



**HAL**  
open science

## The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation

Chris Papadopoulos, Hugo Arbes, David Cornu, Nicolas Chevrollier, Sandra Blanchet, Paul Roginski, Camille Rabier, Safiya Atia, Olivier Lespinet, Olivier Namy, et al.

### ► To cite this version:

Chris Papadopoulos, Hugo Arbes, David Cornu, Nicolas Chevrollier, Sandra Blanchet, et al.. The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation. *Genome Biology*, 2024, 25 (1), pp.268. 10.1186/s13059-024-03403-7 . hal-04765906

**HAL Id: hal-04765906**

**<https://hal.science/hal-04765906v1>**

Submitted on 4 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH

Open Access



# The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation

Chris Papadopoulos<sup>1,2</sup>, Hugo Arbes<sup>1†</sup>, David Cornu<sup>1†</sup>, Nicolas Chevrollier<sup>3</sup>, Sandra Blanchet<sup>1</sup>, Paul Roginski<sup>1</sup>, Camille Rabier<sup>1</sup>, Safiya Atia<sup>1</sup>, Olivier Lespinet<sup>1</sup>, Olivier Namy<sup>1</sup> and Anne Lopes<sup>1\*</sup> 

<sup>†</sup>Hugo Arbes and David Cornu contributed equally to this work.

\*Correspondence: anne.lopes@i2bc.paris-saclay.fr

<sup>1</sup>Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, Cedex 91198, France

<sup>2</sup>Hospital del Mar Research Institute, Barcelona, Spain

<sup>3</sup>Independent Investigator, Paris, France

## Abstract

**Background:** Pervasive translation is a widespread phenomenon that plays a critical role in the emergence of novel microproteins, but the diversity of translation patterns contributing to their generation remains unclear. Based on 54 ribosome profiling (Ribo-Seq) datasets, we investigated the yeast Ribo-Seq landscape using a representation framework that allows the comprehensive inventory and classification of the entire diversity of Ribo-Seq signals, including non-canonical ones.

**Results:** We show that if coding regions occupy specific areas of the Ribo-Seq landscape, noncoding regions encompass a wide diversity of Ribo-Seq signals and, conversely, populate the entire landscape. Our results show that pervasive translation can, nevertheless, be associated with high specificity, with 1055 noncoding ORFs exhibiting canonical Ribo-Seq signals. Using mass spectrometry under standard conditions or proteasome inhibition with an in-house analysis protocol, we report 239 microproteins originating from noncoding ORFs that display canonical but also non-canonical Ribo-Seq signals. Each condition yields dozens of additional microprotein candidates with comparable translation properties, suggesting a larger population of volatile microproteins that are challenging to detect. Our findings suggest that non-canonical translation signals may harbor valuable information and underscore the significance of considering them in proteogenomic studies. Finally, we show that the translation outcome of a noncoding ORF is primarily determined by the initiating codon and the codon distribution in its two alternative frames, rather than features indicative of functionality.

**Conclusion:** Our results enable us to propose a topology of a species' Ribo-Seq landscape, opening the way to comparative analyses of this translation landscape under different conditions.

**Keywords:** Noncoding genome, Pervasive translation, Genome evolution, Non-canonical translation signals, De novo coding products



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## Background

All organisms undergo molecular innovation to adapt to their environment. Typically, molecular innovation involves the creation of novel products (i.e., genes, proteins) or the formation of novel interactions between already existing products. The former case implies the existence of a “reservoir” of neutrally and/or fast-evolving sequences, which allows for the sampling of extensive sequence spaces that could not be reached under natural selection. In theory, to ensure the efficiency of the functional processes that occur in the cell, this reservoir is expected to be locked (i.e., not expressed), to avoid unselected products that could interfere with established ones. Nevertheless, from an evolutionary perspective, the pervasiveness of biological processes (i.e., transcription, translation...) enables the expression of small fractions of novel sequence spaces, thereby exposing them to selection and allowing, from time to time, the passage from this reservoir to the world of established, and selected products.

In fact, noncoding regions can be seen as a reservoir of unselected sequences, hosting thousands of small Open Reading Frames (ORFs) that could give rise to novel products if translated [1–6]. Precisely, OMICS technologies have provided a huge amount of data revealing the “omnipresence” of biological noise which has turned out to result from the pervasiveness of biological processes. As a matter of fact, noncoding regions have been shown to be pervasively transcribed and translated, exposing non-genic sequences to selection [4, 7–26]. Notably, hundreds of novel peptides or microproteins resulting from presumed noncoding regions have been confirmed by proteomics [19, 27–35]. Finally, many studies report examples of *de novo* gene birth from noncoding regions in eukaryotic species [24, 36–61]. These *de novo* genes exhibit clear regulation patterns, may be subject to purifying selection, and have reported function for some of them, confirming that they could undoubtedly be associated with the coding world [37, 38, 46, 60, 62, 63]. If different models of *de novo* gene emergence have been proposed so far, most of them share the hypothesis of an early stage as a protogene or small peptide that results from the translation of noncoding regions [10, 20, 21, 47, 50, 53, 57]. These models consequently attribute an important role to pervasive translation in *de novo* gene birth. Indeed, pervasive translation constitutes somehow the last step for a noncoding ORF to reach the protein state, a prerequisite, though not sufficient, to enter the coding world. All the aforementioned studies, therefore, (i) reveal that the passage from the noncoding to the coding world is much more frequent than previously thought and (ii) place the noncoding genome, but also pervasiveness at the center of the emergence of genetic novelty enabling, from time to time, the passage from unselected products to functional and selected ones.

Each RNA molecule contains three reading frames, each hosting distinct ORFs that encode different amino acid sequences. In coding regions, translation is strongly regulated to ensure the production of the functional product. This regulation results in a highly specific translation process biased towards the frame hosting the coding sequence (CDS). Consequently, ribosome profiling (Ribo-Seq) reads exhibit a remarkable triplet periodicity, with the two alternative frames of the CDS being substantially depleted in reads. As such, the canonical Ribo-Seq patterns of regulated translation are easily detectable by Ribo-Seq analysis tools. However, in noncoding regions undergoing pervasive and, therefore, nonregulated translation, the diversity of Ribo-Seq patterns remains

elusive, while key to apprehending their potential for novel protein products. Yet, pervasive translation is usually studied with methods tailored to coding regions, and the focus is put on canonical Ribo-Seq signals that resemble those of coding sequences undergoing regular translation. Nonetheless, the discarded non-canonical Ribo-Seq signals might hold valuable insights, potentially uncovering translation events associated with de novo products.

Therefore, we inventoried and characterized the entire diversity of Ribo-Seq patterns associated with the noncoding regions of *Saccharomyces cerevisiae*, with the aim to further explore the extent to which pervasive translation can give rise to novel protein products. To do so, we devised a representation framework that maps all the ORFs lying in ribosome-associated RNAs onto a 2D plane, irrespective of their translation status. Each ORF is mapped based on the fractions of reads in its frame and the two alternative ones without any a priori on its translation outcome, providing a comprehensive inventory and a rational classification of all detected ribosome footprints of an ORFeome of interest. As such, we characterized the Ribo-Seq patterns associated with both yeast non-coding regions and coding ones (i.e., CDSs and their alternative frames). The Ribo-Seq signals of coding regions not only served as a reference for studying the characteristics of pervasive translation but also allowed us to reveal non-canonical translation events in coding regions. We then conducted mass spectrometry (MS) to investigate whether the different types of Ribo-Seq signals that we identified, including non-canonical ones, could be associated with a protein product. Finally, we examined the properties of non-coding ORFs with respect to their Ribo-Seq patterns to better understand the rules that dictate their translation outcome and that, ultimately, underlie pervasive translation.

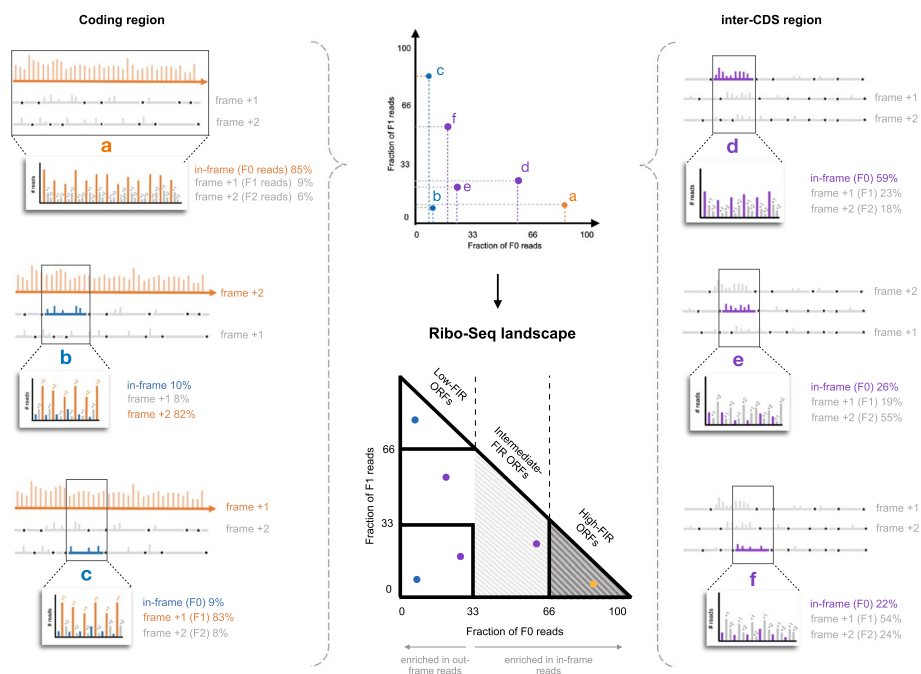
## Results

### Collection of Ribo-Seq experiments

Ribo-Seq involves sequencing, after RNA digestion, the fragments protected by the translating ribosomes (i.e., reads), thereby identifying the RNA regions that were bound by ribosomes [8]. High-quality Ribo-Seq experiments can detect the translation signal at the single nucleotide resolution, allowing for the precise detection of the translated codons and, therefore, of the translated RNA frame. In coding regions, since the coding frame is known, the translated frame is straightforward to detect, even when dealing with intermediate or poor-quality Ribo-Seq experiments. In contrast, in noncoding regions, identifying the translated frame requires high-quality Ribo-Seq data since there is no prior knowledge of the frame that is translated [64]. Detecting pervasively translated ORFs is challenging due to their short size and typically low expression levels. Furthermore, these ORFs are not expected to be expressed constitutively but, instead, are likely to be translated in only a few experiments. To increase the ribosome profiling signal and capture both occasional and constitutive translation events, we assembled a significant collection of 89 publicly available Ribo-Seq datasets from yeast. Specifically, we focused on datasets performed under standard conditions on the wild-type S288C or BY4741 strains of *S. cerevisiae* so as not to be biased by experimental conditions or mutants that could induce translation deregulation. After filtering out poor-quality datasets that could affect our conclusions, we retained 54 high-quality datasets (see the “[Methods](#)” section).

### Representation framework

In this study, we aimed to capture the full diversity of Ribo-Seq footprints associated with noncoding regions without prior knowledge of their translation status. This diversity may encompass canonical Ribo-Seq signals indicative of highly specific translation, as previously reported in [21, 22, 26, 33, 64]. It may also include Ribo-Seq patterns reflecting varying translation specificities or scanning ribosomes without translation outcome. As a reference, we also aim to examine the ribosome footprints associated with coding regions, including CDSs and the ORFs lying in their alternative frames (aORFs). CDSs and aORFs are anticipated to show a limited diversity of Ribo-Seq patterns. Indeed, CDSs are expected to display canonical Ribo-Seq patterns characterized by consistent in-frame read triplet periodicity (see the metagene of ORF “a” in Fig. 1). In contrast, aORFs are likely to exhibit shifted periodicity with reads predominantly lying in the +1 or +2 frames, reflecting the translation of the overlapping CDS (see metagenes of ORFs “b” and “c” in Fig. 1). Comparing the ribosome footprint diversity of coding and noncoding regions will comprehensively characterize the ribosome activity associated



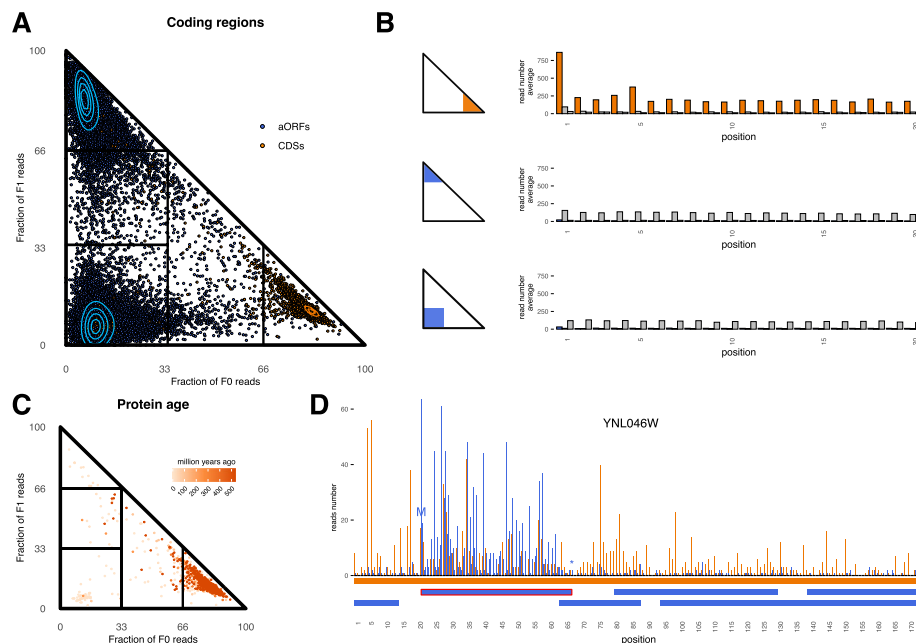
**Fig. 1** Construction of the Ribo-Seq landscape. Left: example of calculation of the fractions of F0, F1, and F2 reads of one CDS (a in orange) and two aORFs (b and c in blue). Each Ribo-Seq read (vertical bars) can be assigned to a unique nucleotide according to its genomic coordinates and associated with a specific frame. For each considered ORF, we calculate its metagene, reflecting the distribution of reads over the ORF’s F0, F1, and F2 frames, as well as the read fractions in the three frames (see Methods). The ORF is then represented on a 2D graph according to its fractions of reads in its F0 and F1 frames (the fraction of reads in its +2 frame (F2) can be calculated as  $F2 = 1 - (F0 + F1)$ ) (middle—top part). As such, the CDS a is associated with the coordinates (85,9). In contrast, the aORF b whose metagene exhibits a +2 shifted read periodicity will be associated with the coordinates (10, 8). Right: The same is done for inter-CDS regions. Theoretically, scanning ribosomes are expected to lead to reads evenly distributed across the three frames, leading approximately to  $F0 = F1 = F2 = 33\%$ . Above 33% of in-frame reads, ORFs are considered enriched in in-frame reads. Despite this enrichment, we do not pretend that all these ORFs are translated. ORF Ribo-Seq status (low-, intermediate-, and high-FIR) are defined according to the fractions of reads in their F0 frame, reflecting different levels of enrichment for F0 reads (middle—bottom part)

with these distinct genomic regions. Therefore, we categorized the yeast genome into two genomic classes: (i) coding regions, comprising the yeast's 6669 CDSs and the 96,873 aORFs hosted in their alternative frames, and (ii), noncoding regions, consisting of 78,989 ORFs (referred to as iORFs) located in the inter-CDS regions, and which do not overlap any annotated feature of *S. cerevisiae* genome (see “Methods” section). Consistent with our previous work and to ensure enough statistical power for the analysis of Ribo-Seq read patterns or other feature analyses, both aORFs, and iORFs have a minimum size of 20 codons (see the “Methods” section). Additionally, to detect non-canonical translation initiation events, iORFs, and aORFs are not required to start with ATG codons, and are defined from STOP to STOP.

All reads extracted from the 54 Ribo-Seq runs were pooled together and mapped to the CDSs, aORFs, and iORFs with our pipeline ORFribo (see Methods and Additional file 1: Fig. S1). The 90th percentile of them has fewer than 49 reads. Since we aim to describe non-canonical patterns, we need the best possible description of their read distribution along the ORF. Therefore, we decided to focus on the 10% of iORFs with the most information and retained all the ORFs associated with at least 50 reads, regardless of the frame, and whose read codon coverage was above 30%. This allowed us to ensure enough signal to estimate the read distribution along the ORF and discard cases where reads accumulate on very few codons that may reflect sequencing and/or library artifacts. For each ORF, we calculated its fraction of in-frame reads (reads mapping to the frame of the ORF, named by convention frame 0 (F0)), and that of its out-of-frame reads mapping to its +1 and +2 frames (F1 and F2, respectively) (Fig. 1). We then devised a representation framework that provides the Ribo-Seq landscape of a set of ORFs of interest. In this 2D Ribo-Seq landscape, each ORF is positioned according to its fractions of reads in the F0 and F1 frames with no a priori on its translation outcome (Fig. 1). It is worth noting that the fraction of F2 frame reads can be directly deduced from those of F0 and F1 with  $F2 = 1 - (F0 + F1)$ . The  $x$ -axis is divided into three equivalent regions to distinguish (i) high-FIR ORFs (i.e., high fraction of in-frame reads) whose reads are mostly in-frame, reflecting ORFs translated with high specificity (e.g., CDSs), and (ii) low-FIR ORFs whose reads are mostly out-of-frame, which, in theory, reflect untranslated ORFs overlapping translated ones (e.g., aORFs). They may nevertheless be associated with a few in-frame reads due to misassignment of the translated frame. Finally, the third category, (iii), corresponds to intermediate-FIR ORFs that display enrichment in in-frame reads compared to a random distribution of reads along the RNA ( $F0$  fraction  $\geq 0.33$ ), though this enrichment is lower than that of ORFs translated with high specificity. As such, the  $x$ -axis reflects the specificity of an ORF's Ribo-Seq reads, while the  $y$ -axis illustrates the ribosome occupancy in its two alternative frames. Overall, the resulting Ribo-Seq landscape offers a rational and comprehensive representation of the diversity of the Ribo-Seq footprints of an ORFeome of interest that can be associated with translation outcomes or not. This landscape offers a promising representation for exploring the relationship between overlapping ORFs, as exemplified in Fig. 1, with CDSs and aORFs.

### Ribo-Seq landscape of coding regions

As expected, most of the read-associated CDS (5404—93%) fall within the high-FIR region of the Ribo-Seq landscape ( $F0 \geq 0.66$ ), highlighting their highly specific



**Fig. 2** Ribo-Seq landscape of coding regions. **A** Ribo-Seq landscape of the coding regions' ORFs, including CDSs and aORFs, colored orange and blue, respectively. Each ORF with at least 50 reads and 30% of codon coverage is represented as a dot on the Ribo-Seq landscape according to its fractions of F0 reads (x-axis) and F1 reads (y-axis). **B** Metagenes of the three extremities of the Ribo-Seq landscape. Each metagene is calculated for the first 20 codons from all the ORFs associated with a given Ribo-Seq area (see the "Methods" section) and represents the average number of reads per nucleotide over all the ORFs of the metagene. The corresponding Ribo-Seq landscape areas are colored on the schematic Ribo-Seq landscapes of each plot. Nucleotide positions that are covered by in-frame reads relative to the ORFs of the metagene are colored in orange or blue, while positions covered by out-of-frame reads (F1 and F2) are colored in gray. **C** Ribo-Seq landscape of the CDSs colored with respect to their evolutionary age (in Mya) as estimated by phylostratigraphy (see the "Methods" section). **D** Top: Metagene of the CDS YNL046W representing the number of reads of each nucleotide summed over the 54 Ribo-Seq datasets. Positions that are covered by reads in frame with the CDS YNL046W are colored in orange, while those covered by reads in its +1 and +2 frames are colored in blue. Bottom: Colored rectangles represent the CDS (orange) and its aORFs (blue). The start and stop codons of aORF chrXIV:542,365–542,502 (surrounded in red) are indicated by an "M" and a star, respectively

translation (Fig. 2A, Additional file 1: Table S1). Their metagene profile further supports their translation specificity, with reads predominantly accumulating in their F0 frame (Fig. 2B). Conversely, 97% (76,416/78,428) of the read-associated aORFs exhibit high fractions of out-of-frame reads ( $F0 < 0.33$ ) and are located at the two left extremities of the Ribo-Seq landscape. The out-of-frame reads of aORFs are likely not indicative of a translation outcome but instead mirror the translation of the overlapping CDSs since the profiles of overlapping ORFs are inherently correlated (Fig. 2B). Overall, the three extremities of the Ribo-Seq landscape highlight RNA regions associated with an unbalanced ribosome occupancy between the three RNA frames in favor of that of the CDS.

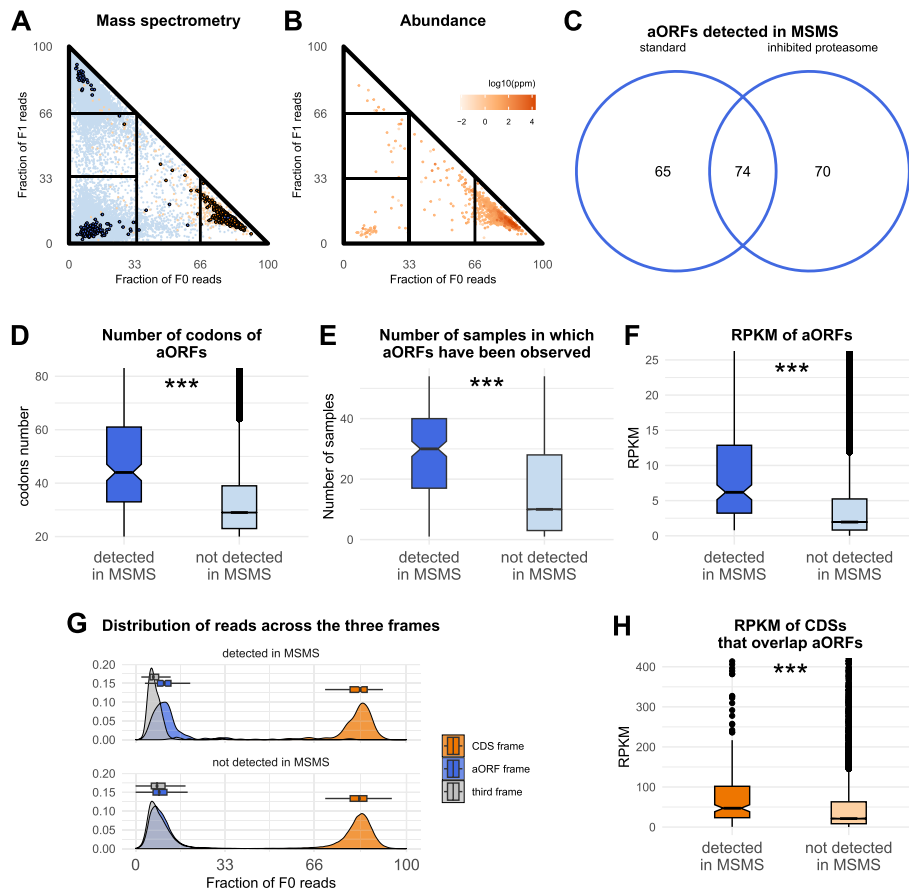
Interestingly, 3% of CDSs (176 cases) fall within the intermediate- and low-FIR regions of the Ribo-Seq landscape ( $F0 < 0.66$ ). 66% of them are specific to *Saccharomyces sensu stricto* (Fig. 2C), among which 48% overlap on the same strand, another CDS that is older and specifically translated. We can hypothesize that these young or emerging ORFs have not yet acquired the regulatory elements necessary for efficient expression (e.g., transcription or translation levels). However, 66% of the CDSs associated with high fractions

of out-of-frame reads do not overlap another CDS. We carefully analyzed their individual metagene profiles and present an illustrative example where the ribosome footprints are shared between a CDS and one of its aORFs (Fig. 2D). Interestingly, if in-frame reads cover the whole sequence of the CDS YNL046W, reads in the +1 frame accumulate between the codons 21 and 67. This region precisely corresponds to the genomic coordinates of the aORF chrXIV:542,365–542,502, which is enriched in in-frame reads (fraction of F0 reads: 55.6%). The first peak of the in-frame reads of the aORF lies at an ATG codon and probably indicates the translation initiation. The fact that the high density of the out-of-frame reads of the gene YNL046W overall correlates with the borders delineated by the first ATG and the STOP codon of the aORF strongly supports the translation of this aORF. In fact, a significant portion (~38%, 67/176) of the low- or intermediate-FIR CDSs overlap aORFs associated with intermediate or high fractions of in-frame reads, which could indicate an alternative translation outcome, although this remains to be fully demonstrated. It is unclear whether these cases reflect different RNA molecules that are independently translated or a competition occurring at the translation level between overlapping ORFs. Nonetheless, our results suggest that despite their low fractions of in-frame reads, these ORFs might be associated with a translation outcome and, therefore, warrant further investigation.

#### Detection of translated products with mass spectrometry

To determine whether CDSs and aORFs with intermediate or low fractions of in-frame reads could be associated with a protein product, we conducted MS experiments under standard conditions or conditions where the proteasome is inhibited (see the “Methods” section). aORFs are short, may encode unstable peptides, and if translated, the abundance of their resulting peptides is expected to be very low, thereby rendering their detection very challenging. Inhibiting the proteasome should, therefore, increase the detection of unstable and/or short-lived peptides that are typically rapidly degraded by the proteasome. Furthermore, these low-abundance peptides are typically very difficult to detect with classical MS analysis protocols, especially when the database being screened (all read-associated aORFs in this case) is much larger than the number of anticipated candidates [65, 66]. Indeed, we recall that most aORFs, despite being associated with reads, are likely not actually translated but instead reflect the translation of the overlapping CDS. Consequently, the proportion of truly expressed aORFs detectable by MS is expected to be very low. To address this, we devised a two-step procedure that involves an initial search on a large database composed of all candidates associated with Ribo-Seq reads. This step is then followed by a second search on a much smaller database, primarily consisting of candidates identified in the first round (see the “Methods” section for details). Doing so, our MS experiment identified 3505 CDSs, 2785 of which were detected under both conditions. The latter are associated with higher read counts and translated in almost all Ribo-Seq experiments, indicating highly and constitutively expressed CDSs that are easy to detect with MS (Additional file 1: Fig. S2). By inhibiting the proteasome, we could identify an additional 627 CDSs. The latter exhibit lower half-lives than those detected in standard conditions (Additional file 1: Fig. S2). This likely explains our difficulty in detecting them when the proteasome is active and presents



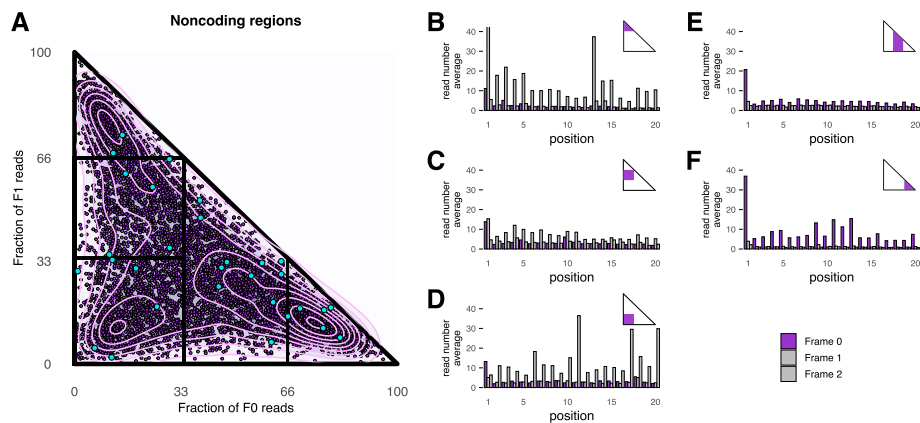


**Fig. 3** **A** Ribo-Seq landscape of the aORFs and CDSs observed in our MS data. Same representation as Fig. 2A, but aORFs and CDSs for which we detected a peptide product are colored in dark blue or dark orange, respectively. **B** Ribo-Seq landscape of the CDSs colored according to their protein abundance in ppm as provided by PaxDb [67]. **C** Venn diagram of the number of aORFs detected with MS by the two experimental conditions, standard (left) and inhibited proteasome (right). **D** Number of codons of aORFs detected with MS experiments (blue) and aORFs that were not detected (light blue). **E** Number of Ribo-Seq datasets in which an aORF has been observed for detected (blue) and not detected aORFs (light blue). **F** Reads per kilobase of exon model per Million mapped reads (RPKM) computed over all datasets for detected (blue) and not detected aORFs (light blue). **G** Distribution of the reads across the three overlapping frames for the aORFs detected in MS (top) and aORFs not detected in MS (bottom). **H** RPKM computed over all datasets of CDSs that overlap aORFs detected with MS (orange) and CDSs whose overlapping aORFs were not detected (light orange)

proteasome inhibition as promising for detecting short-lived products. Nine CDSs with intermediate and low fractions of in-frame reads were also detected, revealing that non-canonical Ribo-Seq patterns may occasionally be linked to a protein product (Fig. 3A). Interestingly, all of them were detected with proteasome inhibition, with only 2 being also detected under standard conditions. We then extracted protein abundances from the PaxDb database [67], a curated meta-resource that integrates multiple mass spectrometry datasets for which abundances were reprocessed, unified, and scored (Fig. 3B). 80% of the low- and intermediate-FIR CDSs were detected in the MS data of PaxDb. Although they display lower abundances than high-FIR CDS (median comparison empirical  $P$ -value  $< 1e - 04$ —see the “Methods” section for

details), these results strengthen the hypothesis that low fractions of in-frame reads may reflect effective translation.

We detected a total of 209 aORFs under standard conditions and/or with protease inhibition. Equivalent amounts of aORFs were detected by each experiment, with 74 identified under both conditions (Fig. 3C, Additional file: Table S2). Even though the portion of aORFs detected under both conditions is significant (~50% of the aORFs detected under each condition), this portion is lower than what was observed for CDSs (~80%). This probably results from the difficulty in capturing such short, lowly abundant, and volatile ORFs whose expression may be inconsistent across experiments. Interestingly, the aORFs detected under a single condition display similar properties to those identified under both conditions while being significantly distinguishable from those that were not detected. Notably, the detected aORFs are longer (median comparison empirical  $P$ -value  $< 1e - 04$ ) (Fig. 3D). While this may increase their number of theoretical tryptic peptides and thus lead to more false identifications, their longer size may also contribute to greater overall stability, further enhancing their detectability. Interestingly, they also exhibit distinct Ribo-Seq properties compared to other aORFs. They are detected in more Ribo-Seq experiments (median comparison empirical  $P$ -value  $< 1e - 04$ ) (Fig. 3E). They display higher RPKM values, are associated with more reads than the other alternative frame, and overlap CDSs also characterized by higher RPKM values (median comparison empirical  $P$ -values all  $< 1e - 04$ ) (Fig. 3EH, Additional file 1: Fig. S3, Fig. S4 for the analysis per condition). Finally, we applied our proteomic analysis pipeline to another dataset [68]. We found 99 aORFs, among which seven were also detected in our data. While the overlap is smaller than that of the two conditions tested in the present study, it is unlikely to happen by chance (during the proteomic search, we screened a database of 76 k aORFs, see Methods). In particular, the 99 ORFs detected in He et al. [68] exhibit similar properties to those detected in our experiment (Additional file 1: Fig. S5). While confidently confirming the detection of such volatile and low-abundance non-canonical ORFs is a non-trivial task, these observations altogether seem to further support their detection and suggest that we may be underestimating their numbers. Overall, these results suggest that (i) benefiting from the higher expression of their overlapping CDS, these aORFs may be translated at higher rates, and (ii) regions with high expression activity may be more likely to produce, through pervasive expression, microproteins in sufficient quantity to be detected with proteomics. This leads us to hypothesize that regions of high expression activity are more likely to generate novel coding products. However, the relationship between the quantity of a microprotein and its capacity to ensure a biological role and ultimately to be fixed requires further investigation. For instance, these aORFs are not more conserved among yeast's neighboring species than the others (ORF age median comparison empirical  $P$ -value  $> 0.05$ ), suggesting that their detection does not reveal emerging ORFs but may rather result from the high expression of their overlapping CDSs. Finally, 96% of these aORFs are located in the low-FIR area of the Ribo-Seq landscape, again supporting that this region can occasionally be associated with translation outcomes (Fig. 3A). Nevertheless, this result does not imply that all low-FIR ORFs are translated. Indeed, their metagenes strongly suggest a non-translational outcome for the majority of them, as

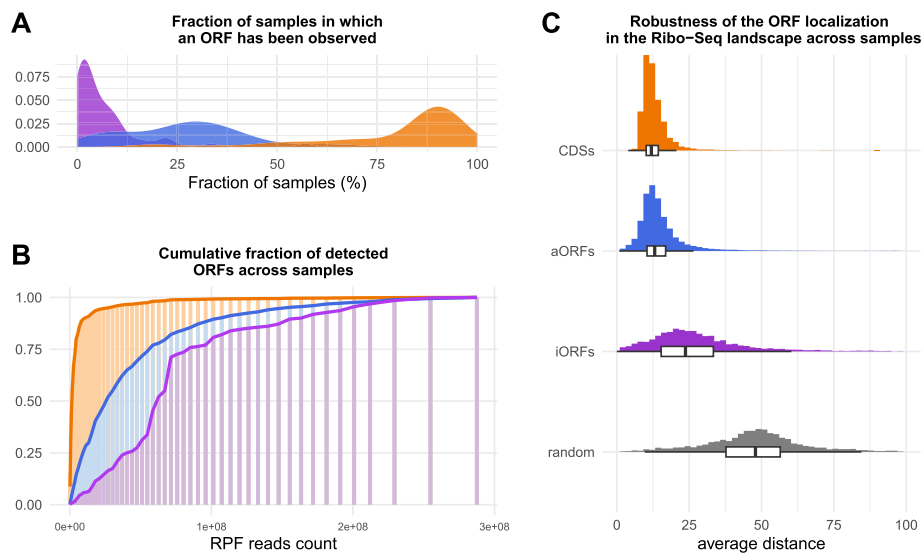


**Fig. 4** Ribo-Seq landscape of noncoding regions. **A** Ribo-Seq landscape of iORFs. Each read-associated iORF is represented as a purple dot according to its fractions of F0 reads (x-axis) and F1 reads (y-axis). Cyan dots correspond to iORFs detected with MS. **B–F** Metagenes of the five regions of the Ribo-Seq landscape representing the average number of Ribo-Seq reads of each nucleotide for the first 20 codons. Nucleotide positions covered by in-frame reads (F0) for the considered iORFs are represented in purple, while those covering reads in the +1 and +2 frames are colored in gray

observed for most aORFs of the gene YNL046W (Fig. 2D). Our MS data, however, seem to indicate that a small fraction of them is translated in enough quantities to be detected at the protein level.

#### Ribo-Seq landscape of noncoding regions

We identified 6814 iORFs with at least 50 reads and a read codon coverage of 30%. Contrarily to coding regions, noncoding regions are associated with a wide diversity of Ribo-Seq signals that populate the entire yeast Ribo-Seq landscape. Specifically, we reveal a continuum in the Ribo-Seq signals that is absent from coding regions and appears to be a hallmark of the pervasive expression of noncoding regions (Fig. 4A). Higher densities of iORFs are nonetheless observed at the three extremities of the Ribo-Seq landscape, suggesting RNA regions translated with high specificity. Notably, we report 1055 high-FIR iORFs and 2262 low-FIR iORFs located at the two left extremities of the Ribo-Seq landscape. The metagene of the former is typical of canonical CDSs, while those of the latter resemble those of aORFs, and rather mirror the translation activity occurring in their alternative frames (Fig. 4BDF). 33% of read-associated iORFs fall within the intermediate-FIR region. Theoretically, scanning ribosomes are expected to be associated with reads evenly distributed across the three frames (i.e., fractions of reads in F0, F1, F2  $\sim 0.33$ ). Interestingly, this twilight zone also includes iORFs enriched in in-frame reads, which may reflect translation. Their metagene reveals in-frame reads that prevail over the other frames, though the contrast is less pronounced than for high-FIR iORFs (Fig. 4E). These iORFs precisely recall the hundred CDSs observed in the same Ribo-Seq landscape region that were translated with a detectable protein product despite their intermediate translation specificity (Fig. 3AB). The remaining area corresponds to low-FIR ORFs, with intermediate fractions of F1 and F2 reads that mostly overlap intermediate-FIR iORFs ( $0.33 \leq [F1, F2] < 0.66$ ) (Additional file 1: Fig. S6). Similar to intermediate-FIR iORFs, their metagene profile is less contrasted than that of the other



**Fig. 5** Robustness of Ribo-Seq signals across individual datasets. **A** Distributions of the fractions of datasets in which the ORFs present in the global Ribo-Seq landscape were detected. Distributions for iORFs, aORFs, and CDSs are colored purple, blue, and orange, respectively. **B** Cumulative fractions of the ORFs of the global Ribo-Seq landscape that were detected when cumulating the reads of the different datasets. Fractions for iORFs, aORFs, and CDSs are colored purple, blue, and orange, respectively. Datasets are ranked according to their number of reads. **C** Distributions of the Euclidean distances calculated between the 2D coordinates of an ORF in the individual datasets

low-FIR iORFs (Fig. 4C). We also refer to this region as the twilight zone. Finally, our MS experiments enabled us to detect 30 iORFs considering both standard and inhibited proteasome conditions, among which 10 were detected under both conditions (“Methods”, Fig. 4A and Additional file 1: Table S2). Interestingly, these iORFs include high-, intermediate-, but also low-FIR ORFs. They tend to be longer and associated with more reads than their counterparts. However, the trend does not hold when normalizing by the ORF length, possibly due to the lack of statistical power (Additional file 1: Fig. S7). Again, these iORFs are not more conserved among yeast species (one-proportion  $z$ -test for every age group, all  $P > 0.05$ ), suggesting that the majority of these products may not be functional but are probably linked to specific properties (i.e., physico-chemical and/or abundances) that enabled their detection.

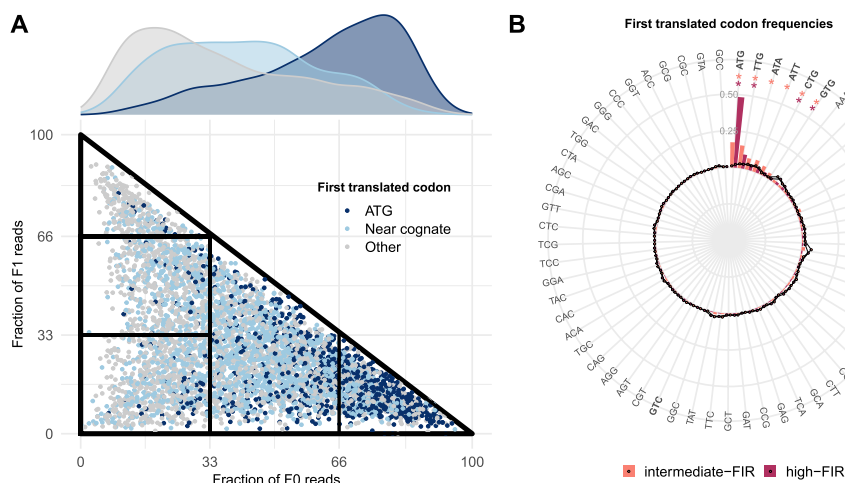
#### Robustness of Ribo-Seq signals across Ribo-Seq datasets

We investigated the contribution of each dataset to the global Ribo-Seq landscape. Figure 5A shows the distribution of the fractions of datasets in which the CDSs, aORFs, and iORFs of the global Ribo-Seq landscape were detected. For each dataset, an ORF was retained if it was associated with at least 20 reads, regardless of the frame (Additional file 1: Fig. S8). If a threshold of 50 reads may be too strict to identify lowly translated ORFs, 20 reads should ensure enough statistical power to estimate fractions of in-frame reads reliably. Most datasets display high coverage for the CDSs present in the global Ribo-Seq landscape, each detecting, on average, 80.2% of them (Fig. 5A). In contrast, this coverage drops significantly to 26.4% and 6.8% for aORFs and iORFs, respectively. In fact, only  $11 \times 10^6$  reads are necessary to detect 90% of the CDSs of the global Ribo-Seq landscape, while it requires at least  $107 \times 10^6$  and  $163 \times 10^6$  reads to cover 90% of the

global Ribo-Seq landscapes of aORFs and iORFs, respectively (Fig. 5B). This emphasizes the importance of pooling multiple datasets to characterize pervasive translation events typically associated with low read levels. Since ORFs are detected in individual samples based on their associated reads, irrespective of the frame, aORFs provide a means to partially deconvolute the impact of ORF length on their detection. Indeed, although the latter are associated with the reads of the overlapping CDS, they are detected in far fewer datasets than CDSs. Their lower detection, therefore, likely results from their shorter size, suggesting that the detection of iORFs in individual datasets is significantly underestimated. Indeed, iORFs are typically short and lowly expressed, affecting their detection in individual samples, particularly when the sequencing coverage of the experiment is low (Fig. 5B, Additional file 1: Fig. S8). Altogether, these findings show that while using multiple Ribo-Seq datasets is relevant for qualitative analyses aimed at uncovering the diversity of Ribo-Seq signals, caution is warranted for accurate estimation of the frequency of translation of lowly expressed short noncoding ORFs across datasets. Finally, we examined the robustness of the ORF localization across the individual samples (Fig. 5C). As expected, the localizations of CDSs and aORFs are highly conserved across the datasets, reflecting the strong optimization of CDSs' translation to ensure the production of the coding product (all Euclidean distance median comparison empirical  $P$ -values  $< 1e-04$ ). Interestingly, the localization of iORFs is also not random, being significantly conserved across the experiments compared to what would be expected by chance (gray distribution) (Euclidean distance median comparison empirical  $P$ -values  $< 1e-04$ ). Overall, these results suggest that, even if associated with pervasive expression, the Ribo-Seq signals of iORFs are consistent under standard conditions.

#### Rules determining the translation signal in pervasive translation

To elucidate the rules that determine the Ribo-Seq signal of iORFs and assess whether iORFs translated with high specificity indicate emerging functional ORFs, we examined several sequence and structural properties previously identified as relevant signatures of de novo gene emergence [21, 42, 54]. iORFs translated with high specificity are not more conserved across neighboring species and do not distinguish themselves according to their length, Kozak scores, or foldability potential (Additional file 1: Fig. S9). However, they display slightly higher GC content and are slightly more abundant outside UTRs than intermediate-FIR ones, but the effect is small and likely does not account for the overall increase in their translation specificity (median comparison empirical  $P$ -values  $< 1e-04$ , one proportion  $z$ -test (two-sided),  $P = 7.3e-14$ , respectively—Additional file 1: Fig. S9). We then undertook dN/dS analyses and identified only three iORFs with significant evidence of purifying selection, similar to the findings reported by Wacholder et al. [22]. While two of them are classified as high-FIR, this result overall suggests that high-FIR iORFs probably do not reflect emerging functional ORFs or are too young to bear a detectable selection signature. We then investigated whether the sequence composition or genomic context could affect the Ribo-Seq patterns of iORFs. Interestingly, while the different iORF categories display comparable amino acid compositions, high-FIR iORFs are significantly enriched in methionine (Additional file 1: Fig. S10). Methionine being an efficient translation start, we sought to identify iORF translation initiation codons.



**Fig. 6** Impact of the initiation codon on the Ribo-Seq signals. **A** Ribo-Seq landscape of the iORFs for which we detected the translation initiation codon. iORFs are colored according to the type of the predicted start codon with dark blue, light blue, and gray dots corresponding to initiation with ATG, near cognate codons, or other codons, respectively. Densities of the three different types of codons with respect to the fraction of in-frame reads are represented on the top of the plot. **B** Barplots representing the frequencies of the 61 codons at the predicted start codon for intermediate-FIR (salmon) and high-FIR iORFs (dark red). The small circles indicate the expected frequency of each codon according to its frequency in the intermediate- or high FIR iORFs. Codons that are significantly enriched at the predicted start codon are indicated with a star (one proportion z-test,  $P < 0.05$ ). Near-cognate codons are represented in bold

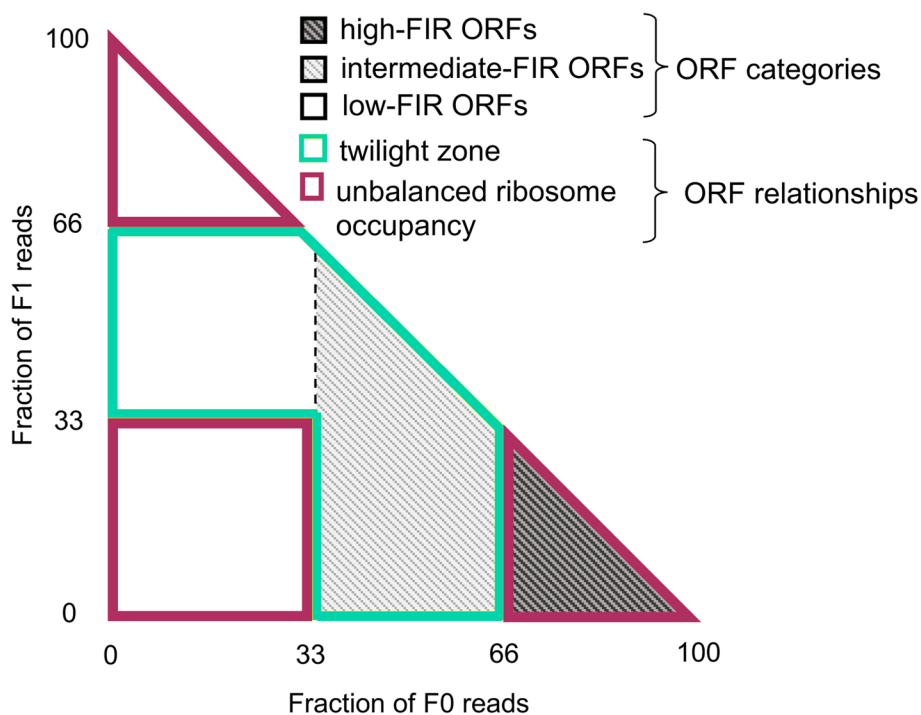
We predicted a start for nearly all high- and intermediate-FIR iORFs (98% and 94%, respectively), while this was the case for only 55% of the low-FIR iORFs. This further supports that although low-FIR ORFs may be associated with translation outcomes, a significant portion likely consists of untranslated ORFs overlapping translated ones. Overall, high-FIR ORFs exhibit a strong tendency to initiate with ATG or TTG, consistent with previous reports [69–71] (49% and 10%, respectively) (Fig. 6AB). While intermediate-FIR iORFs mostly initiate with ATGs or near-cognate codons (62%), they display a wider diversity of start codons. Furthermore, a significant portion (32%) is predicted not to initiate with ATG or near-cognate codons. These cases may reflect noisy Ribo-Seq patterns not associated with translation, but they could also indicate errors in translation start prediction. These ORFs are associated with fewer reads than their counterparts, which may challenge the accurate prediction of the start codon (Additional file 1: Fig. S11). Finally, to further depict the impact of the genomic context of the start codon on translation specificity, we examined the codon content of the alternative frames of intermediate- and high-FIR iORFs. For each intermediate- or high-FIR iORF, we calculated the propensity of each codon to be in the iORF frame compared to its two alternative frames (see the “Methods” section). Additional file 1: Fig. S12 shows that high-FIR iORFs are enriched in ATG and TTG codons with respect to their overlapping frames, while the effect is less pronounced for intermediate-FIR iORFs. These results suggest that in noncoding regions, translation specificity results from the uneven distribution of ATGs across the different frames. We may hypothesize that during the scanning process of the 40S ribosomal subunit along the RNA, the three RNA frames are inherently in competition for translation, with iORFs enriched in ATGs compared to their alternative frames being more likely to be translated. Whether additional features

(e.g., RNA secondary structure, tRNA abundances...) are involved in the translation outcome of iORFs deserves further investigation.

## Discussion

In this study, we characterized the Ribo-Seq landscape of all yeast ORFs of ribosome-bound RNAs, revealing a continuum from non-translation outcomes to highly specific translation. This continuum encompasses a wide diversity of Ribo-Seq signals, including canonical and non-canonical ones. Interestingly, our proteomic analysis seems to indicate that non-canonical Ribo-Seq signals could also lead to a protein product. Despite being associated with higher read levels and highly translated CDSs, aORFs detected with proteomics exhibited Ribo-Seq patterns similar to those of their untranslated counterparts (Additional file 1: Fig. S13). While our work highlights the challenge of distinguishing untranslated ORFs lying on ribosome-associated RNAs from those translated with low specificity, it underscores the importance of considering the entirety of Ribo-Seq signals. Precisely, we also considered ORFs with a priori no translation outcome under the studied conditions (most low-FIR ORFs), but which were nevertheless associated with ribosome-protected fragments. Indeed, their translation status may be unstable, as iORFs not translated today under standard conditions may become translated under other conditions or in the future. Moreover, it is worth noting that the population of untranslated iORFs is heterogeneous since untranslated iORFs located in ribosome-bound RNAs are more likely to be translated later or under other conditions than those lying in RNAs that do not access the translation machinery, or worse, those that are not transcribed. In fact, we may hypothesize that if low-FIR iORFs may have lost the competition for translation (if any) with their overlapping iORFs, they still participate in this competition and, therefore, may play a “passive” role in the translation outcome of the region. All these reasons support the need for characterizing the Ribo-Seq signals of all ORFs associated with ribosome-protected fragments (i.e., which access to the translation machinery) to delineate the translome but also the translation potential of a species.

We introduce a topology of the Ribo-Seq landscape of pervasive translation that enables the rational delineation and classification of the entire diversity of associated Ribo-Seq signals (Fig. 7). In particular, we unveiled a twilight zone that is highly populated in noncoding regions and appears to be a hallmark of pervasive translation. Our results suggest that an important fraction of the ORFs of this region are translated as supported by their metagenes and the MS data. We also highlighted three Ribo-Seq landscape areas of uneven ribosome occupancy reflecting RNA regions associated with high translation specificity. These areas are populated in both coding and noncoding regions. However, their status in these regions is not the same. While the translation specificity of CDSs mostly results from their regulation to ensure successful protein synthesis, the specificity of translation in noncoding regions is mainly due to the codon used to initiate translation and its surrounding genomic context. Consequently, if the translation status of CDSs is expected to be long-lasting throughout evolution, the one of iORFs, even if specifically translated, is expected to be less stable. Upon mutation, high-FIR ORFs may become poorly translated or untranslated and vice-versa. In other words, the specific translation of noncoding regions is a property that is “innate” and probably short-lived, contrary to the “acquired” and stable translation specificity of CDSs.



**Fig. 7** Topology of the yeast Ribo-Seq landscape. Schematic representation of the entire diversity of Ribo-Seq signals that allows (i) their comprehensive and comprehensible classification and (ii) the identification of areas that are either specific to pervasive translation or present in both pervasive and regular translation

Indeed, our findings suggest that, overall, the high translation specificity of an iORF does not result from the translation optimization of a functional product. Specifically translated iORFs are not characterized by features usually indicative of the emergence of function but instead display significant enrichments in ATGs and TTGs relative to their alternative frames. Whether iORFs enriched in ATGs compared to their overlapping counterparts are more likely to reach the protein state and, therefore, to be fixed as novel coding products deserve further investigation. Indeed, our MS data seem to indicate that a microprotein could be detected for aORFs and iORFs associated with different Ribo-Seq landscape areas, highlighting the complex relationship between the specificity of translation and the likelihood of being detected with MS. Moreover, the aORFs that we were able to detect with our MS protocol displayed higher read levels regardless of their translation specificity (Fig. 3, Additional file 1: Fig. S4), suggesting, this time, a complex relationship between the quantity and the specificity of translation. Interestingly, the comparison of different MS datasets suggested a high degree of complementarity, each uncovering new aORF detections. The detected aORFs shared common properties in terms of Ribo-Seq features, being associated with higher read levels and overlapping highly expressed CDSs. This finding suggests a homogeneous population of volatile ORFs whose expression may be inconsistent across experiments, thereby affecting their identification from one experiment to another. Unlike CDSs, inhibiting the proteasome was not accompanied by an increase in non-canonical ORF detection. This raises questions about whether this reflects detection difficulty or uncovers an



alternative surveillance pathway. Overall, it is important to note that confidently asserting the detection of non-canonical ORFs with proteomics remains very challenging. However, the fact that the aORFs we detected display specific Ribo-Seq patterns, which are derived from a technique independent of proteomics, can increase our confidence in their detection. Furthermore, a significant portion of aORFs was retrieved under both conditions, with each condition providing several dozen additional ORFs with comparable properties. This further supports their identification, but it also suggests that the number of these volatile ORFs may be significantly underestimated. Finally, these aORFs are indistinguishable from the others according to sequence or structural properties typical of coding products, suggesting that most of them are not functional but simply result from the high translation activity of their overlapping CDS (Fig. 3). In fact, leading to a detectable product differs from being functional and selected, though the latter necessarily involves being produced at the protein level. Overall, these results show that the relationship between the expression features of an ORF (quantity and specificity), its capacity to reach the protein state, and its potential to be fixed remains unclear and deserves to be carefully investigated.

## Conclusion

The fact that iORFs translated with high specificity, or aORFs and iORFs detected with proteomics, are indistinguishable from others, according to features indicative of the emergence of function, suggests that most pervasively translated products are not functional (at least not in the traditional sense, i.e., not under selection). Nevertheless, we acknowledge that assessing whether an ORF is functional is experimentally and computationally challenging. The former involves identifying the condition under which the ORF affects the phenotype. The latter implies that its function and the conditions under which it is functional persist sufficiently during evolution that we can detect evolutionary traces of selection. Expressly, we do not exclude that our data include young emerging ORFs that do not exhibit traces of selection yet or short-lived ORFs that play a biological role in specific or ephemeral conditions like the evolutionarily transient sequences reported in Wacholder et al. (2023) [22]. Here, we propose that iORFs and pervasive translation are functional “collectively,” providing the cell with the raw material for selection. Of these, from time to time, a handful of functional products (i.e., that will last throughout evolution) may emerge, similar to the dozen candidates under selection identified in Wacholder et al. (2023) [22]. This collective function, which is mainly related to a process rather than individual ORFs, precisely relies on the quantity and diversity of the reservoir of freely evolving sequences that can be translated by pervasive translation. In line with Guerra-Almeida and Nunes-da-Fonseca (2020) [5], these results call for revisiting the concept of function in the context of the emergence of novel coding products. The landscape and diversity of pervasive translation under stress conditions or in tissues associated with the emergence of genetic novelty (e.g., testis or brain [72–75]), therefore deserves to be carefully investigated. Our representation framework precisely enables large-scale and quantitative analyses of the dynamics of the Ribo-Seq landscape of a species of interest upon condition changes or mutations and, hence, opens the way for future exciting comparative analyses.

## Methods

### Extraction of CDSs, aORFs, and noncoding ORFs

All ORFs were extracted with our program ORFtrack [76] from *S. cerevisiae* strain S288C, based on the genome annotation of the Saccharomyces Genome Database (genome version R64-2-1) [77]. iORFs and aORFs are defined from STOP-to-STOP with a minimum size of 60 nucleotides. While relevant candidates may exist within smaller ORFs, very short sequences can lead to inaccuracies in Ribo-Seq or sequence and structure analyses since most analysis programs used in this study are designed for longer sequences. iORFs do not overlap any annotated feature (e.g., tRNA, rRNA, CDS) by more than 5% of their length. As such, we minimize the overlap between iORFs and CDSs which is essential to prevent any bias from CDSs when analyzing their different properties.

### Ribosome profiling analyses

From the Sequence Read Archive [78], we manually collected 89 Ribo-Seq datasets of wild-type *S. cerevisiae* (strain S288C or BY4741) that were conducted under standard conditions (see Additional file 2 for the complete list—i.e., 104 technical replicates from 89 biological datasets). Ribo-Seq raw data were processed using our own pipeline ORFribo from our ORFmine suite (Additional file 1: Fig. S1). Adapter sequences were removed using cutadapt [79]. The trimmed reads were separated into ribosome profiling footprints (RPFs) ranging from 25 to 35 nucleotides and mapped to the CDSs using both HISAT2 [80] and Bowtie2 [81] (default parameters, with only one mismatch being allowed). While read lengths of 28–30 nucleotides generally ensure efficient P-site identification, the optimal read length may vary with the sample and nuclease digestion method. Considering alternative read lengths can provide more comprehensive information. However, beyond the 25–35 nucleotide range, confidence in ribosome activity decreases, making data more uncertain. Furthermore, Bowtie2 typically detects a small portion of additional reads that are not identified by HISAT2 (1.6% of all the considered reads in the present study, but this value depends on the dataset and can be higher). Given that Bowtie2 is not very computationally expensive, we deemed it worthwhile to consider this additional information. For each dataset, we estimated the ribosome's P-site of each RPF length using riboWaltz [82] and subsequently calculated the number of in-frame and out-of-frame reads of each CDS. For each dataset, only the RPF lengths with a median of CDSs' in-frame reads exceeding 70% were retained [64] (Supplementary Data 1 and Additional file 1: Fig. S14). In doing so, we retained 54 datasets where at least one RPF length met this condition. All retained RPF lengths from the different datasets were then pooled together and realigned on the whole genome this time. Then, for each read, the translated codon was estimated according to the previously predicted P-site for that read length (RPF length) in the dataset it originated from. The number of in-frame and out-of-frame reads was subsequently computed for each iORF and aORF.

### Calculations of metagenes

Metagenes were calculated for different subsets of ORFs (e.g., intermediate-, high-FIR ORFs...) as follows: for each ORF of a given subset, we summed over the 54 Ribo-Seq

datasets, the number of reads associated with each nucleotide (i.e., reads whose predicted P-site corresponds to the position of the considered nucleotide). We then averaged the count of reads per nucleotide position over all the ORFs of a given subset. The resulting metagene, therefore, represents the read number average per nucleotide.

### Calculation of sequence and structural properties of the peptides encoded in noncoding ORFs

The HCA foldability score was calculated using the pyHCA tool [83–86] from our program ORFold [76]. We estimated the evolutionary ages of the CDSs by phylostratigraphy using our ORFdate program. Half-life data was extracted from the SGD database [77]. Thus, BLASTp [87] searches with the CDSs of *S. cerevisiae* (i.e., the focal species) were performed against the complete proteomes of 10 saccharomyces (see Additional file 1: Fig. S15 for the complete list and the associated phylogenetic tree). We retained BLASTp high scoring pairs (HSPs) with an e-value less than 0.001 and a minimum query coverage of 70%. For each CDS, we identified the most distant species with respect to the focal that was associated with a positive HSP match. Since horizontal gene transfers are rare events in Eukaryotes, we defined the CDS origination node, the last node shared by the focal and the outgroup species with a positive HSP. This node is then used to estimate the CDS's age according to TimeTree [88, 89] (Additional file 1: Fig. S15). As we are interested in the early ages of CDSs, the node shared by *S. cerevisiae* and *S. pombe* is considered as the upper limit for the age estimation, and all CDSs with a match in *S. pombe* are associated with the same upper bounded age regardless of the fact that the CDS could have additional matches with more distant species. Since iORFs evolve faster than CDSs, we estimated their age, focusing on the Saccharomyces sensu stricto (Additional file 1: Fig. S15). BLASTp were performed with the iORFs of *S. cerevisiae* (focal) as queries and those of the Saccharomyces sensu stricto species as targets using the same parameters. The remaining steps for age estimation were the same as those used for CDSs. Kozak scores were calculated as follows: CDSs were first used to derive a reference scoring table of position-specific nucleotide frequencies. To do so, we computed for each position around the start codon (positions –5 to 6 excluding the start codon) the nucleotide frequencies and calculated the odd ratios between the observed and background frequencies. The latter are calculated as the position-independent relative frequencies of the same sequences. We then calculated the Kozak score of each evaluated iORF by adding the individual position-specific values of all nucleotides observed for each position. To detect iORFs under selection, we defined their orthologous sequences among the neighbor species using the RBH method (aligner:Diamond v2 [90], e-value  $\leq 10^{-3}$ , query coverage  $\geq 70\%$ ) and ran multiple CODEML branch models (from the PAML package) [91]. We followed the protocol of Zhang (2019) [55] although with a distinct subset of trees for the two-ratios models. This subset was populated with all possible tree versions where a single branch or a single clade is labeled. The label indicates a distinct ratio from the rest of the tree. We corrected the chi-squared test *p*-values with the Bonferroni method and applied a significance threshold of 0.05.

The propensity of each codon in iORFs with respect to their alternative frames is defined as the log ratio of the codon frequency in the frame of the ORF versus its

frequency in the ORF's two alternative frames ( $\pm 30$  nucleotides around the considered ORF). It is calculated as follows:

$$\log_{10} \left( \frac{\text{freq}(\text{codon}_i)_{\text{in-frame}}}{\text{freq}(\text{codon}_i)_{\text{out-of-frame}}} \right)$$

### Prediction of the first translated codon

The translation start codon typically exhibits a distinct peak of reads. Since in noncoding regions the number of reads per codon is generally low, we defined as the start codon, the first 5' codon with a peak of at least 5 in-frame reads and for which the two surrounding codons had a lower number of reads.

### Statistical analyses

All statistical analyses were performed in R (4.0.3) [92]. To address *P*-value issues with large samples [93], we empirically estimate a global *p*-value for median comparisons in large samples using an iterative sampling strategy. This method involves repeatedly sampling from the data to calculate the proportion of instances where the median differences from these random samples exceeded the observed median difference. The procedure was applied when one group exceeded 1000 individuals as follows: we computed the observed difference (Dobs) between the medians of the two compared groups. For groups larger than 1000, we generated 10,000 samples of 1000 individuals each and pooled both groups. If one group had fewer than 1000 individuals, to overcome the imbalance between the two groups, we used the entire group for the smallest one and generated 10,000 samples from the larger group, matching the size of the smallest one. We then divided the pooled sample into two equal subsamples, calculated the difference (Dsamp) between sample medians, and derived the *P*-value as the proportion of Dsamp values exceeding Dobs.

### Mass spectrometry experiments

Yeast strain BY4742 was grown under two different conditions, either in complete synthetic media or in complete synthetic media with MG132 proteasome inhibitor. Cells were lysed in an extraction buffer containing 20 mM HEPES, 110 mM KOAc, 10 mM MnCl<sub>2</sub>, 0.5% triton, 0.1% tween and protease inhibitors. Lysates were cleared by centrifugation and loaded on a gradient of 4% to 12% SDS-PAGE. Proteins were extracted from the gels and analyzed by mass spectrometry. For each condition, proteins were concentrated in a single band containing the whole sample or fractionated into 3 different mass regions (1–15 kDa, 15–35 kDa, and > 35 kDa). The bands of about 2 mm were subjected to in-gel trypsin digestion, as previously described in Szabo et al. [94], before submission to mass spectrometry analysis. Trypsin-generated peptides from the three separation regions of the gels were analyzed separately by nanoLC–MSMS using a nanoElute liquid chromatography system (Bruker) coupled to a timsTOF Pro mass spectrometer (Bruker). Peptides were loaded on an Aurora analytical column (ION OPTIK, 25 cm × 75 μm, C18, 1.6 μm) and separated with a gradient of 0–35% of solvent B for 100 min. Solvent A was 0.1% formic acid and 2% acetonitrile in water, and solvent B was acetonitrile with 0.1% formic acid. MS and MS/MS spectra were recorded from *m/z* 100 to 1700 with

a mobility scan range from 0.6 to 1.4 V s/cm<sup>2</sup>. MS/MS spectra were acquired with the PASEF (Parallel Accumulation Serial Fragmentation) ion mobility-based acquisition mode using a number of PASEF MS/MS scans set as 10. MS and MSMS raw data were processed and converted into mgf files with DataAnalysis software (Bruker).

### MS statistical validation of peptides resulting from iORF translation

We first applied an established standard protocol based on class-specific FDRs. To do so, the spectra were searched using the MASCOT search engine (Matrix Science, London, UK, [95]) against the CDS, aORF, and iORF databases separately. Since we have no certainty about the start codon, aORFs, and iORFs were provided as STOP-to-STOP, ATG-STOP, and predicted START-to-STOP when different from the most 5' ATG. Database searches were performed using trypsin cleavage specificity with two possible missed cleavages. Carbamidomethylation of cysteines was set as a fixed modification, and oxidation of methionine as a variable modification. Peptide and fragment tolerances were set at 10 ppm and 0.05 Da, respectively. The procedure was applied to a total of four proteomic runs: two conditions (standard and inhibited proteasome), each associated with two SDS-PAGE gel migration modes. Doing so, no candidates could be detected at 1% FDR. We were able to detect a peptide for only a few non-canonical ORFs (19 aORFs and 5 iORFs) at the cost of substantially increasing local FDRs (Additional file 1: Table S3 for the number of detected ORFs in each migration and the corresponding FDRs). In fact, it is well-acknowledged that with large databases, strict filtering to avoid false-positive matches leads (i) to more false negatives and (ii) overestimating FDRs, as well-explained in Verbruggen et al. (2021) [65]. The issue of dealing with large databases in proteomics is a complex and active research area, and numerous studies have highlighted the difficulty of estimating FDRs when searching large databases [65, 66, 96, 97]. The task becomes even more complicated when attempting to identify entities representing only a tiny portion of the searched database. Specifically, in our configuration, we are searching an extensive database that includes all non-canonical ORFs that match our Ribo-Seq criteria without prior knowledge of whether they are indeed translated. We thus anticipate only a small portion of positives (i.e., translated microproteins abundant and stable enough to be detected) since the majority is not expected to be translated or, when translated, expected to be unstable and short-lived. This situation, therefore, differs from the classical one for which TDC has been initially proposed, where the portion of positives in the searched database is usually high compared to the matches in the decoy database. In our configuration, FDRs are likely to be overestimated because of (i) the large dataset that will inherently be associated with a large number of decoy matches and (ii) the relatively small number of positives in non-canonical ORFs. We, therefore, devised a two-step protocol that first searches a large database combining all the CDSs and non-canonical ORFs that match our Ribo-Seq criteria (see below) without any a priori. Then, a second search is realized on a much smaller database, resulting from the first screen, which places us in a more classical search context where more reliable FDRs can be estimated. Each step combines two FDR-controlling procedures: Target-Decoy Competition (TDC) [98] and the Benjamini–Hochberg framework (BH) [99].

*First step:* the spectra were searched using the MASCOT search engine against the coding and noncoding databases. All CDSs, aORFs, and iORFs with at least 50 reads

and 30% read coverage were combined into a single Target database to avoid assigning a spectrum twice and mistakenly assigning a peptide to a non-canonical ORF when it actually comes from a CDS. Database searches were performed using the same parameters as the standard protocol (see above). We then applied separate TDC and BH procedures. *TDC*: besides screening the target database, we searched a database of decoys obtained by reversing each target sequence. Peptide Spectrum Matches (PSM) mascot scores were converted into  $p$ -values ( $p = 10^{-S/10}$ ). We retained all PSMs with a  $p$ -value  $\leq 0.05$  and then retained the PSM with the higher score between the best-scoring target peptide and the best-scoring decoy peptide. In case of ties, the PSM was excluded. *BH*: the total non-filtered PSMs were exported, and their  $p$ -values were adjusted using the BH procedure. PSMs were then filtered at 2.5% FDR. Finally, only PSMs validated by both methods were retained. This approach enabled us to exclude PSMs that the single BH procedure would retain, even though their associated spectrum scored better with the decoy database. A list of candidate ORFs was thus derived, requiring at least two distinct peptides for CDSs and one for non-canonical ORFs due to their small size. Raw data generated by MASCOT (without any filter) and our filtered final table are available as Supplemental Material.

*Second step*: the proteome of *S. cerevisiae* was supplemented with the aORFs and iORFs retained after the first search (270 and 35 ORFs, respectively), and the spectra were then screened against the expanded proteome using the same parameters as previously. It should be noted that the  $p$ -values depend on the searched database, and PSMs retained during the first step may be excluded in the second. TDC was applied to all PSMs with a  $p$ -value  $\leq 0.05$ , resulting in FDRs of approximately 1% in each of the four proteomic analyses (Additional file 1: Table S4). It is worth noting that the overall FDR is anticipated to be higher; however, combining TDC and BH procedures is expected to reduce FDRs further. Similar to step 1, PSM  $p$ -values were adjusted with the BH procedure, and PSMs were filtered at 2.5% FDR. The BH filtering appeared very stringent, removing 35% of the PSMs of the non-canonical ORFs retained with the TDC procedure alone.

As a control, we represented the scores of the retained PSMs for CDSs, aORFs, and iORFs after TDC or the combination of both TDC and BH procedures alongside those of the corresponding PSMs in the decoy database (Additional file 1: Fig. S16). We then represented the distribution of the difference between the scores of the retained PSMs for each ORF class and those of their corresponding best-scoring PSMs in the decoy database. While it was anticipated that target PSM scores would be higher than those of their corresponding decoys—given our filtering—the extent of the difference was not expected. Target PSM scores for non-canonical ORFs are clearly higher, generally about twice as high as those for their corresponding decoys, revealing a clear separation between their scores in the target and decoy databases. Additional file 1: Fig. S17 presents 30 spectra of different MASCOT score ranges.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03403-7>.

Additional file 1: Supplementary tables S1-S4; Supplementary figures S1-S17 [104–106].

Additional file 2: Datasets used in this study, with accession numbers.

Additional file 3: Review history.

**Acknowledgements**

We thank J Armengaud, JM Camadro, G Chevreux, and V Redeker for the fruitful discussions on the proteomic analyses. We thank M Gallopin and MH Mucchielli for the fruitful discussions on the statistical analyses. We thank A Baumann and S Herman for the fruitful discussions on the manuscript.

**Peer review information**

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team

**Authors' contributions**

Conceptualization: CP, ON, AL; methodology: CP, HA, NC, SB, DC, ON, AL; investigation and development: CP, HA, NC, SB, DC, PR, CR, SA, AL; writing and editing: CP, NC, SB, PR, OL, ON, AL; supervision: AL.

**Funding**

Works by CP and PR were supported by French government fellowships. HA and SB works were supported by ANR Actimeth (19-CE12-0004-02) and ANR Rescue Ribosome (17-CE12-0024). This work has benefited from the facilities and expertise of the I2BC proteomic platform (Proteomic-Gif, SiCaPS) supported by IBiSA, Ile de France Region, Plan Cancer, CNRS, and Paris-Saclay University.

**Data availability**

The supplementary Figures and Tables are available as Additional file 1 on Zenodo and are published under the MIT license: <https://doi.org/10.5281/zenodo.13734541> [100]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [101] partner repository with the dataset identifier PXD040766 and are available at <https://www.ebi.ac.uk/pride/archive/projects/PXD040766> [102]. The scripts and the associated raw and calculated data used in this study are available on Zenodo: <https://doi.org/10.5281/zenodo.13734541> [100] and are published under the MIT license. The 89 Ribo-Seq datasets can be downloaded from the Sequence Read Archive database [78] (the Accession codes for every dataset are provided as Additional file 2 on Zenodo: <https://doi.org/10.5281/zenodo.13734541> [100]). The extraction of all ORFs (CDSs, aORFs and iORFs), the analysis of their sequence and structural properties (foldability potential, ORF conservation) along with their translation activity were calculated using our programs (ORFtrack, ORFold, ORFdate and ORFribi) available in the ORFmine package (v0.8.7) published under the MIT license at: <https://github.com/i2bc/ORFmine> [103].

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 11 September 2023 Accepted: 26 September 2024

Published online: 14 October 2024

**References**

- Hanada K, Zhang X, Borevitz JO, Li W-H, Shiu S-H. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res.* 2007;17:632–40.
- Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.* 2011;21:634–41.
- Couso J-P, Patraquim P. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol.* 2017;18:575–89.
- Orr MW, Mao Y, Storz G, Qian S-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.* 2020;48:1029–42.
- Guerra-Almeida D, Nunes-da-Fonseca R. Small open reading frames: how important are they for molecular evolution? *Front Genet.* 2020;11: 574737.
- Guerra-Almeida D, Tschoeke DA, Nunes-da-Fonseca R. Understanding small ORF diversity through a comprehensive transcription feature classification. *DNA Res.* 2021;28: dsab007.
- Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science.* 2002;296:916–9.
- Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science.* 2009;324:218–23.
- Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. *PLoS Biol.* 2011;9: e1000625.
- Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. *Nature.* 2012;487:370–4.
- Jensen TH, Jacquier A, Libri D. Dealing with pervasive transcription. *Mol Cell.* 2013;52:473–84.

12. Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development*. 2013;140:2828–34.
13. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife*. 2014;3: e03528.
14. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*. 2014;33:981–93.
15. Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, et al. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep*. 2014;7:1858–66.
16. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 2014;8:1365–79.
17. Ruiz-Orera J, Verdaguier-Grau P, Villanueva-Cañas J, Messeguer X, Albà MM. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol*. 2018;2:890–6.
18. Ruiz-Orera J, Albà MM. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet*. 2019;35:186–98.
19. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of non-canonical human open reading frames. *Science*. 2020;367:1140–6.
20. Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, et al. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat Commun*. 2021;12:1–13.
21. Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, et al. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res*. 2021;31:2303–15.
22. Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, et al. A vast evolutionarily transient translome contributes to phenotype and fitness. *Cell Syst*. 2023;14:363–381.e8.
23. Smith C, Canestrari JG, Wang AJ, Champion MM, Derbyshire KM, Gray TA, et al. Pervasive translation in *Mycobacterium tuberculosis*. Kana BD, Shell S, editors. *eLife*. 2022;11:e73980.
24. Parikh SB, Houghton C, Van Oss SB, Wacholder A, Carvunis A. Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast*. 2022;39:471–81.
25. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. *Nat Biotechnol*. 2022;40:994–9.
26. Patraquim P, Magny EG, Pueyo JJ, Platero AI, Couso JP. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nat Commun*. 2022;13:6515.
27. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat Chem Biol*. 2013;9:59.
28. Prbakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, Dittrich C, et al. Quantitative profiling of peptides from RNAs classified as noncoding. *Nat Commun*. 2014;5:1–10.
29. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, et al. Discovery of human sORF–encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res*. 2014;13:1757–65.
30. Hsu PY, Benfey PN. Small but mighty: functional peptides encoded by small ORFs in plants. *Proteomics*. 2018;18:1700038.
31. Cao X, Khitun A, Na Z, Dumitrescu DG, Kubica M, Olatunji E, et al. Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J Proteome Res*. 2020;19:3418–26.
32. Cuevas MVR, Hardy M-P, Holly J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep*. 2021;34:108815.
33. Zheng EB, Zhao L. Protein evidence of unannotated ORFs in *Drosophila* reveals diversity in the evolution and properties of young proteins. Levine MT, Przeworski M, editors. *eLife*. 2022;11:e78772.
34. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The translational landscape of the human heart. *Cell*. 2019;178:242–260.e29.
35. Sandmann C-L, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, et al. Evolutionary origins and interaction of human, young microproteins and small peptides translated from short open reading frames. *Mol Cell*. 2023;83:994–1011.e18.
36. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* Clade. *Genetics*. 2007;176:1131–7.
37. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci*. 2006;103:9935–9.
38. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*. 2008;179:487–96.
39. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, et al. On the origin of new genes in *Drosophila*. *Genome Res*. 2008;18:1446–55.
40. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. *Genome Res*. 2009;19:1752–9.
41. Siepel A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res*. 2009;19:1693–5.
42. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat Rev Genet*. 2011;12:692–702.
43. Wu D-D, Irwin DM, Zhang Y-P. De novo origin of human protein-coding genes. *PLoS Genet*. 2011;7: e1002379.
44. Wissler L, Godmann L, Bornberg-Bauer E. Evolutionary dynamics of simple sequence repeats across long evolutionary time scale in genus *Drosophila*. *Trends in Evolutionary Biology*. 2012;4:e7–e7.
45. Murphy DN, McLysaght A. De novo origin of protein-coding genes in murine rodents. *PLoS ONE*. 2012;7: e48650.
46. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*. 2014;343:769–72.
47. Schlötterer C. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet*. 2015;31:215–9.
48. Bornberg-Bauer E, Schmitz J, Heberlein M. Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. *Biochem Soc Trans*. 2015;43:867–73.
49. Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol Evol*. 2016;8:2190–202.



50. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol.* 2017;1:1–6.
51. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The goddard and saturn genes are essential for *Drosophila* male fertility and may have arisen de novo. *Mol Biol Evol.* 2017;34:1066–82.
52. Vakirlis N, Hebert AS, Oplente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular portrait of de novo genes in yeasts. *Mol Biol Evol.* 2018;35:631–45.
53. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature Ecol Evol.* 2018;2:1626–32.
54. Van Oss SB, Carvunis AR. De novo gene birth. *PLoS Genet.* 2019;15:e1008160.
55. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nature Ecol Evol.* 2019;3:679–90.
56. Prabh N, Rödelsperger C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *pristionchus* nematodes. *G3: Genes, Genomes, Genetics.* 2019;9:2277–86.
57. Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun.* 2020;11:1–18.
58. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife.* 2020;9:e53500.
59. Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. *J Mol Evol.* 2020;88:382–98.
60. Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, et al. Structural and functional characterization of a putative de novo gene in *Drosophila*. *Nat Commun.* 2021;12:1–13.
61. Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol.* 2021;68:175–83.
62. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De Novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* 2013;9: e1003860.
63. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. Perry GH, Weigel D, Perry GH, Menke DB, editors. *eLife.* 2019;8:e44392.
64. Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, et al. What can ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? *Mol Cell Proteomics.* 2023;22. <https://doi.org/10.1016/j.mcpro.2023.100631>.
65. Verbruggen S, Gessulat S, Gabriels R, Matsaroki A, Van de Voorde H, Kuster B, et al. Spectral prediction features as a solution for the search space size problem in proteogenomics. *Mol Cell Proteomics.* 2021;20. <https://doi.org/10.1016/j.mcpro.2021.100076>.
66. Wacholder A, Carvunis A-R. Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biol.* 2023;21: e3002409.
67. Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, et al. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics.* 2012;11:492–500.
68. He C, Jia C, Zhang Y, Xu P. Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. *J Proteome Res.* 2018;17:2335–44.
69. Hinnebusch AG. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol Mol Biol Rev.* 2011;75:434–67.
70. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* 2017;31:1717–31.
71. Cao X, Slavoff SA. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. *Exp Cell Res.* 2020;391: 111973.
72. Gardner LB. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol.* 2008;28:3729–41.
73. Zetoune AB, Fontanière S, Magnin D, Anczuków O, Buisson M, Zhang CX, et al. Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genet.* 2008;9:1–11.
74. Heinen TJ, Staubach F, Häming D, Tautz D. Emergence of a new gene from an intergenic region. *Curr Biol.* 2009;19:1527–31.
75. Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genetics.* 2012;8:e1002942.
76. Papadopoulos C, Chevrollier N, Lopes A. Exploring the peptide potential of genomes. In: Simonson T, editor. *Computational peptide science methods in molecular biology*. New York: Springer; 2022. p. 63–82.
77. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. *Saccharomyces Genome Database: the genomics resource of budding yeast*. *Nucleic Acids Res.* 2012;40:D700–5.
78. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research.* 2011;39:D19–21.
79. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17:10–2.
80. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37:907–15.
81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–9.
82. Lauria F, Tebaldi T, Bernabò P, Groen EJM, Gillingwater TH, Viero G. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Comput Biol.* 2018;14: e1006169.
83. Faure G, Callebaut I. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics.* 2013;29:1726–33.
84. Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput Biol.* 2013;9:e1003280.

85. Bitard-Feildel T, Callebaut I. HCAtk and pyHCA: a toolkit and python API for the hydrophobic cluster analysis of protein sequences. *bioRxiv*. 2018;249995. <https://doi.org/10.1101/249995>.
86. Lamiable A, Bitard-Feildel T, Rebehmed J, Quintus F, Schoentgen F, Mornon J-P, et al. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie*. 2019;167:68–80.
87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
88. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*. 2006;22:2971–2.
89. Kumar S, Suleski M, Craig JM, Kasprowitz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. *Mol Biol Evol*. 2022;39: msac174.
90. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12:59–60.
91. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*. 2007;24:1586–91.
92. Team R Core RC. R: A language and environment for statistical computing. 2020. Available from: <https://www.R-project.org/>.
93. Lin M, Lucas HC Jr, Shmueli G. Research commentary—too big to fail: large samples and the p-value problem. *Inf Syst Res*. 2013;24:906–17.
94. Szabó Á, Papin C, Cornu D, Chélot E, Lipinski Z, Udvardy A, et al. Ubiquitylation dynamics of the clock cell proteome and TIMELESS during a circadian cycle. *Cell Rep*. 2018;23:2273–82.
95. Perkins D, Pappin D, Creasy D, Cottrell J. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20:3551–67.
96. Jouffret V, Miotello G, Culotta K, Ayrault S, Pible O, Armengaud J. Increasing the power of interpretation for soil metaproteomics data. *Microbiome*. 2021;9:195.
97. Fancello L, Burger T. An analysis of proteogenomics and how and when transcriptome-informed reduction of protein databases can enhance eukaryotic proteomics. *Genome Biol*. 2022;23:132.
98. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4:207–14.
99. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (Methodol)*. 1995;57:289–300.
100. Papadopoulos C, Arbes H, Cornu D, Chevrollier N, Blanchet S, Roginski P, et al. The Ribosome Profiling landscape of yeast reveals a high diversity in pervasive translation [Data set]. Zenodo; 2024. <https://doi.org/10.5281/zenodo.13734541>.
101. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res*. 2022;50:D543–52.
102. Cornu D. Diversity in pervasive translation. A new translational landscape of yeast. PRIDE. 2023. <https://doi.org/10.6019/PXD040766>.
103. Papadopoulos C, Chevrollier N, Arbes H, Roginski P, Lopes A. ORFmine. 2023. Available from: <https://github.com/i2bc/ORFmine>.
104. Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global Proteome turnover analyses of the yeasts *S. cerevisiae* and *S. pombe*. *Cell Rep*. 2014;9:1959–65.
105. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol*. 2015;16:179.
106. Fesenko I, Kirov I, Kniazev A, Khazigaleeva R, Lazarev V, Kharlampieva D, et al. Distinct types of short open reading frames are translated in plant cells. *Genome Res*. 2019;29:1464–77.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.