predicted by ribosome profiling, but also its substantial limitations in finding many biologically relevant lowly-expressed proteins.

Introduction

Ribosome profiling (ribo-seq) experiments indicate that genomes are pervasively translated outside of annotated coding sequences.[1] This “noncanonical” translatome primarily consists of small open reading frames (ORFs), located on the UTRs of annotated protein-coding genes or on separate transcripts, that potentially encode thousands of small proteins missing from protein databases.[2] Several previously unannotated translated ORFs identified by ribo-seq have been shown to encode microproteins that play important cellular roles.[3–6] The number of translated noncanonical ORFs identified by ribo-seq analyses is typically very large, but many are weakly expressed, poorly conserved[7–9], and not reproduced between studies[10], suggesting that they may not all encode functional proteins. There has thus been considerable interest in proteomic detection of the predicted products of noncanonical ORFs.[11–15] Detection of a noncanonical ORF product by mass spectrometry (MS) confirms that the ORF can generate a stable protein that is present in the cell at detectable concentrations and thus might be a good candidate for future characterization.

Over the past decade, numerous studies have attempted to identify noncanonical proteins using bottom-up “shotgun” proteomics in which MS/MS spectra from a digested protein sample are matched to predicted spectra from a protein database.[16,17] These studies report hundreds of peptides encoded by noncanonical ORFs with evidence of detection in mass spectrometry data.[13–15,18–20] However, these detections typically represent only a small fraction of the noncanonical ORFs found to be translated using ribo-seq. It is unclear whether most proteins translated from noncanonical ORFs are undetected by MS because they are absent from the cell, for example owing to rapid degradation, or because they are technically difficult to detect. Both the short sequence length and low abundance of noncanonical ORFs pose major challenges for detection in typical bottom-up MS analysis.[17] Alternative techniques for protein detection, such as microscopy[21] and targeted proteomics[22], are more sensitive at detecting small proteins, but lack the convenience of untargeted bottom-up MS in being able to readily search for unannotated proteins predicted from an entire genome, transcriptome or translatome of a species.

Several recent MS studies have aimed to improve detection of short, lowly-expressed proteins in *S. cerevisiae*. He et al. 2018[23] used a combination of techniques to enrich for small proteins and detected 117 microproteins, including three translated from unannotated ORFs. Gao et al. 2021[24] also used a combination of strategies to detect many small and low abundance proteins. Sun et al. 2022[25] searched for unannotated microproteins in a variety of stress conditions and found 70, all expressed from alternative reading frames of canonical coding sequences. At the same time as these studies provided increased coverage of the yeast proteome, Wacholder et al. 2023[7] integrated ribo-seq data from hundreds of experiments in over 40 published studies and assembled a high-confidence yeast reference translatome including 5372 canonical protein-coding genes and over 18,000 noncanonical ORFs. Here we leveraged these recent technical advances in MS and ribo-seq analysis to investigate the factors limiting detection of noncanonical proteins using *S. cerevisiae* as a model organism.

Results

Noncanonical peptides and decoys detected at comparable rates

Using the MSFragger program[26], we searched the three aforementioned published MS datasets optimized for detection of short, lowly expressed proteins[23–25] against a sequence dataset that included all 5,968 canonical yeast proteins on Saccharomyces Genome Database (SGD)[27] as well as predicted proteins from 18,947 noncanonical ORFs (including both unannotated ORFs and ORFs annotated as "dubious") inferred to be translated in Wacholder et al. 2023[7] on the basis of ribosome profiling data. The spectra from the three studies were pooled and false discovery rates (FDR) were estimated separately for canonical and noncanonical ORFs using a target-decoy approach.[28] MSFragger expect scores were used to assess confidence in peptide-spectrum matches (PSMs), with lower values indicating stronger matches. Among canonical ORFs, 4021 of 5968 had proteins detected at a 1% FDR (**Figure 1A**). For noncanonical ORFs, it was not possible to generate a substantial list of detected proteins at a 1% FDR because too many decoys were detected relative to targets at all confidence thresholds (**Figure 1B**).

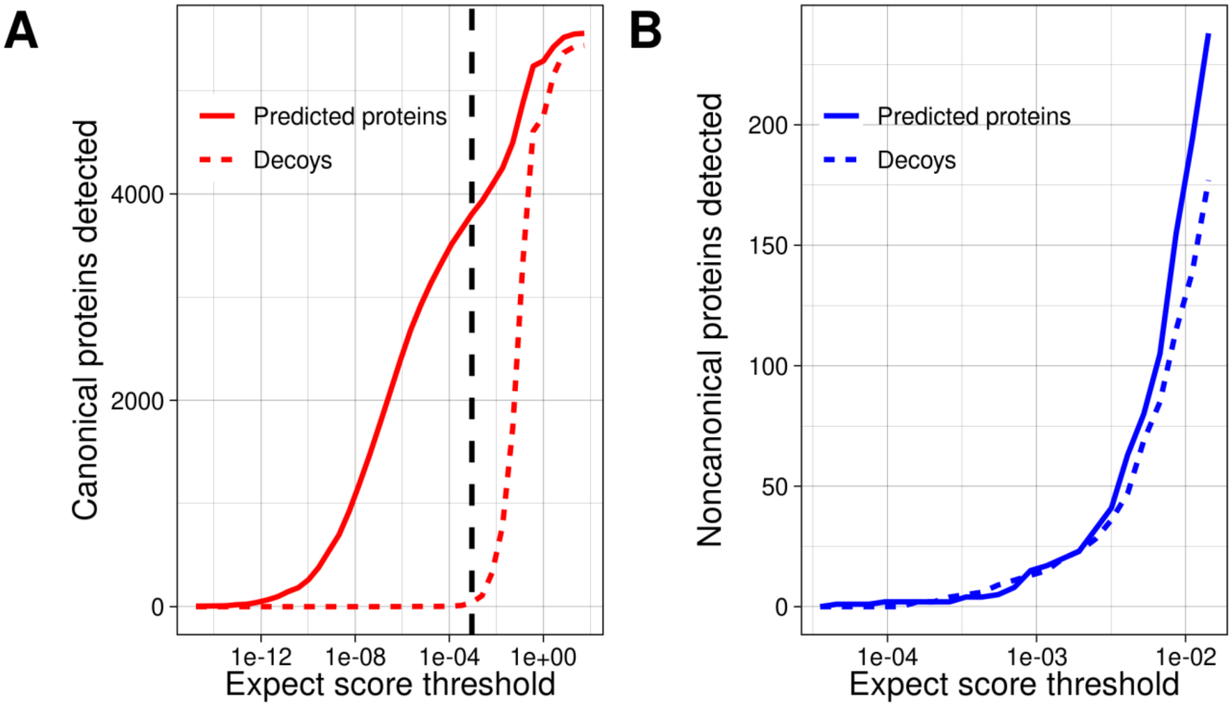


Figure 1: Few noncanonical proteins are confidently detected in MS data. A) The number of predicted proteins and decoys detected in MS data at a range of confidence thresholds among canonical yeast proteins. The dashed line signifies the 1% FDR threshold. B) The number of predicted noncanonical proteins and decoys detected in MS data at a range of confidence thresholds.

Decoy bias among noncanonical ORF products leads to inaccurate FDR estimates

In general, there is a trade-off in target-decoy approaches such that setting a weaker confidence threshold results in a longer list of proteins inferred as detected, but with a higher FDR. In the case of yeast noncanonical ORF peptides, the decoy/target ratio never went below 60% for any list of inferred detected target proteins larger than 10, and this ratio also did not converge to 1 even with thresholds set to allow 10,000 target proteins to pass (**Figure 2A**). The small enrichment of targets above decoys gives little confidence in detection of noncanonical ORF products at the level of individual proteins but leaves open the possibility that MS data could contain a weak biological signal.

However, there is an alternative explanation for why targets are found at somewhat higher rates than decoys across a large range of confidence thresholds: decoy bias.[28] The accuracy of FDR calculations require that target and decoy false positives are equally likely at any threshold, but this assumption could be violated if there are systematic differences between targets and decoys. Decoy bias has been assessed in previous work by comparing the number of target and decoy PSMs below the top rank for each spectra: if a peptide is genuinely detected, it will usually be the best match to its spectra, and so lower-ranked matched peptides will be false and should appear at approximately equal numbers for both targets and decoys.[28] Among canonical ORFs, this expected pattern is observed (**Figure 2B**). In contrast, targets substantially outnumber decoys at all ranks for noncanonical ORFs (**Figure 2C**). We reasoned that this bias could be explained by the short length of noncanonical proteins. Indeed, many predicted peptides derived from noncanonical ORFs include the starting methionine, while decoys, consisting of reversed sequences from the protein database, are more likely to end with methionine

(**Figure 2D**). To eliminate this large systematic difference, we constructed an alternative decoy database in which decoys for noncanonical proteins were reversed only after the leading methionine. When this database is used, the number of noncanonical targets and decoys at each rank is close to equal (**Figure 2E**) and the target/decoy ratio converges to one as confidence thresholds are lowered (**Figure 2F**). This behavior is consistent with expectations for a well-constructed decoy set. We therefore repeated our initial analysis using the alternative decoy set (**Figure 2G-H**) and used it for all subsequent analyses.

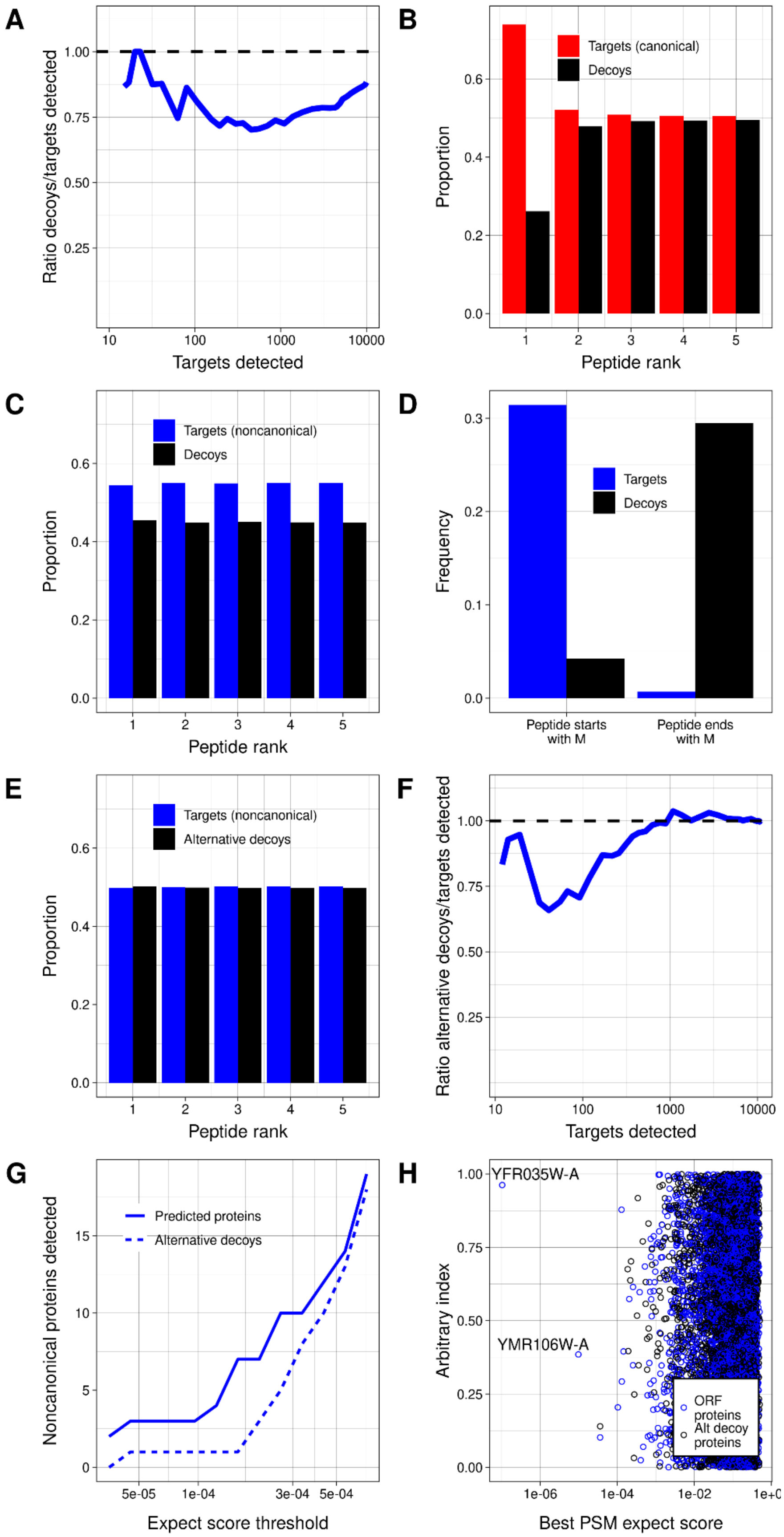

A
Ratio decoys/targets detected
1.00
0.75
0.50
0.25
10
100
1000
10000
Targets detected
B
Targets (canonical)
Decoys
Proportion
0.6
0.4
0.2
0.0
1
2
3
4
5
Peptide rank
C
Targets (noncanonical)
Decoys
Proportion
0.6
0.4
0.2
0.0
1
2
3
4
5
Peptide rank
D
Targets
Decoys
Frequency
0.3
0.2
0.1
0.0
Peptide starts with M
Peptide ends with M
E
Targets (noncanonical)
Alternative decoys
Proportion
0.6
0.4
0.2
0.0
1
2
3
4
5
Peptide rank
F
Ratio alternative decoys/targets detected
1.00
0.75
0.50
0.25
10
100
1000
10000
Targets detected
G
Predicted proteins
Alternative decoys
Noncanonical proteins detected
15
10
5
0
5e-05
1e-04
3e-04
5e-04
Expect score threshold
H
YFR035W-A
YMR106W-A
ORF proteins
Alt decoy proteins
Arbitrary index
1.00
0.75
0.50
0.25
0.00
1e-06
1e-04
1e-02
1e+0
Best PSM expect score

Figure 2: Decoy biases distort false discovery rate estimation. A) Among noncanonical proteins, the ratio of decoys detected to targets detected, across a range of targets detected, which varies with expect score threshold. Decoys are reverse sequences of the noncanonical protein database. B) Across all spectra, the proportion of peptide-spectrum matches of each rank that are canonical peptides vs. decoys. Peptide rank indicates the rank of the strength of the peptide-spectrum match, ordered across all peptides and decoys. C) Across all spectra, the proportion of peptide-spectrum matches of each rank that are noncanonical peptides vs. decoys. D) Among noncanonical ORF and decoy predicted trypsinized peptides that match spectra at any confidence level, the proportion that start or end with a methionine. E) Across all spectra, the proportion of peptide-spectrum matches of each rank that are noncanonical peptides vs. decoys, using the alternative decoy set. Alternative decoys are constructed by reversing noncanonical proteins after the starting methionine, such that all decoy and noncanonical proteins start with M. F) Among noncanonical proteins, the ratio of decoys detected to targets detected across counts of targets detected, using the alternative decoy set. G) The number of predicted proteins and decoys at a range of confidence thresholds, using the alternative decoy set. H) The best peptide-spectrum match expect scores for each noncanonical protein and decoy in the database, using the alternative decoy set.

Two noncanonical proteins show strong evidence of genuine detection

Using the alternative decoy set and standard MSFragger analysis, we remained unable to construct an FDR-controlled list of noncanonical proteins at a 10% FDR threshold because decoys were still detected at a similar rate as targets (**Figure 2G**). We therefore sought to examine the strongest hits to determine if we could identify evidence that any were genuine detections. Two noncanonical proteins had peptides with stronger expect scores than any decoys (**Figure 2H**; standard MSFragger approach in **Table 1**). We gave the ORFs encoding these proteins systematic names YMR106W-A and YFR035W-A following SGD conventions. Both proteins matched to two distinct spectra at thresholds stronger than the best decoy match. Moreover, YMR106W and YFR035W-A both had ribo-seq read counts greater than 99.9% of noncanonical ORFs in the Wacholder et al. dataset. The identification of multiple matching spectra for these noncanonical proteins and their relatively high rates of translation provide strong support that these are genuine detections.

Table 1: Noncanonical ORFs possibly detected in mass spectrometry data

Systematic name	Approaches used to find	Coordinates	Peptides detected (spectra count)	Best expect score	Quantile of ribo-seq read count	Evidence of conservation	Strength of evidence**
YMR106W-A*	Standard MSFragger, MS-GF+	chrXIII:480924-481187	MISMEAINNFIK (1), ISMEAINNFIK (1)	9.82e-06	0.99958	None	Strong
YFR035W-A*	Standard MSFragger, MS-FG+	chrVI:226260-226550	HLNIPDLRFEK (2)	1.04e-07	0.99974	Conserved within genus	Strong
YPR159C-A	Acetylation	chrXVI: 857598-857660	IVACTICVQVCATKVVR (1)	8.48e-06	0.858	None	Weak
YIL059CW-A*	Non-enzymatic end	chrIX:246550-246915	EFDFDVGYEEFVR (1)	4.74e-07	0.987	Conserved with *S. jurei*	Strong
YNL155C-A*	Same-strand overlap	chrXIV: 341911-342135	KQHTEWPIEENR (2), MIGLIVVPILFAIK (8)	1.06e-08	0.99968	Conserved within genus	Strong

*Assigned in this study.

**Assessed based on proteomic, translation and evolutionary evidence

YMR106W-A is located 27 nt away from a Ty1 long terminal repeat. No homologs outside *S. cerevisiae* were found using BLASTP or TBLASTN against the NCBI non-redundant and nucleotide databases or against the 332 budding yeast genomes collected by Shen et al. 2018.[29] It is thus plausible that this ORF was brought into the *S. cerevisiae* genome through horizontal transfer mediated by Ty1 retrotransposition.[30] YFR035W-A overlaps the canonical ORF YFR035C on the opposite strand. However, YFR035C was not detected in our canonical protein MS analysis. YFR035C deletion was reported to increase sensitivity to alpha-synuclein[31], but this observation stemmed from a full ORF deletion that would also have disturbed YFR035W-A. While YFR035C has 287 in-frame ribo-seq reads mapping to the ORF in the Wacholder et al. 2023[7] dataset, YFR035W-A has 22,523, greater by a factor of 79 (**Figure 3A**). In a multiple sequence alignment with other species in the Saccharomyces genus, the full span of the YFR035W-A amino acid sequence aligns between all species (**Figure 3B)**, while other species have an early stop preventing alignment with most of the YFR035C amino acid sequence (**Figure 3C**). Thus, evolutionary, translation and proteomics evidence all indicate that unannotated ORF YFR035W-A is a better candidate for a conserved protein-coding gene than annotated ORF YFR035C.

Alternative strategies for MS search yield two additional noncanonical peptide detections

Aside from YMR106W-A and YFR035W-A, the standard MSFragger approach did not confidently detect proteins encoded by noncanonical ORFs supported by ribo-seq. We therefore considered some reasons we could miss noncanonical proteins present in the data and employed alternative approaches to test these possibilities. For each approach, we determined whether a substantial list of noncanonical ORFs could be constructed with FDR of 10%. If not, we further investigated peptides with expect scores < 10^{-5}, similar to the level at which YMR106W-A was detected.

First, we hypothesized that a mismatch between the environmental conditions in which the ribo-seq and MS datasets were constructed may explain the low number of detected noncanonical proteins. To investigate this possibility, we reduced our analysis to consider only ribo-seq and MS experiments conducted on cells grown in YPD at 30° C. The target/decoy ratio looked similar to the analysis on the full dataset, with no peptide list generatable with a 10% FDR (**Figure 4A**). The only noncanonical proteins detected at a 10^{-5} expect score threshold were the same two as in the standard analysis.

Next, to ensure our results were not specific to the search program MSFragger, we repeated our analysis using MS-GF+.[32] The pattern of target vs. decoy detection was again similar to the standard MSFragger analysis, with no peptide list generatable with a 10% FDR (**Figure 4B**). The only noncanonical proteins detected at a 10^{-5} e-value threshold were YMR106W-A and YFR035W-A, also found by MSFragger. We then applied the machine learning based MS^{2}Rescore algorithm[33] to rescore the MSGF+ results, as this has been shown to improve peptide identification rates in some contexts. However, this also did not improve target-decoy ratios (**Figure 4C**) and the strongest rescored match was to a decoy.

Next, we hypothesized that noncanonical proteins could have been missed from our searches due to post-translational modification or cleavage. Allowing for phosphorylation of threonine, serine, or tyrosine as variable modifications did not improve the decoy/target ratio or yield detection of any noncanonical phosphorylated peptides at a 10^{-5} expect score threshold (**Figure 4D**). Adding acetylation of lysine or N-terminal acetylation as variable modifications did not improve target/decoy ratios overall **(Figure 4E**), but a single hit with an expect score of 8.48 x 10^{-6} was found, which we named YPR159C-A. The corresponding peptide was encoded from an ORF on the opposite strand of the canonical gene YPR159W. However, this hypothetical protein was identified from a peptide found only once, showed no evidence of conservation in the *Saccharomyces* genus, and was translated at lower levels than other noncanonical protein detections (**Table 1**); we therefore conclude that it may not be a genuine detection.

Allowing for peptides to have one end that is not an enzymatic cut site to search for potential cleavage products did not improve target/decoy ratios overall (**Figure 4F**), but a single additional noncanonical peptide was identified with a relatively strong expect score of 4.7 x 10^{-7}. This peptide was from the ORF YIL059C, annotated as “dubious” on SGD, indicating that, in the view of SGD, the ORF is “unlikely to encode a functional protein.” YIL059C is in the 98th percentile of ribo-seq read count and 99th percentile of length among noncanonical ORFs, at 366 nt (**Table 1**). It overlaps on the opposite strand the ORF YIL060W, classified as “verified” on SGD. However, the references listed in support of YIL060W are all based on full deletion experiments which would disturb both ORFs and therefore do not distinguish between them.[34–36] YIL060W may have been considered the more likely gene as its ORF is longer, at 435 nt. But as in the case of YFR035C and YFR035W-A discussed above, both ribo-seq and MS data provide more support for the noncanonical ORF than the canonical ORF on the opposite strand: YIL059C has 1741 ribo-seq reads compared to only 7 reads for YIL060W (**Figure 5A**), and YIL060W was not detected in our MS analysis of canonical ORFs. Given that the YIL059C peptide had one non-enzymatic end, we tested whether it could be a signal peptide using the TargetP program.[37] YIL059C has a predicted signal peptide cleavage site corresponding exactly to the detected peptide (**Figure 5B**), providing additional support that this is a genuine detection. Searching for homologs using TBLASTN, BLASTP and BLASTN in the NCBI databases and in *Saccharomyces* genus genomes at a 10^{-4} e-value threshold, YIL059C and YIL060W have detected DNA homologs only in *Saccharomyces* species *S. paradoxus*, *S. mikatae* and *S. jurei*. There was an intact protein alignment of YIL059C between *S. cerevisiae* and *S jurei* (**Figure 5C**) while YIL060W has no homologs that fully align in any species (**Figure 5D**). YJL059C is located adjacent, and on the opposite strand, to a Ty2 long terminal repeat. These observations are consistent with a transposon-mediated horizontal transfer of YIL059C prior to divergence between *S. cerevisiae* and *S. mikatae*, followed by loss in *S. paradoxus* and *S. mikatae* and preservation in *S. cerevisiae* and *S. jurei*. We do not rule out a role for YIL060W, but all considered evidence provides greater support for the biological significance of YIL059C.

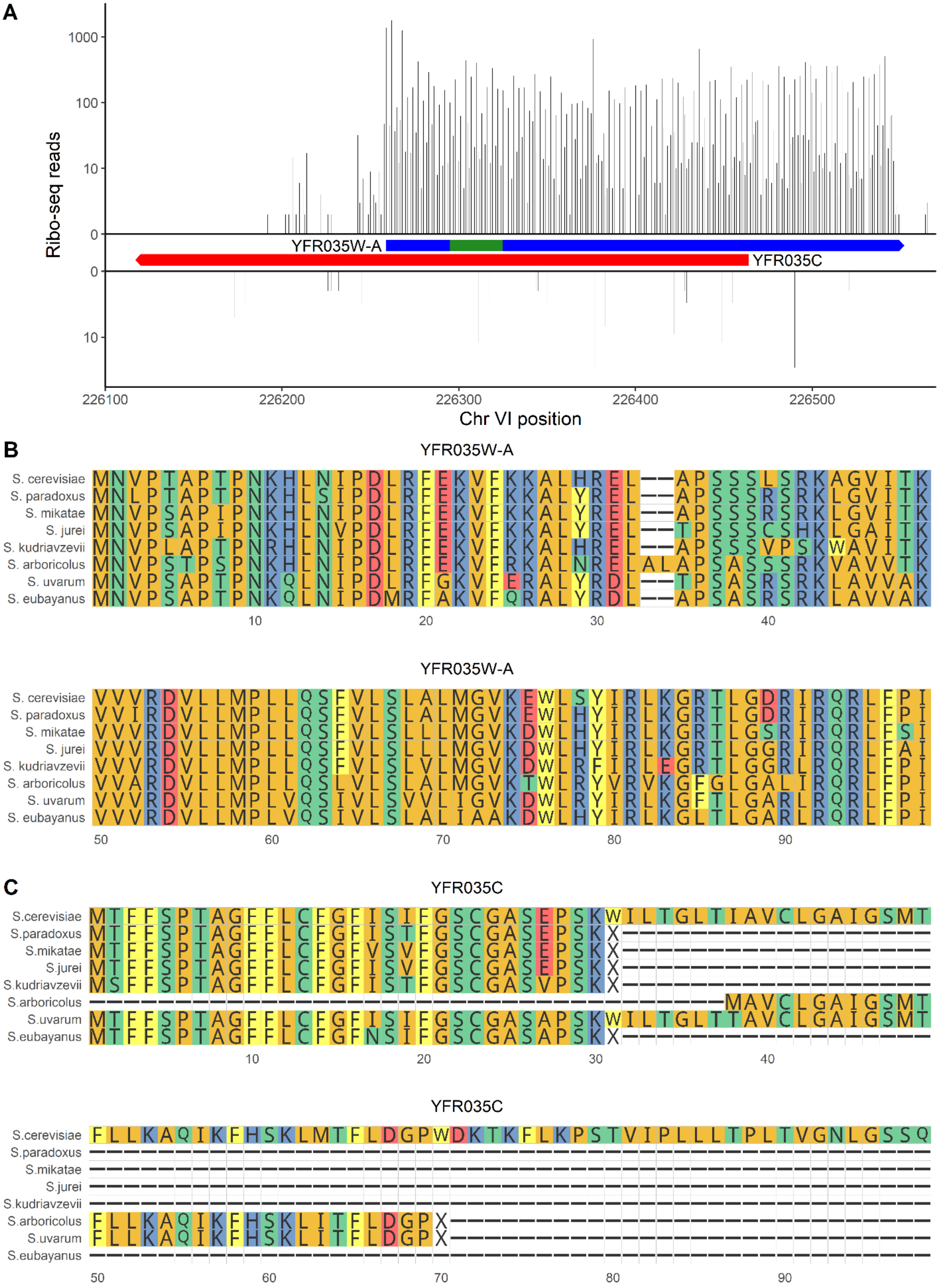

A
Ribo-seq reads
1000
100
10
0
0
10
YFR035W-A
YFR035C
226100
226200
226300
226400
226500
Chr VI position
B
YFR035W-A
S. cerevisiae MNVPTAPTPNKHLNIPDLRFEKVFKKALHREL--APSSSLSRKAGVITK
S. paradoxus MNLPTAPTPNKHLSIPDLRFEKVFKKALYREL--APSSSRSRKLGVITK
S. mikatae MNVPSAPIPNKHLNIPDLRFEKVFKKALYREL--APSSSRSRKLGVITK
S. jurei MNVPSAPIPNKHLNVPDLRFEKVFKKALYREL--TPSSSCSHKLGAITK
S. kudriavzevii MNVPLAPTPNRHLNIPDLRFEKVFKKALHREL--APSSSVPSKWAVITK
S. arboricolus MNVPSTPSPNKHLNIPDLRFEKVFRKALNRELALAPSASSSRKVAVVTK
S. uvarum MNVPSAPTPNKQLNIPDLRFGKVFERALYRDL--TPSASRSRKLAVVAK
S. eubayanus MNVPSAPTPNKQLNIPDMRFAKVFQRALYRDL--APSASRSRKLAVVAK
10
20
30
40
YFR035W-A
S. cerevisiae VVVRDVLLMPLLQSFVLSLALMGVKEWLSYIRLKGRTLGDRIRQRLFPI
S. paradoxus VVIRDVLLMPLLQSFVLSLALMGVKEWLHYIRLKGRTLGDRIRQRLFPI
S. mikatae VVVRDVLLMPLLQSFVLSLLLMGVKDWLHYIRLKGRTLGSRIRQRLFSI
S. jurei VVVRDVLLMPLLQSFVLSLLLMGVKDWLHYIRLKGRTLGGRIRQRLFAI
S. kudriavzevii VVVRDVLLMPLLQSFVLSLVLMGVKDWLRFIRLEGRTLGGRLRQRLFPI
S. arboricolus VVARDVLLMPLLQSLVLSLALMGVKTWLRYVRVKGFGLGALIRQRLFPI
S. uvarum VVVRDVLLMPLVQSIVLSVVLIGVKDWLRYIRLKGFTLGARLRQRLFPI
S. eubayanus VVVRDVLLMPLVQSIVLSLALIAAKDWLHYIRLKGLTLGARLRQRLFPI
50
60
70
80
90
C
YFR035C
S.cerevisiae MTFFSPTAGFFLCFGFISIFGSCGASEPSKWILTGLTIAVCLGAIGSMT
S.paradoxus MTFFSPTAGFFLCFGFISTFGSCGASEPSKX------------------
S.mikatae MTFFSPTAGFFLCFGFVSVFGSCGASEPSKX------------------
S.jurei MTFFSPTAGFFLCFGFISVFGSCGASEPSKX------------------
S.kudriavzevii MSFFSPTAGFFLCFGFISTFGSCGASVPSKX------------------
S.arboricolus ------------------------------------MAVCLGAIGSMT
S.uvarum MTFFSPTAGFFLCFGFISIFGSCGASAPSKWILTGLTTAVCLGAIGSMT
S.eubayanus MTFFSPTAGFFLCFGFNSIFGSCGASAPSKX------------------
10
20
30
40
YFR035C
S.cerevisiae FLLKAQIKFHSKLMTFLDGPWDKTKFLKPSTVIPLLLTPLTVGNLGSSQ
S.paradoxus ---
S.mikatae ---
S.jurei ---
S.kudriavzevii ---
S.arboricolus FLLKAQIKFHSKLITFLDGPX----------------------------
S.uvarum FLLKAQIKFHSKLITFLDGPX----------------------------
S.eubayanus ---
50
60
70
80
90

Figure 3: Translation and evolutionary evidence indicates that unannotated ORF YFR035W-A is likely a conserved gene. A) ribo-seq reads on unannotated ORF YFR035W-A (top) and annotated ORF YFR035C (bottom). The bounds of each ORF are indicated in boxes. The location of the detected peptide is indicated in green. B) Alignment of the amino acid sequence of YFR035W-A with its homologs across the *Saccharomyces* genus. C) Amino acid alignment of the annotated ORF YFR035C and its homologs in *Saccharomyces*.

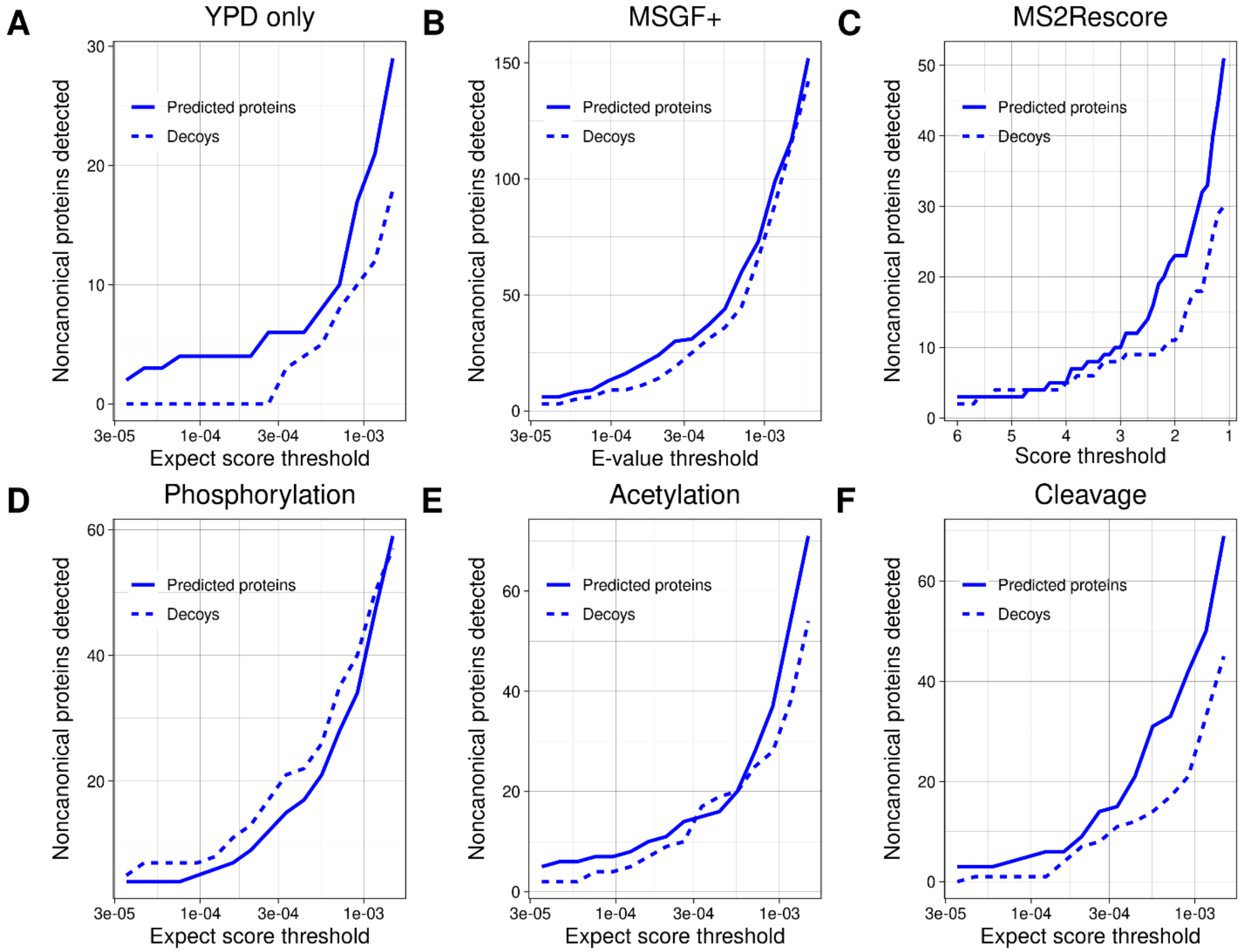


Figure 4: Alternative strategies for detecting noncanonical ORF products yield few additional discoveries. A-F) The number of predicted proteins and decoys detected across a range of thresholds, using a variety of strategies for detection. Aside from the specific changes indicated, all searches were run using the same parameter settings (described in Methods). A) Analysis using only ribo-seq and MS data taken from yeast grown in YPD at 30° C. B) Analysis using the program MSGF+. C) Analysis using the rescoring algorithm MS^{2}Rescore on MSGF+ results. D) Analysis allowing for phosphorylation of threonine, serine or tyrosine as variable modifications. E) Analysis allowing for acetylation of lysine or n-terminal acetylation as variable modifications. F) Analysis allowing detection of peptides with one end as a non-enzymatic cut site.

Finally, we wanted to investigate a class of noncanonical ORFs not present in the Wacholder et al. translated ORF dataset: noncanonical ORFs that overlap a canonical ORF on the same strand. These ORFs are difficult to identify by ribo-seq because it is challenging to distinguish noncanonical ORF-associated ribo-seq reads from those of the canonical gene; however, some proteins encoded by noncanonical ORFs that overlap canonical ORFs have been identified in previous MS analyses, including in the Sun et al. dataset included in our MS analysis.[25] We therefore constructed a sequence database consisting of all canonical ORFs as well as noncanonical ORFs that overlap canonical ORFs on the same strand, with ORFs determined only from the genome sequence rather than expression evidence.

Running this database against the full set of MS data, we again observed that, among noncanonical ORFs, decoys were detected at a high fraction of the rate of predicted peptides and so a list of confident noncanonical detections could not be established at reasonable false discovery rates (**Figure 6A).** Only one overlapping ORF had associated PSMs with expect scores stronger than 10^{-5}. We assigned it systematic name YNL155C-A following SGD conventions (**Table 1**).

The stable translation product of YNL155C-A was supported by two distinct peptides which together were detected 10 times with expect scores below the best decoy score of 5.12×10^{-7}, with the strongest value of 1.06×10^{-8}. This 255 bp ORF overlaps canonical gene YNL156C for 57 of 255 bases. Its translation product was not identified in the Sun et al. analysis.[25] A clear pattern of ribo-seq read triplet periodicity was observed in the frame of YNL155C-A (i.e., reads tend to match to the first position of a codon) before the overlap with YNL156C, indicating translation in this frame (**Figure 6B**). There also appears to be a triplet periodic pattern in a frame distinct from both YNL156C and YNL155C-A at the locus, suggesting that all three frames may be translated. Excluding the overlapping region, there are 14,741 reads on the ORF that map to the first position of a codon in the YNL155C-A reading frame; this would put it in the 99.95th percentile of read count among translated noncanonical ORFs in the Wacholder et al. dataset. No homologs were found in more distantly related species in a TBLASTN search against the NCBI non-redundant protein database, but YNL155C-A was well conserved across *Saccharomyces* (**Figure 6C**). Thus, proteomic, translation and evolutionary evidence all support YNL155C-A as a protein-coding gene.

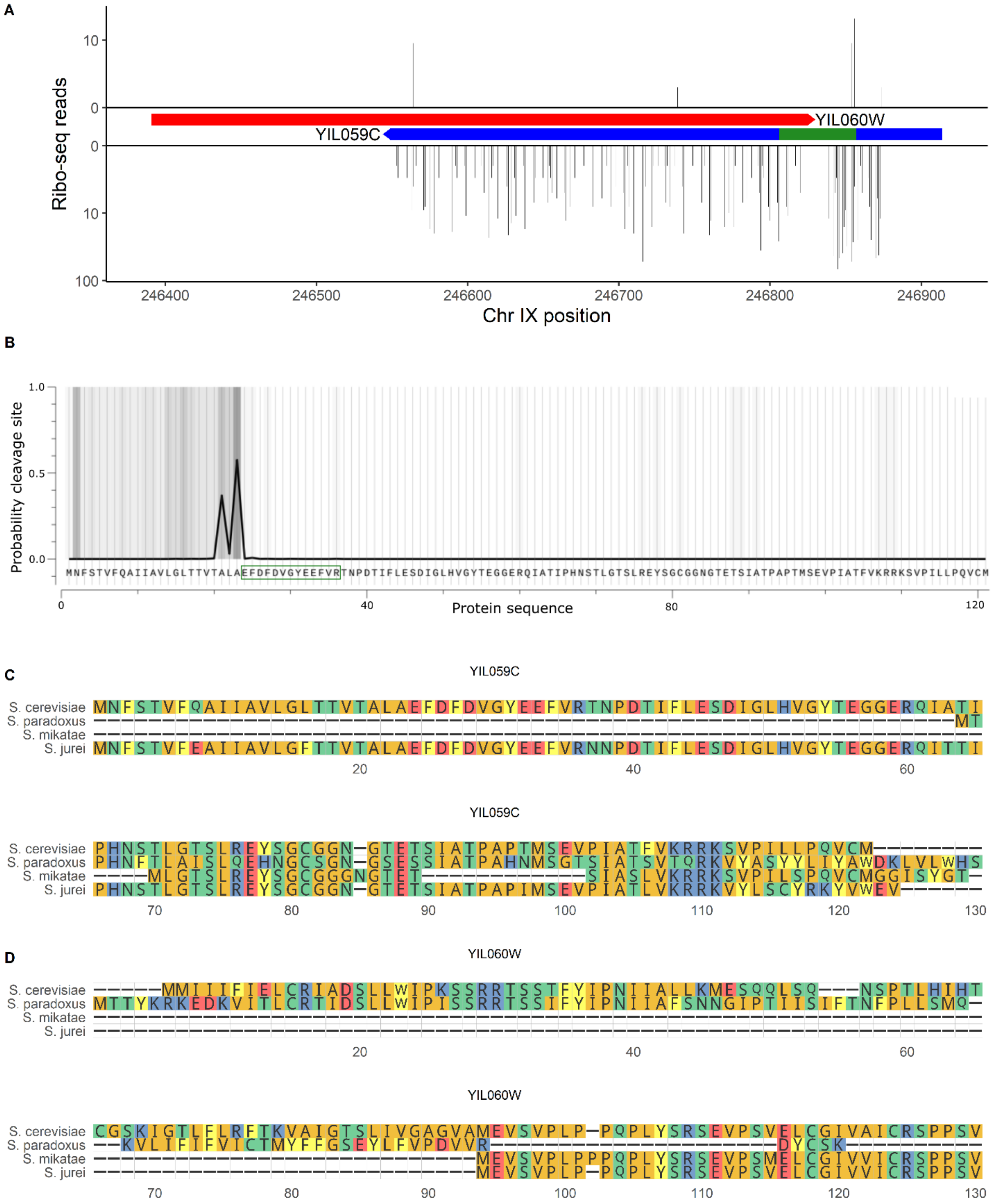


Figure 5: Dubious ORF YIL059C encodes a signal peptide. A) Ribo-seq reads on canonical ORF YIL060W (top) and "dubious" ORF YIL059C (bottom). The bounds of each ORF are indicated in boxes. The location of the detected peptide is indicated in green. B) Probability of a signal peptide cleavage site across the YIL059C sequence, as predicted by TargetP.[37] The peptide detected in MS analysis is indicated by a green box. C) Alignment of YIL059C with the highest identity protein matches at the homologous locus in *Saccharomyces* species. Only species with a homologous locus (at the DNA level) are shown. D) Alignment of YIL060W, the

canonical gene antisense to YIL059C, with its highest identity protein matches at the homologous locus in *Saccharomyces* species.

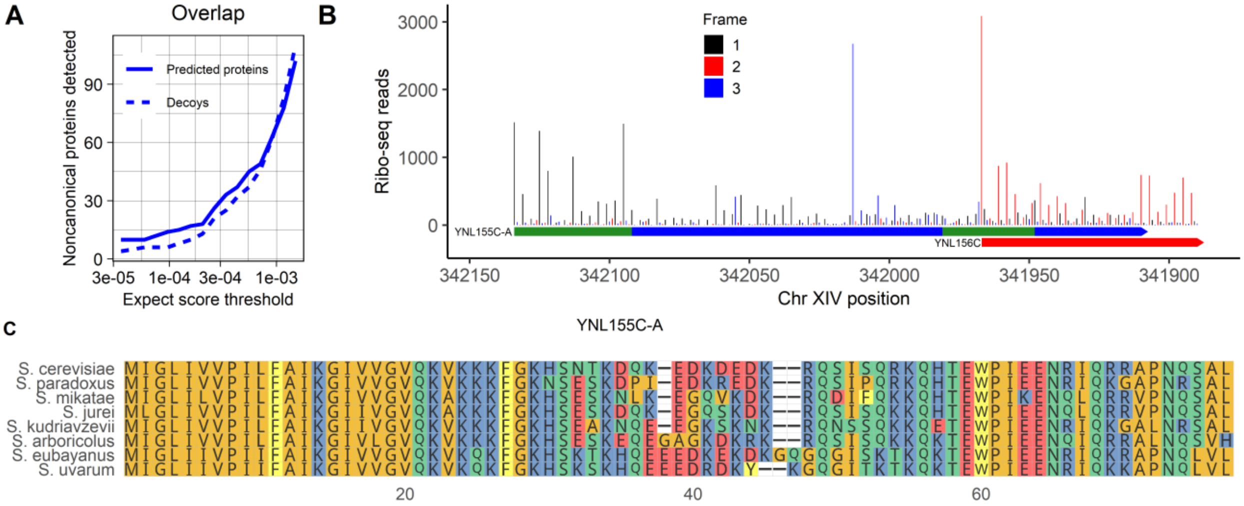


Figure 6: Noncanonical protein YNL155C-A, detected by MS, is well-translated and conserved in *Saccharomyces* genus. A) Predicted proteins and decoys detected in MS data at a range of expect-score thresholds, among noncanonical proteins that could be encoded by ORFs that overlap canonical ORFs on alternative frames. B) Ribo-seq reads across the YNL155C-A ORF. Reads are assigned to the reading frame in which the position they map to is the first position in a codon. The full span of YNL155C-A and the start of YNL156C are shown. The position of the two peptides found in MS are in green. C) Multiple sequence alignment of YNL155C-A with its homologs in the *Saccharomyces* genus.

The low detectability of noncanonical proteins can be explained by their short lengths and low translation rates

We sought to understand why the large majority of peptides predicted from translated noncanonical ORFs remained undetected across multiple computational search strategies. A major difference between canonical and noncanonical proteins is length: the average canonical protein is 503 residues compared to only 31 among noncanonical proteins. Short size can affect protein detection probability through distinct mechanisms: the sample preparation steps of the MS experiment may be biased against small proteins[17], and shorter sequences also provide fewer distinct peptides when digested. To distinguish these mechanisms, we related detection probability to ORF length at the level of peptides rather than proteins. We computationally constructed all possible enzymatic peptide sequences that could be theoretically detected from the proteins in the sequence database given their length and mass. We then calculated the peptide detection rate, out of all theoretically detectable peptides, among different ORF size classes (**Figure 7A**). We observe a division between canonical ORFs shorter vs. longer than 150 nt. Among 27 canonical yeast ORFs shorter than 150 nt, none of the 269 theoretically detectable peptides were detected at a 10^{-6} expect score threshold (a high-confidence detection threshold). This detection rate is significantly below expectation given the overall 5.5% rate at which canonical peptides are detected (binomial test, p = 5.5 x 10^{-7}), suggesting that there may be technical biases limiting detection of proteins that are this short. As 83% of noncanonical ORFs (15,717) are shorter than 150 nt, short length can partially explain the low detectability of noncanonical ORF products. In contrast, however, among canonical ORFs longer than 150 nt, shorter lengths were associated with higher probabilities that a peptide was detected. This is likely due to a trend of higher translation rates among shorter ORFs (**Supplementary Figure 1A**), which is also observed among noncanonical ORFs (**Supplementary Figure 1B**). This observation suggests that short size should not be a barrier to detection of proteins encoded by noncanonical ORFs longer than 150 nt. There are 3,080 such ORFs, potentially encoding 32,728 detectable peptides, yet only one was found at a 10^{-6} expect score threshold (the peptide from YFR035W-A, Table 1).

Besides length, a major difference between canonical and noncanonical ORFs is expression level, and this too can affect the probability a protein is detected in MS data.[17] We therefore evaluated the relation between translation level and detection probability using the ribo-seq data from Wacholder et al. The number of in-frame ribo-seq reads that map to a canonical ORF is strongly associated with the probability of detecting the ORF product at a 10^{-6} expect score threshold, at both the protein (**Figure 7B**) and peptide (**Figure 7C**) level. As with protein length, we can use the canonical ORFs to infer an approximate detection limit: among 267 canonical ORFs with fewer than 1000 in-frame mapped reads, only 2 of 8388 theoretically detectable peptides were detected at a 10^{-6} threshold. Thus, almost all canonical peptides, with only these two exceptions, are found among ORFs with at least 1000 reads and longer than 150 nt. Yet, only 80 noncanonical ORFs (0.4% of total) are in this category (**Figure 7D**). Thus, almost all noncanonical ORFs are outside the limits in which canonical ORF products are detected by MS.

For the 80 noncanonical translated ORFs displaying length and expression levels amenable to detection (longer than 150 nt and detected with more than 1000 ribo-seq reads), we estimated the probability a peptide would be detected at a 10^{-6} expect score threshold under the assumption that detection probability depends only on read count. This probability was estimated as the peptide detection rate among canonical ORFs with a similar read count to the transient ORF (a natural log of read count within 0.5). Given these estimates, the expected total count of detected peptides for the 80 ORFs was 2.68. In reality, a single peptide was detected (the peptide from YFR035W-A, Table 1). To see whether observing only a single detection was surprising, we simulated the distribution of peptide detection counts under the estimated detection probabilities (**Figure 7E**). The observed count of one peptide detection was obtained in 28% of 100,000 simulations, and in 15% of simulations there were no detections. Thus, the single observed detection of a noncanonical peptide at a 10^{-6} expect score threshold is within range of expectations.

No evolutionarily transient ORFs detected in MS data, even annotated ORFs with established roles

A majority of translated ORFs identified in the Wacholder et al. dataset are classified as “evolutionarily transient”, indicating that they are of recent evolutionary origin and do not show signatures of purifying selection. Of 18,947 noncanonical ORFs analyzed here, 17,471 (91%) are inferred to be evolutionarily transient in Wacholder et al.; an additional 103 canonical ORFs are also classified as transient. As these ORFs comprise such a large portion of the noncanonical translatome, we wanted to assess whether any could be detected in MS data.

No evolutionarily transient noncanonical ORFs were detected in our analyses, as none of the noncanonical proteins we identified (listed in Table 1) were classified as evolutionarily transient. Among the 103 evolutionarily transient canonical ORFs, none were detected at a 10^{-5} expect score threshold, and similar numbers of ORFs and decoys were found at weaker thresholds (Supplementary Figure 2).

Five transient canonical ORFs have been characterized in some depth[7], including MDF1, a well-established *de novo* gene specific to *S. cerevisiae* that plays a role in the yeast mating pathway.[38] Yet none of these show any evidence of detection in the MS datasets examined here, with expect scores far higher than what would constitute even weak evidence (**Table 2**). These results indicate that MS detection appears to miss the entire class of evolutionary transient ORFs, whether canonical or not.

Table 2

Canonical transient ORF	Major publication	Minimum expect score
MDF1	Li et al. 2010[38]	1.85
YBR196C-A	Vakirlis et al. 2020[39]	.99
HUR1	Omidi et al. 2018[40]	1.62
YPR096C	Hajikarimlou et al. 2020[41]	0.10
ICS3	Alesso et al. 2015[42]	0.03

Discussion

Bottom-up mass spectrometry is an attractive approach for validating noncanonical ORFs supported by ribosome profiling due to the ease of testing large lists of predicted proteins but is limited by low sensitivity. Analyzing three mass spectrometry experiments optimized to find small proteins, we identified three noncanonical proteins expressed from ORFs identified as translated in a recent analysis of yeast ribosome profiling studies (YMR106W-A, YFR035W-A, and YIL059C). We additionally found MS evidence for an ORF not initially identified by ribo-seq, YNL155C-A, due to overlapping a canonical ORF on the same strand. All four proteins were translated at rates much higher than typical noncanonical ORFs, providing independent evidence that they are genuine protein-coding genes; three also showed evidence of evolutionary conservation. These findings illustrate the power of using proteomic, translation, and evolutionary evidence in combination to identify undiscovered genes at high confidence even in a well-annotated model organism.

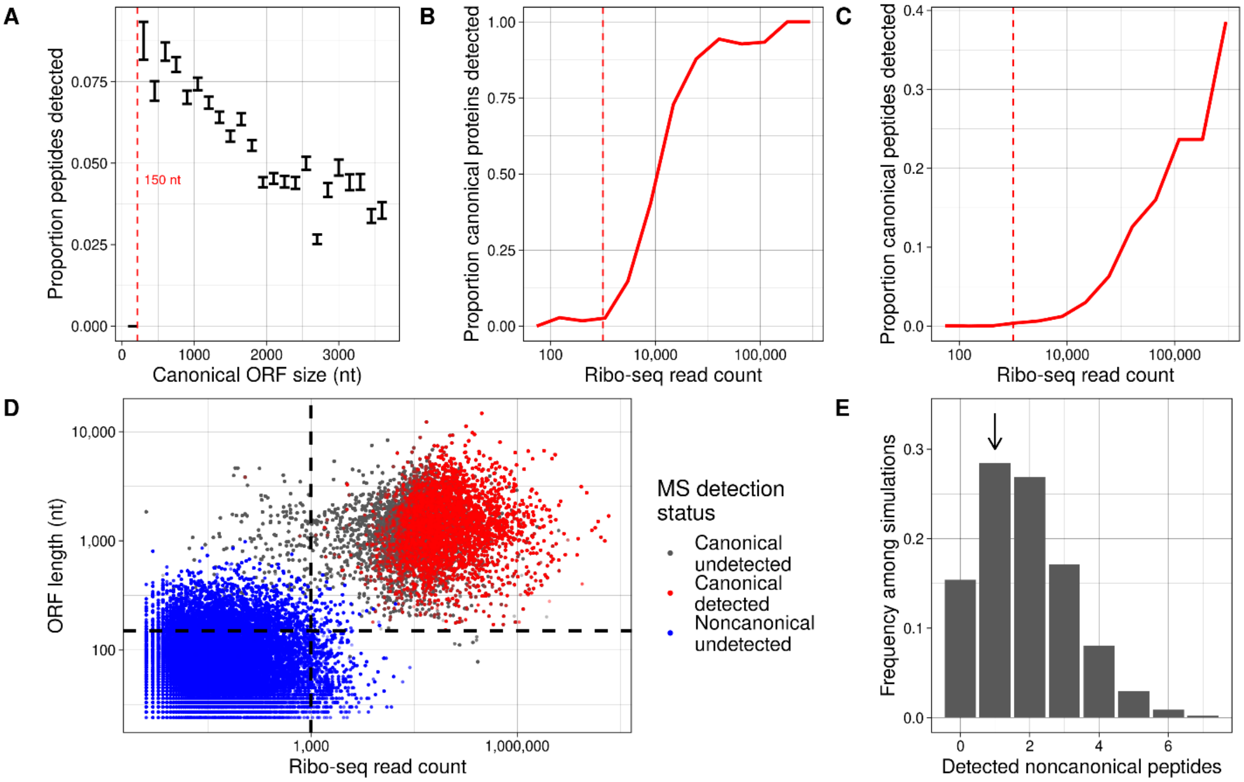


Figure 7: Lack of detection of noncanonical proteins can be explained by their low translation rate. A) The proportion of canonical peptides detected, among all eligible for detection, for ORFs of different size classes. Bars indicate a range of one standard error. A dashed line is drawn at 150 nt, below which no canonical peptides are detected. B) Proportion of canonical proteins detected within bins defined by total count of in-frame ribo-seq reads mapping to the ORF. A dashed line is drawn at 1000 reads, below which few canonical proteins are detected. C) Proportion of canonical peptides detected, out of all eligible, within bins defined by total count of in-frame ribo-seq reads mapping to the ORF. A dashed line is drawn at 1000 reads, below which few canonical peptides are detected. D) For all peptides predicted from canonical and noncanonical translated ORFs with detectable mass and length, the in-frame ribo-seq read count and ORF length is plotted. Nearly all detectable peptides are restricted to the top right quadrant, where ORF length > 150 nt and ribo-seq read count > 1000. E) The distribution of counts of noncanonical ORF peptides detected in 100,000 simulations, with peptide detection probabilities for each peptide estimated from canonical peptides encoded by ORFs with similar read counts. An arrow points to the number detected in actuality.

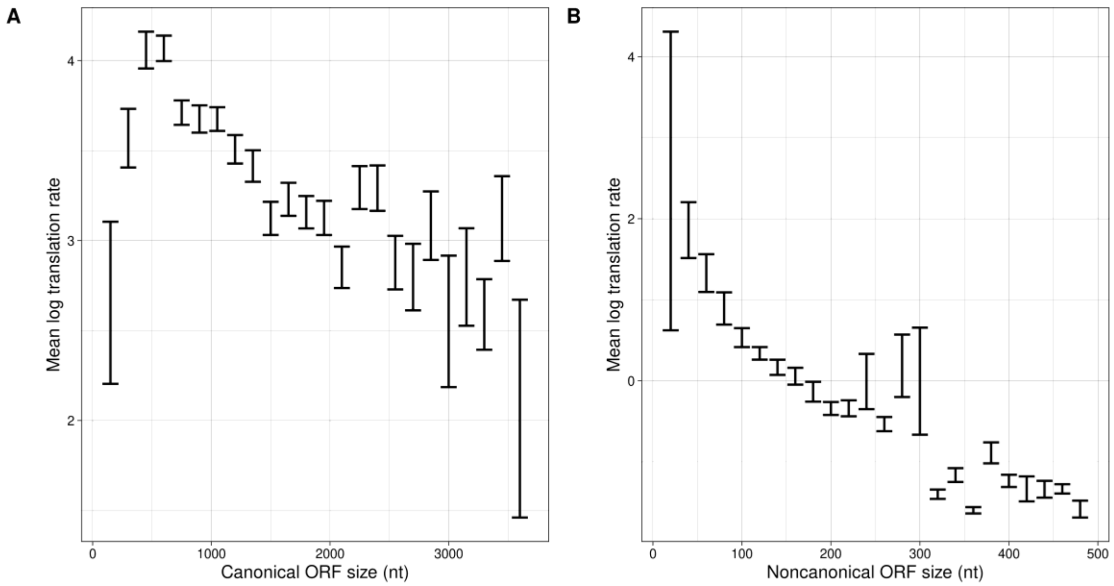
Nevertheless, the vast majority of ribo-seq supported noncanonical ORFs showed no evidence of detection in MS datasets. We show that the low rates of detection of noncanonical ORFs can be explained by their short size and low translation rate: canonical ORFs with similar levels of translation are also very rarely detected. As size and translation rate alone can explain the differences in detectability between canonical and noncanonical ORFs, little else about the biology of noncanonical ORFs can be inferred from their lack of detection in MS data. We cannot conclude that proteins expressed from noncanonical ORFs are less stable than canonical proteins, that they are targeted for degradation at higher rates, or that they are less likely to be functional, except to the extent that low expression already justifies these inferences.

A majority of the yeast noncanonical translatome, and a small portion of the canonical, consist of evolutionarily young ORFs with little evolutionary conservation, classified as “evolutionary transient ORFs” in the Wacholder et al. dataset.[7] No transient ORFs were detected in MS data, not even canonical transient ORFs that are well characterized. Evolutionary transient ORFs are both abundant in the genome and biologically significant, with some playing important roles in conserved pathways despite their short evolutionary lifespans.[7] Though we were unable to detect them in MS data, numerous proteins expressed from evolutionarily transient ORFs are found to be present in the cell in microscopy studies.[7] The biology of the vast majority of these ORFs are poorly understood; most have never been studied in any depth. Bottom-up MS, using currently available studies, does not appear useful for identifying the evolutionarily transient ORFs most likely to have interesting biological roles.

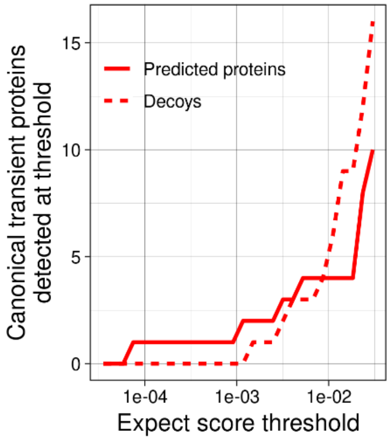
There is considerable variability across studies that attempt to detect noncanonical proteins using MS, with some reporting detection of hundreds of proteins while others, as in this study, find many fewer.[10,13,15,18,22,25,43–46] This could partly reflect biological differences between the cell types and species analyzed. However, there is also great variation in statistical approach. For example, though it is recommended for studies of noncanonical proteins to estimate a class-specific FDR among the noncanonical proteins themselves[47,48], some studies control confidence using a whole-proteome FDR (including both canonical and noncanonical), which may allow many false discoveries among the noncanonical proteins. There is a need to adopt a more consistent standard that will limit the number of false positive detections. We believe the approach employed here, in which the distribution of confidence scores among predicted noncanonical proteins and their unbiased decoys is directly compared, provides a clear picture of the extent to which noncanonical proteins can be genuinely detected.

We conclude that, while MS analysis of yeast ribo-seq supported noncanonical ORFs has some utility, it also has major limitations: it misses noncanonical proteins likely to be of biological interest, including an entire class of translated element, the evolutionarily transient ORFs. Targeted approaches such as Western blots, microscopy, and top-down MS, or new technological developments such as protein sequencing[49], are needed to better assess the cellular presence and abundance of the great majority of proteins potentially encoded by the noncanonical translatome.

Supplementary Figures



Supplementary Figure 1: Translation rate declines with ORF size. A) Average log ribo-seq read count per nucleotide among canonical ORFs of different size classes. B) Average log ribo-seq read count per nucleotide among noncanonical ORFs of different size classes.



Supplementary Figure 2: Evolutionarily transient canonical proteins found at similar rates to decoys. Predicted proteins and decoys detected in MS data at a range of expect-score thresholds, among canonical proteins identified as evolutionarily transient in Wacholder et al. 2021[50], using the standard MSFragger approach.

Methods

Mass spectrometry search

All mass spectrometry data files were taken from three studies. The He et al. 2018[23] dataset PXD008586 and Gao et al. 2021 dataset PXD001928 were downloaded from PRIDE. The Sun et al. 2022[25] dataset PXD028623 was downloaded from IPROX. These datasets were searched using all proteins predicted to be encoded from the full reference translatome described in Wacholder et al. 2021.[50] The sequence database was supplemented with all canonical proteins not included in the Wacholder et al. 2021 dataset. Canonical proteins are those annotated as "verified", "uncharacterized" or "transposable element" in the August 3, 2022 update of the Saccharomyces Genome Database annotation.[27]

Searches were conducted using the MSFragger program.[26] Unless otherwise indicated, the following parameters were used: 20 ppm precursor mass tolerance, 20 ppm fragment mass tolerance, two enzymatic termini required, up to two missed cleavages allowed, clipping to the N-terminal methionine as a variable modification, methionine oxidation as a variable modification, cysteine carbamidomethylation as fixed modification, peptide digestion lengths from 7 to 50 nt, peptide masses from 350 to 1800 Daltons, a maximum fragment charge of 2, and all other parameters as default. FDR was calculated in a class-specific manner (i.e., specific to canonical or noncanonical ORFs) by dividing the number of decoys within the class that are below the expect score threshold from the number of targets in the class lower than the threshold. Decoys were either default (reverse of protein database sequence) or reversed after the starting methionine, as indicated. Peptides were excluded if they belonged to more than one predicted protein. Peptide-spectrum matches were excluded if the MSFragger hyperscore was less than 3 above the score for the next best peptide, in order to avoid using peptide-spectrum matches that did not uniquely support a single protein.

In one analysis, searches were instead conducted using the MS-GF+ program.[32] All available parameters were set to be the same as in the MSFragger search, and decoys were reversed after the starting methionine. MS^{2}Rescore[33] was then run on MS-GF+ output files to rescore the results.

Ribo-seq data

All ribo-seq data was taken from the analysis in Wacholder et al. 2021.[50] This data included ribo-seq reads aggregated over 42 published studies and mapped to the *S. cerevisiae* genome. A read was considered to map to an ORF only if the inferred P-site mapped to the first position of a codon in the reading frame of the ORF; the total read count for an ORF is the sum of reads mapping over all first codon positions.

Homology analyses

BLAST analyses were conducted with default settings and a 10^{-4} e-value threshold to consider a match a homolog. BLAST searches conducted on NCBI databases were done on the NCBI website. Searches of the yeast genomes collected in Shen et al.[29] were conducted using the BLAST command line tool on the genomes taken from that study.[51] BLAST searches of *Saccharomyces* species genomes were conducted on genomes acquired from the following sources: *S. paradoxus* from Liti et al. 2009[52], *S. arboricolus* from Liti et al. 2013[53], *S. jurei* from Naseeb et al. 2018[54], and *S. mikatae*, *S. uvarum*, *S. eubayanus* and *S. kudriavzevii* from Scannell et al. 2011.[55] These genome were also used to make sequence alignments. All sequence alignments were generated using the MAFFT tool on the European Bioinformatics Institute website.[56]

Peptide Analysis

For each ORF in the protein database, a set of possible peptides was constructed following the same rules as used for the MSFragger analysis: two enzymatic termini (or protein ends) were required, up to two missed cleavages were allowed, clipping to the N-terminal methionine was a variable modification, and methionine oxidation was a variable modification. As in the MSFragger analysis, peptides were restricted to 7 to 50 nt and peptide masses from 350 to 1800 Daltons. Out of this list of theoretical peptides, the peptides that were detected in the MS analysis at a 10^{-6} expect score threshold in at least one experiment were identified.

Acknowledgments

We thank Jiwon Lee for helpful feedback. This work was supported by funds provided by the Searle Scholars Program to A.-R.C. and the National Institute of General Medical Sciences of the National Institutes of Health grant DP2GM137422 (awarded to A.-R.C.).

Author contributions

Conceptualization, A.W. and A.-R.C. Methodology, A.W., A.-R.C. Investigation, A.W. Writing – Original Draft, A.W. Writing – Review & Editing, A.W., A.-R.C. Supervision, A.-R.C.

Declaration of interests

A.-R.C. is a member of the scientific advisory board for Flagship Labs 69, Inc (ProFound Therapeutics).

References

1. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
2. Wright, B. W., Yi, Z., Weissman, J. S. & Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* **32**, 243–258 (2022).
3. Jackson, R. *et al.* The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434–438 (2018).
4. Pauli, A. *et al.* Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **343**, 1248636 (2014).
5. Herberg, S., Gert, K. R., Schleiffer, A. & Pauli, A. The Ly6/uPAR protein Bouncer is necessary and sufficient for species-specific fertilization. *Science* **361**, 1029–1033 (2018).
6. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).

7. Wacholder, A. *et al.* A vast evolutionarily transient translatome contributes to phenotype and fitness. 2021.07.17.452746 Preprint at https://doi.org/10.1101/2021.07.17.452746 (2023).

8. Durand, É. *et al.* Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **29**, 932–943 (2019).

9. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).

10. Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).

11. Makarewich, C. A. & Olson, E. N. Mining for Micropeptides. *Trends Cell Biol.* **27**, 685–696 (2017).

12. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).

13. Chothani, S. P. *et al.* A high-resolution map of human RNA translation. *Mol. Cell* **82**, 2885-2899.e8 (2022).

14. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).

15. Duffy, E. E. *et al.* Developmental dynamics of RNA translation in the human brain. *Nat. Neurosci.* **25**, 1353–1365 (2022).

16. Wolters, D. A., Washburn, M. P. & Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).

17. Ahrens, C. H., Wade, J. T., Champion, M. M. & Langer, J. D. A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. *J. Bacteriol.* **204**, e00353-21 (2022).

18. Zheng, E. B. & Zhao, L. Protein evidence of unannotated ORFs in Drosophila reveals diversity in the evolution and properties of young proteins. *eLife* **11**, e78772 (2022).

19. Ouspenskaia, T. *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).

20. Lu, S. *et al.* A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res.* **47**, 8111–8125 (2019).

21. Yofe, I. *et al.* One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat. Methods* **13**, 371–378 (2016).

22. van Heesch, S. *et al.* The Translational Landscape of the Human Heart. *Cell* **178**, 242-260.e29 (2019).

23. He, C., Jia, C., Zhang, Y. & Xu, P. Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae. *J. Proteome Res.* **17**, 2335–2344 (2018).

24. Gao, Y. *et al.* Mass-Spectrometry-Based Near-Complete Draft of the Saccharomyces cerevisiae Proteome. *J. Proteome Res.* **20**, 1328–1340 (2021).

25. Sun, Y., Huang, J., Wang, Z., Pan, N. & Wan, C. Identification of Microproteins in Saccharomyces cerevisiae under Different Stress Conditions. *J. Proteome Res.* **21**, 1939–1947 (2022).

26. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).

27. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).

28. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. in *Proteome Bioinformatics* (eds. Hubbard, S. J. & Jones, A. R.) 55–71 (Humana Press, 2010). doi:10.1007/978-1-60761-444-9_5.

29. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545.e20 (2018).

30. Czaja, W., Bensasson, D., Ahn, H. W., Garfinkel, D. J. & Bergman, C. M. Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. *PLOS Genet.* **16**, e1008632 (2020).

31. Willingham, S., Outeiro, T. F., DeVit, M. J., Lindquist, S. L. & Muchowski, P. J. Yeast Genes That Enhance the Toxicity of a Mutant Huntingtin Fragment or α-Synuclein. *Science* **302**, 1769–1772 (2003).

32. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).

33. Declercq, A. *et al.* MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell. Proteomics* **21**, (2022).

34. Herst, P. M., Perrone, G. G., Dawes, I. W., Bircham, P. W. & Berridge, M. V. Plasma membrane electron transport in Saccharomyces cerevisiae depends on the presence of mitochondrial respiratory subunits. *FEMS Yeast Res.* **8**, 897–905 (2008).

35. Wilson, W. A., Wang, Z. & Roach, P. J. Systematic Identification of the Genes Affecting Glycogen Storage in the Yeast Saccharomyces cerevisiae: Implication of the Vacuole as a Determinant of Glycogen Level * S. *Mol. Cell. Proteomics* **1**, 232–242 (2002).

36. Hoepfner, D. *et al.* High-resolution chemical dissection of a model eukaryote reveals targets, pathways and gene functions. *Microbiol. Res.* **169**, 107–120 (2014).

37. Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, (2019).

38. Li, D. *et al.* A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–420 (2010).

39. Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).

40. Omidi, K. *et al.* Uncharacterized ORF HUR1 influences the efficiency of non-homologous end-joining repair in Saccharomyces cerevisiae. *Gene* **639**, 128–136 (2018).

41. Hajikarimlou, M. *et al.* Sensitivity of yeast to lithium chloride connects the activity of YTA6 and YPR096C to translation of structured mRNAs. *PLOS ONE* **15**, e0235033 (2020).

42. Alesso, C. A., Discola, K. F. & Monteiro, G. The gene ICS3 from the yeast Saccharomyces cerevisiae is involved in copper homeostasis dependent on extracellular pH. *Fungal Genet. Biol.* **82**, 43–50 (2015).

43. Wang, S. *et al.* Large-Scale Discovery of Non-conventional Peptides in Maize and Arabidopsis through an Integrated Peptidogenomic Pipeline. *Mol. Plant* **13**, 1078–1093 (2020).

44. Budamgunta, H. *et al.* Comprehensive Peptide Analysis of Mouse Brain Striatum Identifies Novel sORF-Encoded Polypeptides. *PROTEOMICS* **18**, 1700218 (2018).

45. Cao, X. *et al.* Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J. Proteome Res.* **19**, 3418–3426 (2020).

46. Bogaert, A. *et al.* Limited Evidence for Protein Products of Noncoding Transcripts in the HEK293T Cellular Cytosol. *Mol. Cell. Proteomics MCP* **21**, 100264 (2022).

47. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).

48. Woo, S. *et al.* Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.* **14**, 3555–3567 (2015).

49. Floyd, B. M. & Marcotte, E. M. Protein Sequencing, One Molecule at a Time. *Annu. Rev. Biophys.* **51**, 181–200 (2022).

50. Wacholder, A., Acar, O. & Carvunis, A.-R. A reference translatome map reveals two modes of protein evolution. 2021.07.17.452746 Preprint at https://doi.org/10.1101/2021.07.17.452746 (2021).

51. Ye, J., McGinnis, S. & Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **34**, W6–W9 (2006).

52. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).

53. Liti, G. *et al.* High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. *BMC Genomics* **14**, 69 (2013).

54. Naseeb, S. *et al.* Whole Genome Sequencing, de Novo Assembly and Phenotypic Profiling for the New Budding Yeast Species Saccharomyces jurei. *G3 Genes Genomes Genet.* **8**, 2967–2977 (2018).

55. Scannell, D. R. *et al.* The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 Genes Genomes Genet.* **1**, 11–25 (2011).

56. Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580–W584 (2015).