A reference translatome map reveals two modes of protein evolution

Aaron Wacholder[12], Omer Acar[123], and Anne-Ruxandra Carvunis[12*]

1. Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
2. Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
3. Joint CMU-Pitt Ph.D. Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA, 15213, United States

*Corresponding author: **anc201@pitt.edu**

Abstract

Ribosome profiling experiments demonstrate widespread translation of eukaryotic genomes outside of annotated protein-coding genes. However, it is unclear how much of this "noncanonical" translation contributes biologically relevant microproteins rather than insignificant translational noise. Here, we developed an integrative computational framework (iRibo) that leverages hundreds of ribosome profiling experiments to detect signatures of translation with high sensitivity and specificity. We deployed iRibo to construct a reference translatome in the model organism *S. cerevisiae*. We identified ~19,000 noncanonical translated elements outside of the ~5,400 canonical yeast protein-coding genes. Most (65%) of these non-canonical translated elements were located on transcripts annotated as non-coding, or entirely unannotated, while the remainder were located on the 5' and 3' ends of mRNA transcripts. Only 14 non-canonical translated elements were evolutionarily conserved. In stark contrast with canonical protein-coding genes, the great majority of the yeast noncanonical translatome appeared evolutionarily transient and showed no signatures of selection. Yet, we uncovered phenotypes for 53% of a representative subset of evolutionarily transient translated elements. The iRibo framework and reference translatome described here provide a foundation for further investigation of a largely unexplored, but biologically significant, evolutionarily transient translatome.

Introduction

The central role played by protein-coding genes in biological processes has made their identification and characterization an essential project for understanding organismal biology. Over the past decade, the scope of this project has expanded as ribosome profiling (ribo-seq) studies have revealed pervasive translation of eukaryotic genomes.[1,2] These experiments demonstrate that genomes encode not only the "canonical translatome", consisting of the open reading frames (ORFs) identified as coding genes in genome databases like RefSeq[3], but also a large "noncanonical translatome" consisting of coding ORFs that are not annotated as genes. Despite lack of annotation, large-scale studies find that many noncanonical ORFs (nORFs) show evidence of association with phenotypes.[4–6] Additionally, a handful of previously unannotated coding sequences, identified by ribo-seq experiments, have now been characterized in depth, revealing that they play key roles in biological pathways and are important to organism fitness.[7–10] Yet, these well-studied examples represent only a small fraction of the noncanonical translatome. Most noncanonical translation could simply be biologically insignificant "translational noise" resulting from the imperfect specificity of translation processes.[11,12] Alternatively, thousands of missing protein-coding genes could be hidden in the noncanonical translatome.

A common and powerful approach to identifying biologically significant genomic sequences is to look for evidence that the sequence is evolving under selection[13–15]. Many canonical genes were annotated on the basis of such evidence.[16,17] However, in the case of noncanonical translation, this approach is often limited by a lack of sufficient statistical power to confidently detect selection. Many noncanonical translated ORFs are much shorter than canonical genes[5], providing fewer informative variants to use for evolutionary inference. Short coding sequences are sometimes missed by genome-wide evolutionary analyses due to their short length despite long-term evolutionary conservation.[9,18] Power limitations are even more severe for noncanonical ORFs that are evolutionarily novel, as a short evolutionary history also provides less information to distinguish selective from neutral evolution. Several *de novo* genes that evolved recently from noncoding sequences have been discovered from within the noncanonical translatome[19,20].

The challenges in identifying signatures of selection acting on short translated ORFs are compounded by difficulty in establishing unequivocal translation in the first place. Microproteins are often missed by most proteomics techniques, though specialized methods are being developped.[21,22] In ribo-seq data, the most robust evidence of translation comes from a pattern of triplet periodicity in reads across an open reading frame (ORF) corresponding to the progression of the ribosome across codons.[4,23,24]

Translation inference methods have less power to detect translation of short ORFs as they contain fewer positions to use to establish periodicity.[25] The lower expression levels of some noncanonical ORFs further increases the difficulties in identification.[19,26] Perhaps as a result of these power limitations, less than half of the noncanonical ORFs detected as translated in humans are reproducible across studies.[27]

Here, we designed an approach to increase power in detection of both translation and selection among noncanonical ORFs. We address the challenges in detecting translation through the development of an integrative ribo-seq analysis framework (iRibo) that identifies signatures of translation with high sensitivity and high specificity even for sequences that are short or poorly expressed. We address the challenges in detecting selection through a comparative genomics framework that analyzes translated sequences collectively across evolutionary scales within- and between-species.

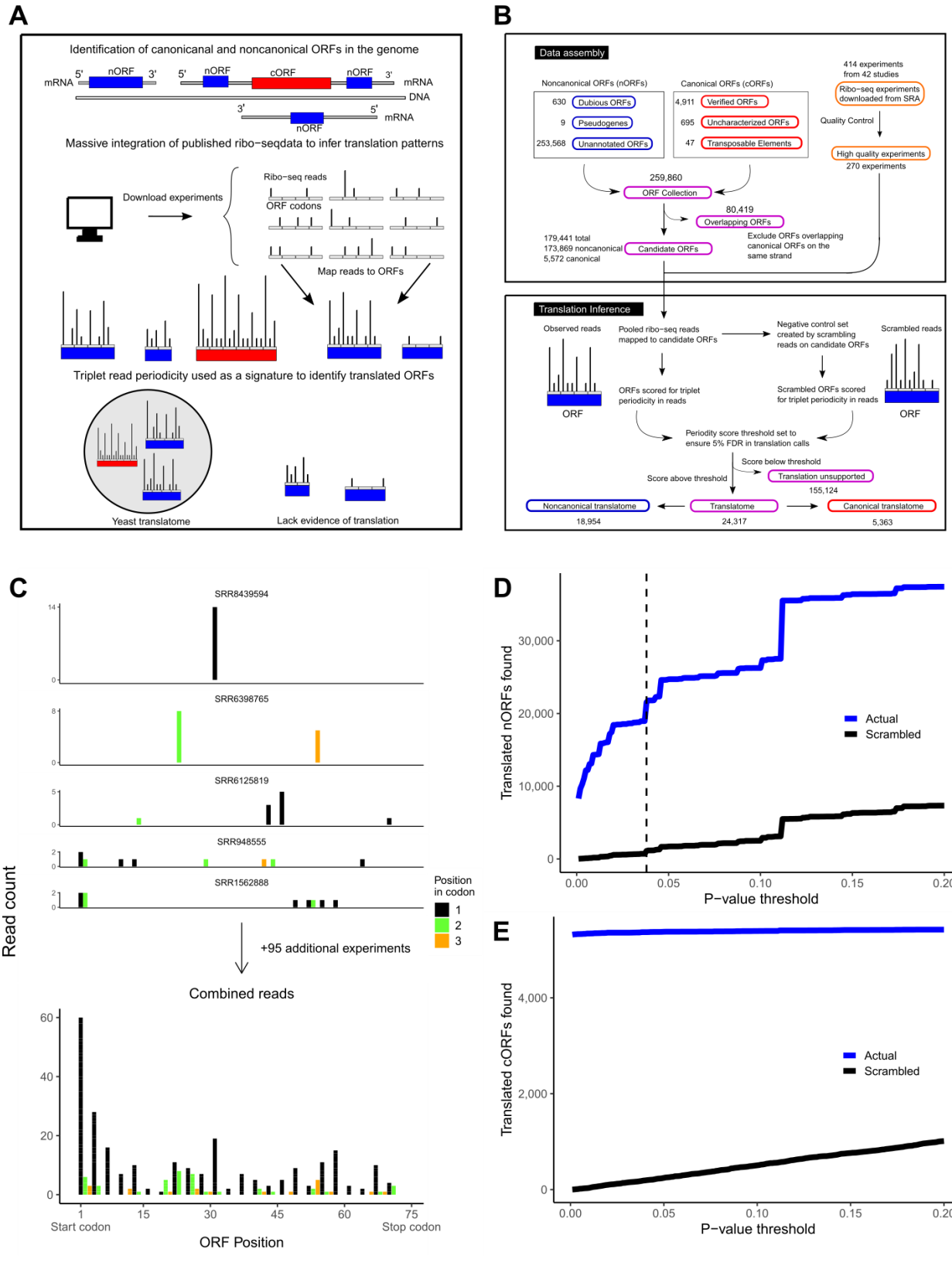
We applied our approach to define a "reference translatome" for the model organism *S. cerevisiae* and to characterize the biological significance of noncanonical translated elements. Using iRibo, we identified ~19,000 noncanonical ORFs translated at high confidence and established the dependence of noncanonical translation on both genomic context and environment condition. Using genomic data at the population level within strains of *S. cerevisiae* and at the species level across the budding yeasts[28,29], we identified a handful of undiscovered conserved genes within the yeast noncanonical translatome. However, the vast majority of the yeast noncanonical translatome consists of evolutionarily transient sequences evolving close to neutrally. Despite lacking signatures of selection, many transient ORFs were associated with phenotypes and cellular pathways. We thus conclude that much of the noncanonical translatome is composed of neither translational noise nor genes in the traditional sense, but rather a distinct class of short-lived coding sequences that play important biological roles.

Results

An integrative approach to defining the translatome

iRibo consists of three components (**Figure 1A;** methods). First, reads from multiple ribo-seq experiments are pooled and mapped to the genome. Second, the translation status of each candidate ORF in the genome is assessed based on the periodicity of ribo-seq reads across the ORF. High-quality ribo-seq data provides single-nucleotide resolution such that reads map to the first position within codons of translated ORFs at much higher frequencies than to the other two positions, generating a pattern of triplet nucleotide periodicity corresponding to the progression of the ribosome codon-by-codon across the transcript. iRibo calls ORFs as translated if they show significant evidence of triplet

89 periodicity in a binomial test. Finally, confidence in the list of ORFs called translated, the identified
90 translatome, is evaluated using an empirical null distribution. A false discovery rate is estimated by
91 assessing triplet periodicity on a dataset generated by shuffling the genomic location of ribo-seq reads
92 across each ORF. iRibo thus defines the translatome with high sensitivity by leveraging the power of
93 integrating multiple ribo-seq experiments, while high specificity is maintained by setting a desired false
94 discovery rate. iRibo can be applied to a set of ribo-seq experiments conducted under a single
95 environmental condition in order to precisely describe translation patterns under that condition.
96 Alternatively, it can be deployed on a broader set of ribo-seq experiments conducted in many different
97 contexts to construct a “reference translatome” consisting of all elements within a genome with
98 sufficient evidence of translation.



99

Figure 1: The iRibo framework enables detection of thousands of noncanonical translated sequences. A) The iRibo framework: both canonical (cORF) and noncanonical (nORFs) are identified in the genome. Reads aggregated from published datasets are then mapped to these ORFs, with translation inferred from triplet periodicity of reads. B) Workflow to identify translated ORFs in the *S. cerevisiae* genome using published datasets. C) Mapped ribo-seq reads across an example nORF on chromosome II. The top five graphs correspond to the individual experiments with the most reads mapping to the ORF, while the bottom graph includes all reads in all experiments. Reads from many distinct experiments are necessary to identify the periodic pattern. D) The number of nORFs found to be translated using the iRibo method at a range of p-value thresholds. Translation calls for a negative control set, constructed by scrambling the actual ribo-seq reads for each nORF, is also plotted. The dashed line signifies a false discovery rate of 5% among nORFs. E) The number of cORFs found to be translated using iRibo at a range of p-value thresholds, contrasted with negative controls constructed by scrambling the ribo-seq reads of each cORF.

We applied iRibo to candidate ORFs across the *S. cerevisiae* genome (**Figure 1B**). The set of candidate ORFs was constructed by first collecting all genomic sequences at least three codons in length that start with ATG and end with a stop codon in the same frame. For ORFs overlapping in the same frame, only the longest ORF was kept. Each candidate ORF can be classified as canonical (cORF) if it is annotated as "verified," "uncharacterized," or "transposable element" in the Saccharomyces Genome Database (SGD) or as noncanonical (nORF) if it is annotated as "dubious," "pseudogene," or is unannotated. We excluded nORFs that overlap cORFs on the same strand. This process generated a list of 179,441 candidate ORFs, of which 173,869 are nORFs and 5,572 cORFs. Translation status for candidate ORFs was assessed using data from 414 ribo-seq experiments across 42 studies, of which 270 experiments across 36 studies were kept after excluding experiments that did not show strong patterns of triplet periodicity among cORFs (**Supplementary Table 1, Supplementary Table 2**).

As expected, combining data from many experiments allowed for identification of translated ORFs that would otherwise have too few reads in any individual experiment (**Figure 1C**). Setting a confidence threshold to ensure a 5% FDR among nORFs, we identified 18,954 nORFs (**Figure 1D**) as translated along with 5,363 cORFs (**Figure 1E**), for a total of 24,317 ORFs making up the yeast reference translatome. This corresponds to an identification rate of 96% for cORFs and 11% for nORFs (**Figure 2A-B**). In general, translated cORFs are much longer (**Figure 2C)** and translated at much higher rates (**Figure 2D**) than translated nORFs.

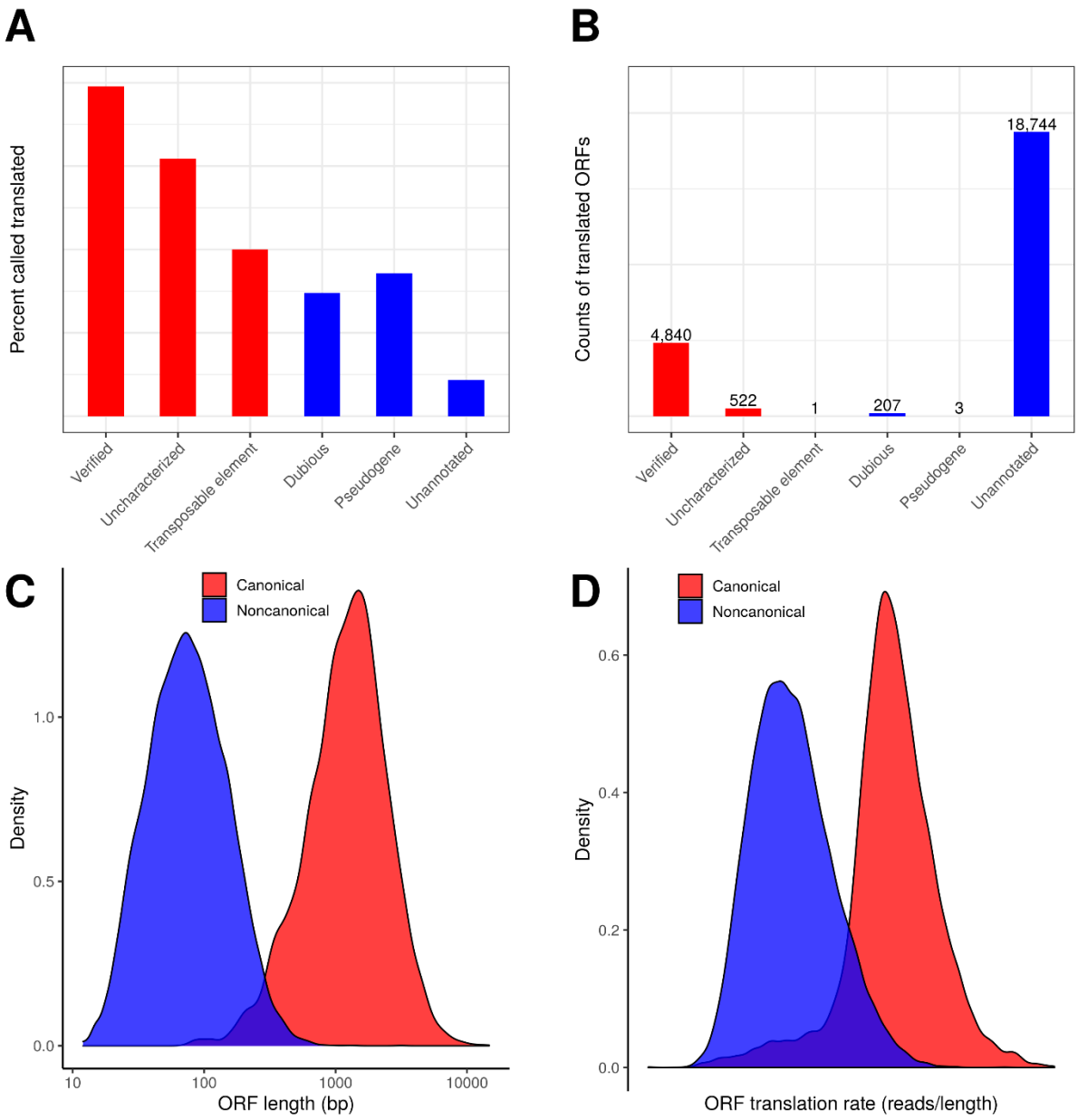


Figure 2: The noncanonical yeast translatome is larger than the canonical. A) The percent of ORFs in each Saccharomyces Genome Database annotation class that are detected as translated by iRibo, with canonical classes indicated in red and noncanonical in blue. B) The number of ORFs of each annotation class that are detected using iRibo. C) ORF length distribution for cORFs and nORFs. D) Distribution of translation rate (in-frame reads per base) for cORFs and nORFs.

To assess the consistency of our ribo-seq datasets, we measured the replicability of translation patterns between studies. In general, ribo-seq coverage among ORFs was highly correlated among studies (**Supplementary Figure 1A-B**). To assess replicability in translation calls for nORFs, we applied iRibo to

each individual study and identified the nORFs that could be inferred to be translated using only the reads in that study. We then determined the proportion of translated nORFs found using each large study that were also found using the largest study, Gerashchenko et al. 2014[30] (**Figure 3A**). All studies had replication rates of at least 75%. These observations demonstrate that non-canonical translation patterns are highly reproducible, suggesting that they are driven by regulated biological processes rather than technical artifacts or stochastic ribosome errors.

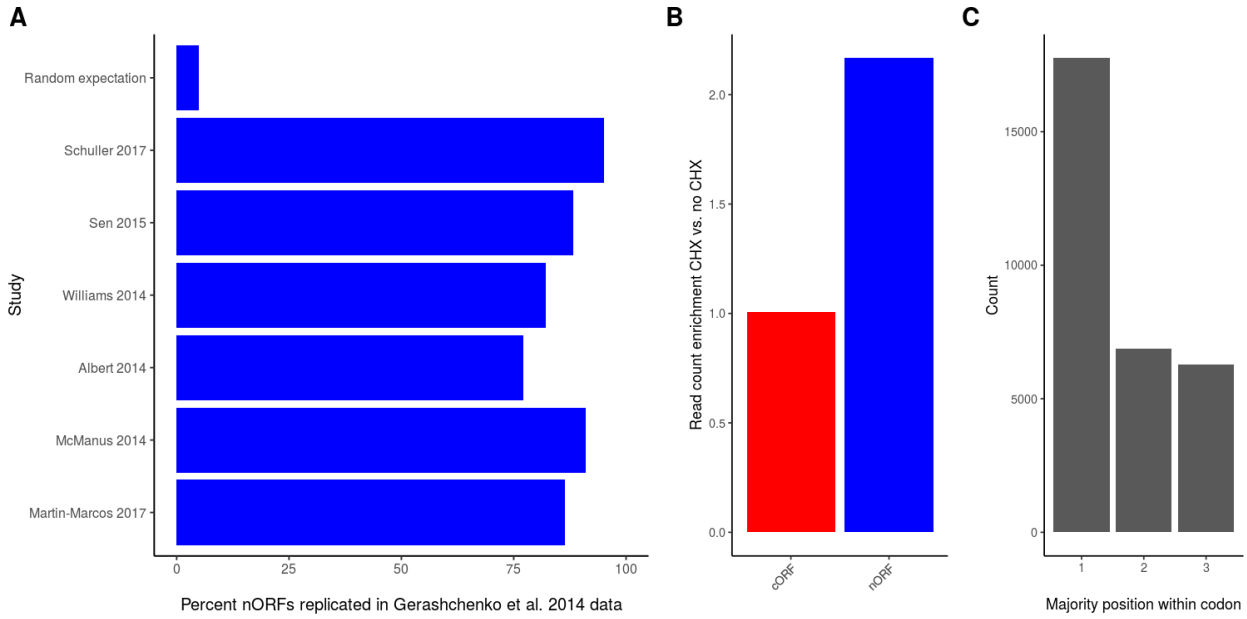


Figure 3: Translation patterns in noncanonical ORFs show high replicability between studies. A) For six large studies in our dataset, the proportion of nORFs identified using reads from that study that are replicated using reads from the largest study, Gerashchenko et al. 2014, is indicated. Random expectation is the proportion that would be expected to replicate by chance. B) Relative enrichment of ribo-seq read counts in the first position of each codon with vs. without CHX treatment. C) Codon position of mapped ribo-seq reads in the no CHX condition among ORFs identified only in the CHX condition. A strong preference for the first codon, characteristic of translation, is observed.

As usage of the translation inhibitor CHX to treat cells in ribo-seq studies has been widely discussed[30–32] as a factor influencing observed noncanonical translation patterns, we wished to specifically examine the consistency between studies in our dataset that differ in usage of this drug. We thus compared translation signatures from experiments with (N=139) and without (N=170) CHX, randomly sampling the same number of reads from both groups. We found a large enrichment in ribo-seq read counts among nORFs with CHX treatment, resulting in more nORFs identified as translated (**Figure 3B**). However,

nORFs identified as translated only in CHX nevertheless displayed strong triplet periodicity in its absence when analyzed as a group (**Figure 3C**), indicating that they are translated under normal conditions.

Noncanonical translation patterns depend on genomic and environmental context

We examined to what extent translation of nORFs depends on genomic context. We classified nORFs as: upstream nORFs (uORFs) located on the 5' untranslated regions of transcripts containing cORFs; downstream nORFs (dORFs) located on the 3' untranslated regions of transcripts containing cORFs; intergenic nORFs that do not share transcripts with cORFs (independent), antisense nORFs located entirely within a cORF (full overlap), and antisense nORFs that overlap the boundaries of a cORF (partial overlap) (**Figure 4A**). Around 35% of identified translated nORFs, including 4,031 uORFs and 2,993 dORFs, shared a transcript with a cORF, while 1.3% (268) were located on an annotated RNA gene (**Figure 4B**). The remaining 64% were located on transcripts that contain no annotated gene (5,958 independent, 4,826 full overlap, 1,779 partial overlap). We compared the frequency at which nORFs were identified as translated relative to expectations based on candidate nORFs between different contexts (**Figure 4C**). Genome-wide, 23% of nORFs on the same transcript as a cORF were identified as translated, significantly higher than the translation frequency of 13% for independent nORFs ($p<10^{-10}$, Fisher's Exact Test). Consistent with prior research[33], the relative position of the nORF on a transcript shared with a cORF affected likelihood of translation, with 28% of uORFs found to be translated compared to only 18% of dORFs ($p<10^{-10}$, Fisher's Exact Test). The nORFs in an antisense orientation to a cORF, and fully overlapping it, were translated at a frequency of 5%, the lowest of any context considered ($p<10^{-10}$ for any comparison).

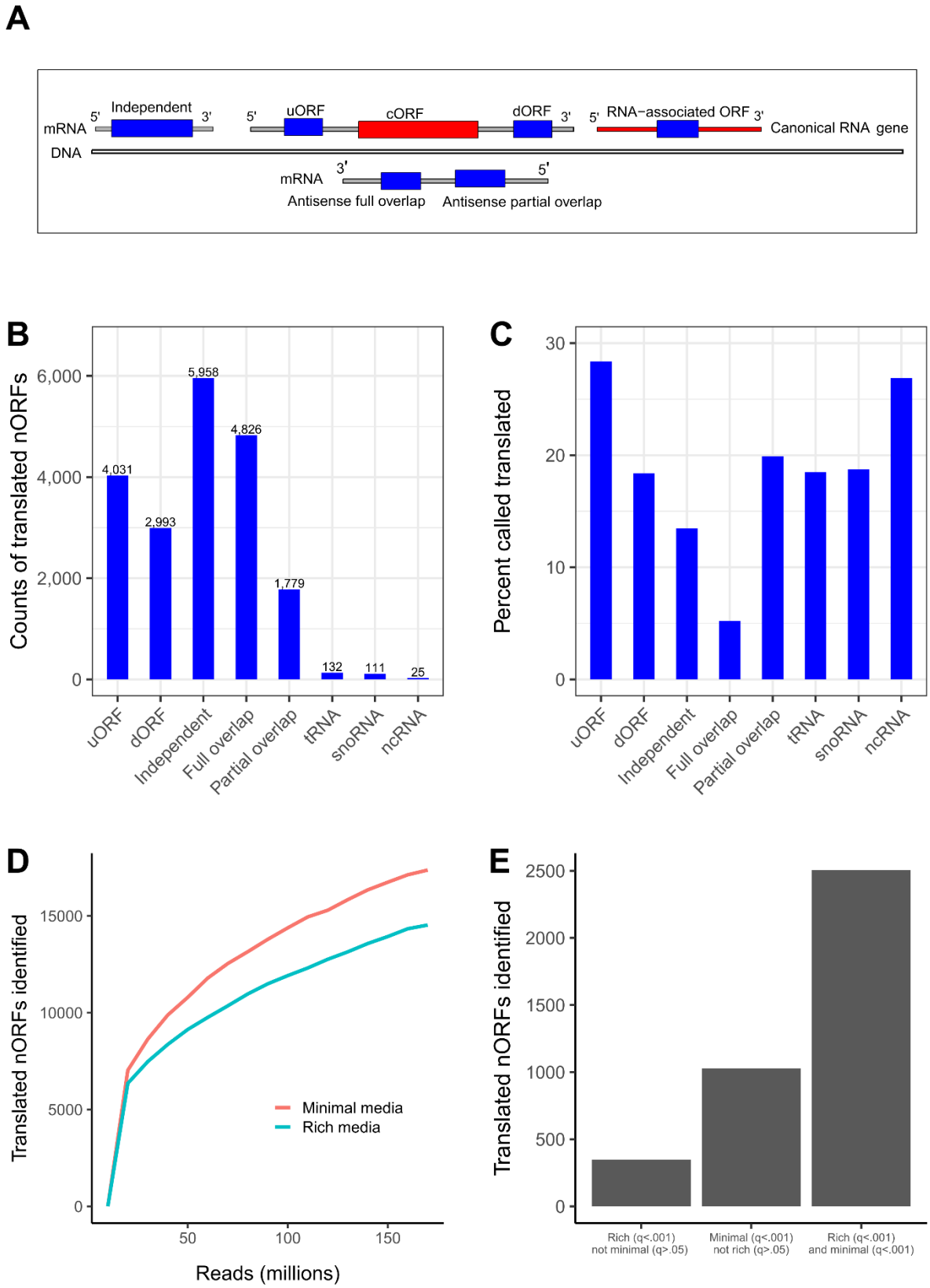


Figure 4: Noncanonical translation patterns depend on both genomic and environmental context. E) Potential genomic contexts for nORFs in relation to nearby canonical genes: on 5' untranslated region (uORF), on 3' untranslated region (dORF), intergenic nORFs that do not share transcripts with annotated genes (independent), antisense nORFs located entirely within an annotated gene (full overlap), and antisense ORFs that overlap the boundaries of an annotated gene (partial overlap). Also considered are nORFs that share a transcript with an RNA gene. B) Proportion of nORFs detected as translated by iRibo in each genomic context considered. For nORFs that share a transcript with RNA genes, the annotation of the RNA gene is specified. C) Counts of translated nORFs identified in each considered genomic context. D) Number of translated nORFs identified for experiments on yeast grown in either minimal (SD) or rich media (YPD), at a range of read counts. E) Number of nORFs identified at high confidence either exclusively in rich media or minimal media (q-value <.001 in one condition and q-value >.05 in the other) or found at high confidence in both conditions.

We next investigated how environmental context affects noncanonical translation. To this aim, we leveraged the power of iRibo to construct separate datasets of nORFs found translated in rich media (YPD) or in nutrient-limited minimal media (SD) (**Supplementary Table 3**). Previous research had reported an increase in noncanonical translation in response to starvation.[1,19] Consistent with these results, at equal read depths, more nORFs were identified as translated in minimal than in rich media (**Figure 4D**). Furthermore, 1,028 nORFs were found as translated with high confidence specifically in minimal media but showed no evidence of translation at all in rich media (q-value < .001 in SD; q-value >.05 in YPD), while only 348 nORFs were found translated specifically in rich media but showed no evidence of translation in minimal media (**Figure 4E**). These results confirm that nORF translation is regulated in response to changing environments.

Two translatomes, transient and conserved

To determine whether the proteins encoded by translated nORFs are being maintained by selection, we performed integrative comparative genomics analyses across three evolutionary scales. At the population level, we analyzed 1011 distinct *S. cerevisiae* isolates sequenced by Peter et al. 2018.[28] At the species level, we compared *S. cerevisiae* ORFs to their orthologs in the *Saccharomyces* genus, a taxon consisting of *S. cerevisiae* and its close relatives[34]. To detect long term conservation, we looked for homologs of *S. cerevisae* ORFs among the 332 budding yeast genomes (excluding *Saccharomyces*) collected by Shen et al[29]. The power to detect selection on an ORF depends on the amount of genetic variation in the ORF available for evolutionary inference, which in turn depends on its length, the density of genetic variants across its length, and the number of genomes available for comparison. Given that many translated nORFs are very short (**Figure 2C**), we employed a two-stage strategy to increase power for detecting signatures of selection. First, we investigated selection in a set of "high information" ORFs for which we have sufficient statistical power to potentially detect selection. Second, we

209 investigated the remaining "low information" ORFs in groups to quantify collective evidence of selection
210 (**Figure 5A**). Group level analysis increases power to detect the presence of selection but does not
211 enable identification of the specific ORFs under selection. The "high information" set consisted of the
212 ORFs that 1) have identified orthologs in at least four other *Saccharomyces* species and 2) have a
213 median count of nucleotide differences between the *S. cerevisiae* ORF and its orthologs of at least 20.
214 We found these criteria are sufficient to distinguish ORFs evolving under selection (**Supplementary**
215 **Figure 2**). Under this definition, 9,453 translated ORFs that do not overlap cORFs (henceforth
216 "nonoverlapping ORFs", including 4,223 nORFs, and 5,230 cORFs) and 3,063 antisense ORFs (3,003
217 nORFs and 60 cORFs) were placed in the "high information" set.

218 To detect selection in the high information set, we first used reading frame conservation (RFC), a
219 sensitive approach developed by Kellis *et al.* 2003[13] to distinguish ORFs under selection from ORFs that
220 exist by happenstance in the yeast genome. RFC ranges from 0 to 1, measuring codon structure
221 conservation between an *S. cerevisae* ORF and potential orthologs in the *Saccharomyces* genus. We
222 found a bimodal distribution of RFC among nonoverlapping ORFs in the yeast translatome: 53.7% have
223 RFC above 0.8 and 44.4% have RFC less than 0.6, with only 1.9% of ORFs intermediate between these
224 values (**Figure 5B)**. The modes of the distribution largely correspond to annotation status, with 96.4% of
225 cORFs having RFC > 0.8 and 96.8% of nORFs falling in RFC < 0.6 category. This bimodal distribution of RFC
226 among translated ORFs was similar to that observed among all candidate ORFs in the yeast genome.[13]
227 High RFC among antisense ORFs does not demonstrate selection on the ORF itself, as it might be caused
228 by selective constraints on the opposite-strand gene, but low RFC still indicates lack of conservation. A
229 majority of antisense translated nORFs (65%) have RFC <0.6, indicating that most are not preserved by
230 selection (**Supplementary Figure 3**).

231 In light of the general correspondence between annotation and conservation, the exceptions are of
232 interest: 126 cORFs (111 nonoverlappping and 15 antisense) showed poor conservation and therefore
233 might not be evolving under purifying selection, while 13 nonoverlapping nORFs had preserved ORF
234 structure and are thus potentially evolving under purifying selection. Several lines of evidence suggest
235 that these preserved nORFs are indeed evolving under purifying selection (**Table 2)**. For nine of the
236 thirteen, we identified a BLASTP or TBLASTN match among 332 budding yeast genomes[29] (excluding
237 *Saccharomyces* genus species), suggesting conservation over long evolutionary time. Two ORFs showed
238 evidence of selection in a pN/pS analysis performed on 1011 *S. cerevisiae* isolates[28], and three others
239 showed evidence of selection by dN/dS performed on the *Saccharomyces* genus species (**Table 2**). We

sought to determine whether selection could be inferred for any additional nORFs on the basis of long-term evolutionary conservation. We searched for distant homologs of translated nonoverlapping *S. cerevisiae* nORFs using TBLASTN within budding yeast genomes outside the *Saccharomyces* genus[29]. After excluding matches that appeared non-genic (**Supplementary Figure 4A-B, Supplementary Table 4**) we identified a single additional ORF with both distant TBLASTN matches and recent signatures of purifying selection: *YBR012C*, annotated as "dubious" on SGD (**Table 2**). Thus, combining the 13 nORFs that appeared conserved by RFC analysis and the single additional ORF with signatures of long-term conservation by TBLASTN, we identified 14 translated nORFs that show evidence of preservation by selection (**Table 2**).

Table 2: Properties of well-conserved nORFs

Systematic Name	Coordinates	BLASTP e-value	TBLASTN e-value	RFC	Length (nt)	pN/pS (p-value)	dN/dS (p-value)	Translation percentile
YBL029W-B	chrII:164192-164368	6.8×10^{-4}	8.0×10^{-3}	0.82	177	1.65 (.33)	0.88 (.68)	0.67
YBL014W-A	chrII:196737-196889	4.3×10^{-5}	1.0×10^{-4}	1	153	0.47 (.11)	0.14 (3.46×10^{-12})	0.86
YBR085W-B	chrII:417494-417556	1	1	0.86	63	0.72 (.48)	1.26 (.62)	0.58
YBR268W-A	chrII:741844-742005	1	1	0.99	162	0.61 (.15)	0.35 (3.18×10^{-7})	0.97
YBR292W-A	chrII:786745-786903	1.9×10^{-7}	5.0×10^{-3}	0.96	159	0.72 (.43)	0.57 (.0026)	0.83
YER186W-A	chrV:565603-565800	6.2×10^{-6}	1	0.92	198	0.55 (.02)	1.0 (1)	0.97
YGL262W-A	chrVII:4663-4872	1	1.0×10^{-3}	0.88	210	0.96 (.86)	1.0 (1)	0.86
YGR238W-A	chrVII:969015-969089	1	1	0.87	75	0.20 (.01)	1.18 (.74)	0.94
YBL049C-A	chrII:126330-126461	8.7×10^{-5}	6.0×10^{-4}	0.84	132	1.36 (.79)	1.5 (.22)	0.75
YBL026C-A	chrII:169634-169870	7.0×10^{-12}	9.0×10^{-10}	0.88	237	1.30 (.6)	0.87 (.42)	0.9996
YJR107C-A	chrX:628457-628693	3.9×10^{-8}	3.0×10^{-18}	0.99	237	0.39 (.005)	1.42 (.13)	0.9991
YLR349C-A	chrXII:828276-828338	1	1	0.81	63	0.30 (.02)	0.73 (.24)	0.73
YNR062C-A	chrXIV:745640-745792	5.2×10^{-14}	5.0×10^{-13}	0.89	153	0.65 (.44)	1.49 (.15)	0.44
YBR012C	chrII:259147-259566	6.51×10^{-59}	$1x10^{-16}$	.70	420	.62 (.1)	.50 (.039)	0.92

To obtain power for analyzing selection among "low information" ORFs, including 8,062 nonoverlapping nORFs and 21 cORFs (8,695 low information antisense ORFs were not analyzed), we analyzed collective evidence of selection within specified groups of ORFs. These groups were constructed as deciles of properties that we expected to be potentially associated with selection. The properties considered were genomic context, rate of translation (as measured by ribo-seq reads mapped to the first position within codons), ORF length, and coding score[35,36]. For each group, we calculated pN/pS ratio among 1,011 *S. cerevisiae* isolates[28]. Low-information cORFs showed pN/pS ratio significantly below 1, indicating that

some ORFs in the group are evolving under purifying selection (**Figure 5C**). In contrast, for all groups of nORFs examined, we observed no significant deviations from neutral expectations in pN/pS (**Figure 5C**). To assess whether each group showed collective evidence of distant homology that could not be established at the individual level with confidence, we also calculated the frequency of weak TBLASTN matches (e-values between 10^{-4} and .05). The frequency of weak matches was compared to a negative control set consisting of scrambled sequences of the ORFs in each group. Applying this strategy to cORFs lacking strong matches, we found a large excess of weak matches relative to controls (**Figure 5D**), demonstrating the ability of this approach to detect faint signals of homology within a group of ORFs. However, we identified no significant difference in the frequency of weak TBLASTN hits between any nORF group and scrambled controls (**Figure 5E**). The lack of a significant result does not exclude the possibility that a small subset of short conserved nORFs could be lost in the noise of a much larger set of nORFs evolving close to neutrally. However, these analyses indicate that ORFs evolving under strong purifying selection are not a major component of the yeast noncanonical translatome.

Overall, our analyses distinguish two distinct yeast translatomes: a conserved, mostly canonical translatome with intact ORFs preserved by selection; and a mostly noncanonical translatome where ORFs are not preserved over evolutionary time. This distinction is rooted in evolutionary evidence rather than annotation history. We thus propose to group the translated ORFs that showed no evidence of selection in either our high-information or low-information set as the “transient translatome”. The transient translatome includes the 4088 nonoverlapping and 1948 antisense nORFs identified as not preserved by selection using RFC analyses, along with 90 nonoverlapping and 15 antisense cORFs matching the same criteria. Also included are all 8041 nonoverlapping nORFs that lack sufficient information to analyze at the individual level but were found to show no selective signal in group-level analyses. Together, this set of 14,203 ORFs that are translated yet evolutionarily transient makes up 58% of the yeast reference translatome (**Figure 5F**).

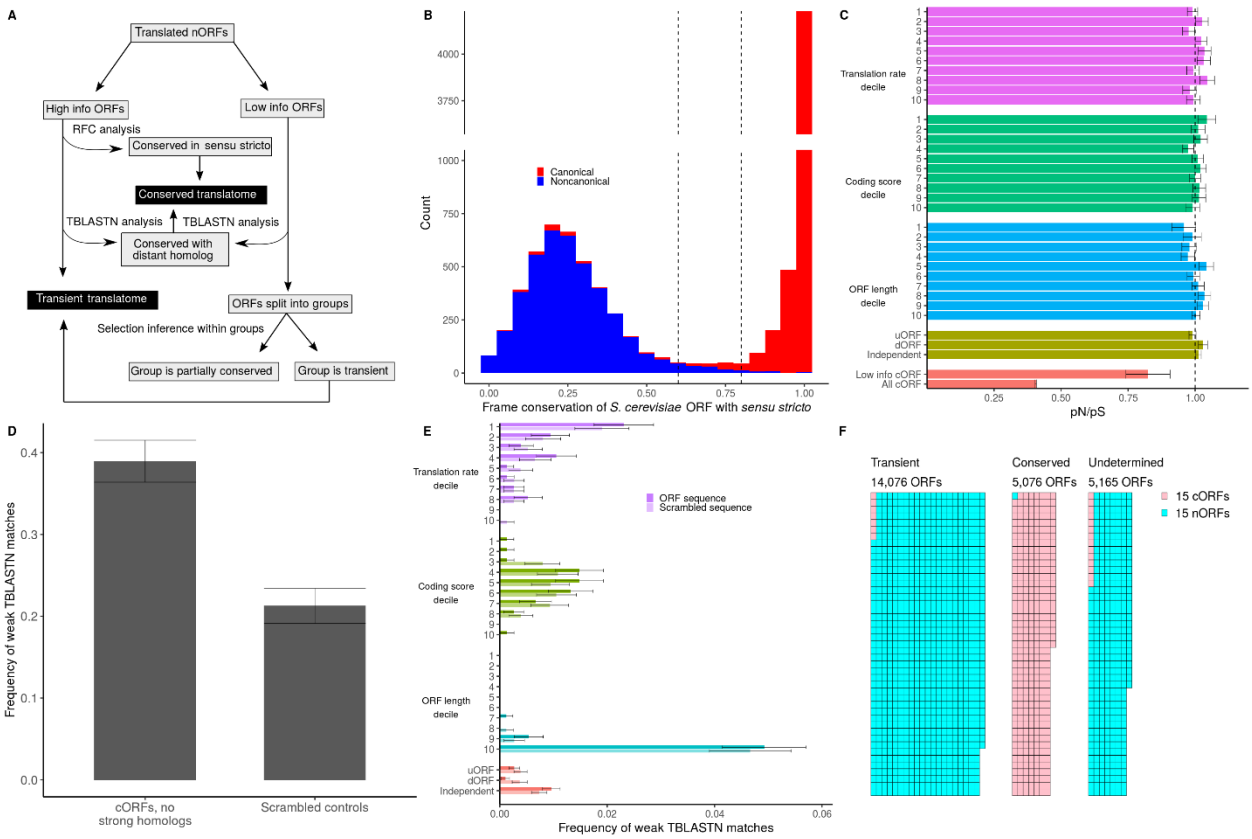


Figure 5: Two distinct translatomes: transient and conserved. A) Selection inference analyses conducted on low-information and high-information nORFs. B) The distribution of reading frame conservation among high information ORFs, separated between noncanonical and (stacked above) canonical. Dashed lines separate RFC < 0.6 and RFC > 0.8, the thresholds use to distinguish ORFs preserved or not preserved by selection. C) pN/pS values are shown for each group of low-information nORFs, representing a decile of translation rate (in-frame ribo-seq reads per base), coding score, or ORF length, as well as ORFs in different genomic contexts. Note that pN/pS values are not averages among ORFs but a ratio reflecting the number of synonymous and nonsynonymous variants pooled over the entire class. Error bars indicate standard errors estimated from bootstrapping. D) The frequency of weak TBLASTN matches (10^{-4} < e-value < .05) among budding yeast genomes for cORFs that lack any strong matches, and controls consisting of the same sequences randomly scrambled. Error bars indicate standard errors estimated from bootstrapping. E) The frequency of nORFs with weak TBLASTN matches (10^{-4} < e-value < .05) in each group of nORFs (dark bars) and negative controls (light bars) consisting of the sequences of the nORFs of each group randomly scrambled. Error bars indicate standard errors estimated from bootstrapping. F) The components of the translatome are represented with area proportional to frequency.

Most annotated transient ORFs appear biologically significant

We have identified a large collection of nORFs that show strong evidence of translation but appear to be evolutionarily transient and have no clear evolutionary signature of selection (**Figure 5A-C**). By general theory and practice in evolutionary genomics, the lack of any selective signal suggests that the

noncanonical transient translatome does not meaningfully contribute to fitness. Surprisingly, however, 105 cORFs also belong to the transient set. If lack of selective signal implies lack of function, why are these ORFs classified as canonical genes? To better understand the potential roles of these ORFs, we examined what has been discovered about each transient cORFs in the *S. cerevisiae* experimental literature.

While most transient cORFs are not well-characterized, five have been studied in depth. Two of these, *MDF1*[37] and *YBR196C-A*[38], have been previously discussed as apparent *de novo* genes; the remaining three have been characterized, but their evolutionary properties were not analyzed in the corresponding studies. *HUR1* plays an important role in non-homologous end-joining repair and its encoded polypeptide physically interacts with conserved proteins[39]. Both deletion and overexpression mutants of *YPR096C* indicate that it regulates translation of *PGM2*.[40] A thorough investigation of *ICS3* mutants demonstrates its involvement in copper homeostasis[41]. These cases indicate the potential for evolutionarily transient ORFs to play important biological roles. For transient cORFs with no described role, we examined all literature listed as associated with the ORF on SGD. Many of these transient cORFs are supported by direct evidence of phenotype (**Supplementary Table 5**). Of particular interest are the 98 transient cORFs with null mutants included in the yeast deletion collection.[42] Of these cORFs, 35 (36%) were associated with phenotypes in at least one screen using the collection. An additional 10 transient ORFs were reported to have null mutant phenotypes in other screens, and 11 to have overexpression phenotypes. Overall, we found phenotypes reported in the literature for 51 of 105 transient cORFs (49%).

In addition to the set of transient cORFs, 144 transient nORFs are annotated as “dubious” on SGD. Though considered unlikely to encode a protein in the current version of the genome annotation, these ORFs have nevertheless been investigated in various studies. To further determine the potential for biological activity in transient nORFs and cORFs, we assessed whether each expressed a stable protein that can be detected in the cell. Fifty transient cORFs were identified among 21 yeast quantification studies assembled by Ho et al. 2018[43]. We examined two microscopy datasets for additional evidence, the CYCLoPs database of GFP-tagged proteins[44] and the C-SWAT tagging library developed by Meurer et al. 2018[45]. Both of these datasets attempted to localize proteins expressed from their native promoters. Together, the CYCLoPs and C-SWAT libraries identified 26 of 36 (72%) transient “dubious” nORFs examined and 71 of 88 (81%) transient cORFs (**Figure 6A**). These results indicate that a majority of the

proteins coded by transient nORFs and cORFs exist stably within the cell and have the potential to affect phenotypes.

Next, we sought to determine how many annotated transient ORFs can affect fitness. To this aim, we leveraged the large yeast genetic interaction network assembled in Costanzo et al. 2016.[46] This dataset includes 81 transient cORFs and 13 "dubious" transient nORFs. Deletion strains for these 94 transient ORFs exhibited an average fitness of 0.99, not significantly different from the wildtype fitness of 1.0 (p=0.06, t-test) (**Figure 6B**). However, despite the lack of substantial single-mutant effects, most transient ORFs participated in strong negative genetic interactions. Out of the 94 transient ORFs in the dataset, 89 (95%) have at least one negative interaction strength at Ɛ<-0.2 and p-value<0.05 (described as a high-stringency cut-off by Costanzo et al.) and 63 (67%) have negative interactions with Ɛ<-0.35, the threshold for synthetic lethality in Costanzo et al.[46] (**Figure 6C**). This was only a slightly lower rate than for conserved non-essential ORFs, 98% of which had interactions with Ɛ<-0.2 and p-value <0.05 (p=0.047, Fisher's exact test), and 77% of which had interactions with Ɛ<-.35 (p=0.026, Fisher's exact test). To further investigate these interactions, we performed GO enrichment analyses on the genetic interactors of each transient ORF. At an Ɛ< -0.2 threshold, 27 transient ORFs were found to interact with groups of related genes enriched in specific GO terms (5% FDR; **Supplementary Table 6**). For example, the interactors of *YER175W-A* are associated with the GO category "cryptic unstable transcript (CUT) metabolic processes" with high confidence, and five of its eleven interactors are components or co-factors of the exosome (**Figure 6D**), indicating likely involvement in CUT degradation or a closely related pathway. The GO associations demonstrate biologically coherent knockout phenotypes for many transient ORFs.

Overall, we uncovered evidence that 131 of 249 (53%) annotated transient ORFs have at least one indicator of biological significance (detection of a protein product, a reported phenotype in a screen, or a genetic interaction in the Costanzo et al. 2016[46] network) (**Figure 6E**). This is likely an underestimate due to study bias. For example, many "dubious" ORFs have been excluded from the gene mutant libraries that are used in genetic screens and localization studies.

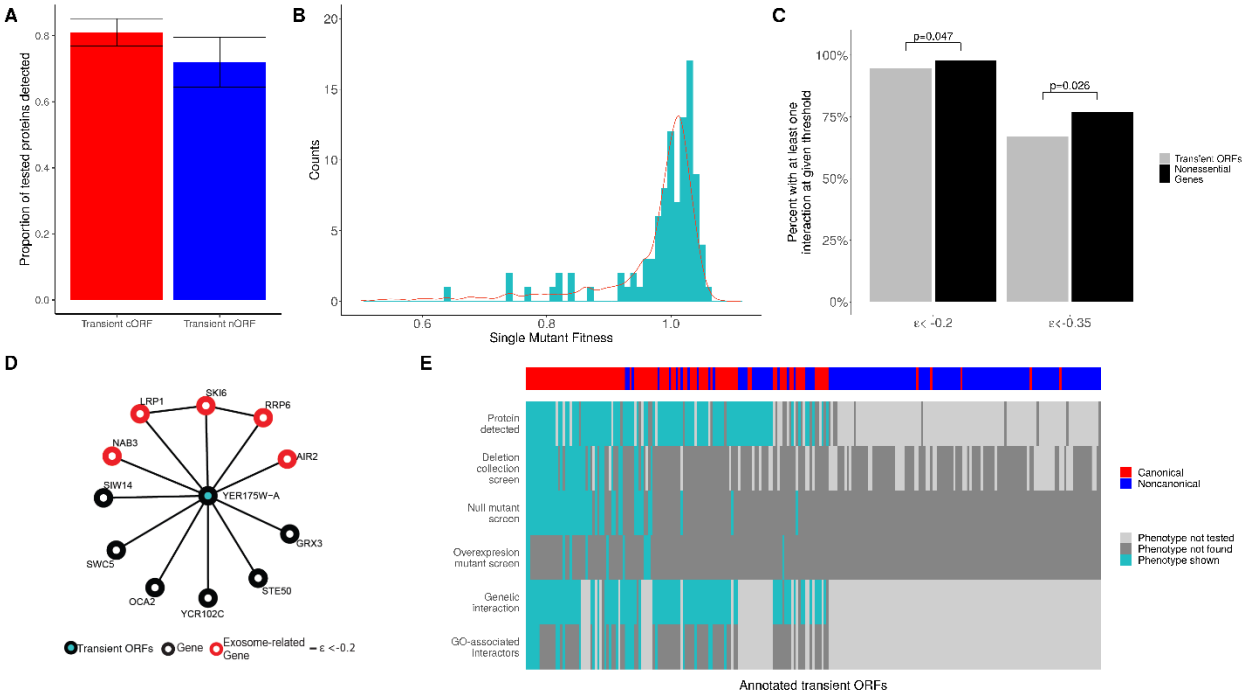


Figure 6: Many annotated transient ORFs have phenotypes indicative of biological roles. A) Proportion of proteins expressed by transient ORFs detected in either the CSWAP or CYCLoPS tagging libraries out of those tested. B) Histogram of single deletion mutant fitness among transient ORFs. The fitness distribution of nonessential genes is plotted in red for comparison. C) The percent of transient ORFs and nonessential genes with at least one genetic interaction at a given threshold. E) Genetic interactions of the transient ORF YER175W-A. Five interactors are related to exosome. F) Presence of phenotypes among all annotated transient ORFs. "Protein detected" indicates that the ORF product was found in either the CSWAP or CYCLoPS database. Phenotypes of deletion collection, null and overexpression screens were taken from reported findings in the yeast experimental literature and are described in Supplemental Table 5. "Genetic interaction" indicates a statistically significant genetic interaction with ε< -0.2, and "GO-associated interactors" indicates a GO enrichment was found among significant interactors at 5% FDR.

Transient annotated ORFs appear to be representative of the transient translatome overall

We sought to determine whether the level of biological significance observed for the annotated subset of the transient translatome could be representative of the transient translatome as a whole. To this aim, we compared the evolutionary properties, translation rate and coding scores of transient cORFs, transient "dubious" nORFs and transient unannotated nORFs. No class of transient ORF showed a pN/pS ratio different from one or from untranslated negative controls (**Figure 7A**), consistent with neutral evolution. Similarly, the average nucleotide diversity within each transient subset was not significantly different than untranslated controls, and much higher than conserved genes (**Figure 7B**). Frame conservation with *S. paradoxus* also showed no difference from the controls (**Figure 7C**). In addition, no

class of transient ORFs showed differences in their rate of translation (**Figure 7D**) or coding score (**Figure 7E)**. The only distinguishing property between annotated and unannotated transient ORFs was their length. Both transient cORFs and "dubious" nORFs are much longer on average than unannotated transient nORFS (**Figure 7F**). This is a consequence of the history of annotation of the *S. cerevisiae* genome, where a length threshold of 300 nt was set for annotation of unknown ORFs[47,48]. The sharp 300 nt threshold is still clearly reflected in annotations. For example, genome annotations include 96% of nonoverlapping transient ORFs in the 300-400 nt range (55/57), but only 4% in the 252-297 nt range (4/101). This cutoff was not set due to a belief that shorter ORFs could not be biologically relevant—118 ORFs annotated as "verified" on SGD are shorter than 300 bp—but due to difficulty in distinguishing potentially significant ORFs from those arising by chance.[49] Thus, given that 300 bp does not represent a threshold for biological significance, and transient unannotated ORFs resemble transient cORFs in all other respects, numerous never-studied transient nORFs likely also play a variety of biological roles.

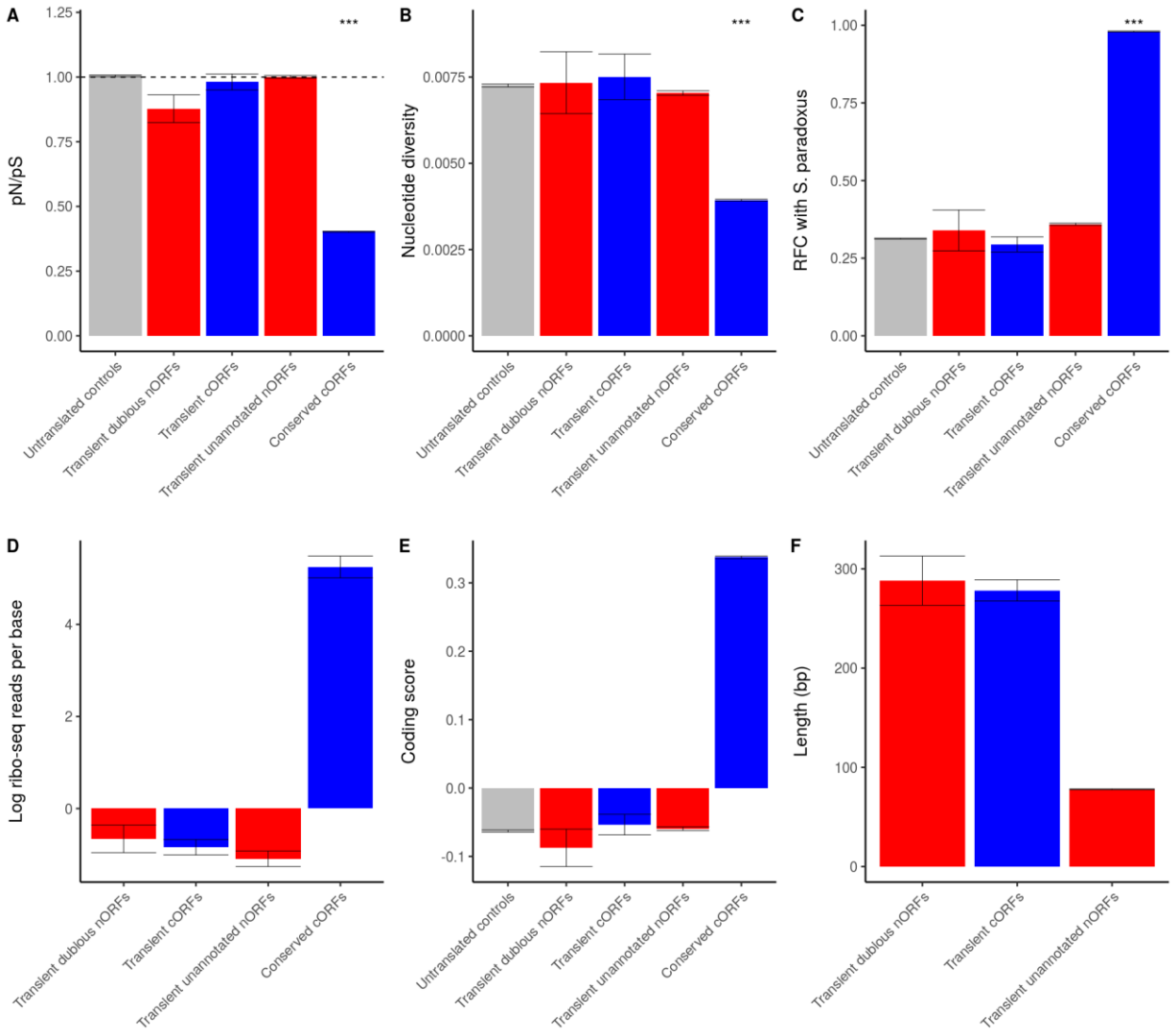


Figure 7: Canonical and noncanonical transient ORFs have similar properties. A-G) Properties of nonoverlapping transient cORFs and nORFs. Untranslated controls consist of nonoverlapping ORFs that would be grouped in the transient class (RFC <.6) but are not inferred to be translated based on ribo-seq evidence. Conserved cORFs are nonoverlapping cORFs with distant homologs and high RFC (>.8). P-value<.05:*. P-value<.01:**. P-value <.001: ***. A) pN/pS values for each group among *S. cerevisiae* strains. B) Average nucleotide diversity (π) among each group. C) Average reading frame conservation between *S. cerevisiae* and *S. paradoxus* ORFs. D) Average ribo-seq reads per base, considering only in-frame reads. E) Coding scores are plotted for ORFs of each group. F) ORF lengths in nucleotides are shown for ORFs of each group.

Discussion

Since the advent of ribosome profiling, it has been evident that large parts of eukaryotic genomes are translated outside of canonical protein-coding genes[1], but the nature and full significance of this translation has remained elusive. To facilitate study of this noncanonical translatome, we developed iRibo, a framework for integrating ribosome profiling data from a multitude of experiments in order to sensitively detect ORF translation across a variety of environmental conditions. Here, we demonstrate

that iRibo is able to identify a high confidence yeast reference translatome almost five times larger than the canonical translatome. This resource can serve as the basis for further investigations into the yeast noncanonical translatome, including the prioritization of nORFs for experimental study.

We used the iRibo dataset to address a fundamental question about the yeast noncanonical translatome: to what extent does it consist of conserved coding sequences that were missed in prior annotation attempts? In a thorough evolutionary investigation, we identified 14 translated nORFs that show evidence of being conserved under purifying selection. Only one of these ORFs, YJR107C-A, appears to have been previously described[22], though it is not annotated on Saccharomyces Genome Database. Thus, even a genome as well-studied as *S. cerevisiae* contains undiscovered conserved genes, likely missed in prior analyses due to difficulties in analyzing ORFs of short length. These 14 nORFs are, however, the exception: the great majority of translated nORF show no signatures of selection whatsoever, comprising a large pool of evolutionarily transient translated sequences.

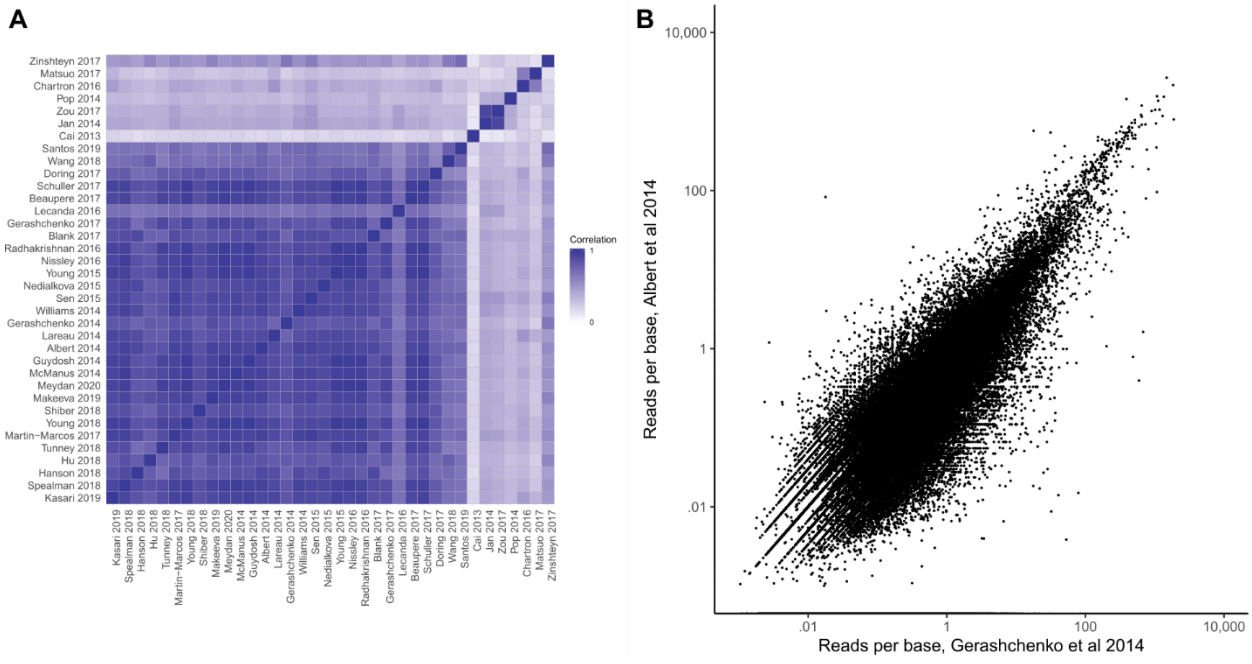
We identified and analyzed a collection of transient annotated ORFs to get a sense of the potential roles played by the much larger set of transient unannoated ORFs. Despite lacking evidence of selection, annotated transient ORFs expressed stable proteins and contributed to cellular processes and phenotypes. These annotated ORFs were representative of the transient translatome as a whole besides being longer on overage, but this difference stems from a decision made early in the annotation of the yeast genome not to annotate most ORFs shorter than 100 codons.[47] As this annotation choice was based only on length and not direct evidence of phenotype, it does not serve as evidence that shorter transient ORFs lack phenotypes observed in larger transient ORFs. Indeed, research on microproteins show clearly that sequences shorter than 100 codons are often biologically important.[5,50]

It is perhaps surprising that a coding sequence can affect organism phenotype despite showing no evidence of selection. However, this result is consistent with evidence from the field of *de novo* gene birth. Species-specific coding sequences have been characterized in numerous species[20]. Xie et al. 2019[51] identify a mouse protein contributing to reproductive success that experienced no evident period of adaptive evolution. Sequences that contribute to phenotype without conservation have also been described outside of coding sequences. Many regulatory sequences, such as transcription factor binding sites, are a mix of relatively well-conserved elements and elements that are not preserved even between close species;[52] it is thus plausible that translated sequences also show such a division. These findings do not imply an absence of selective forces in shaping the patterns of noncanonical translation.

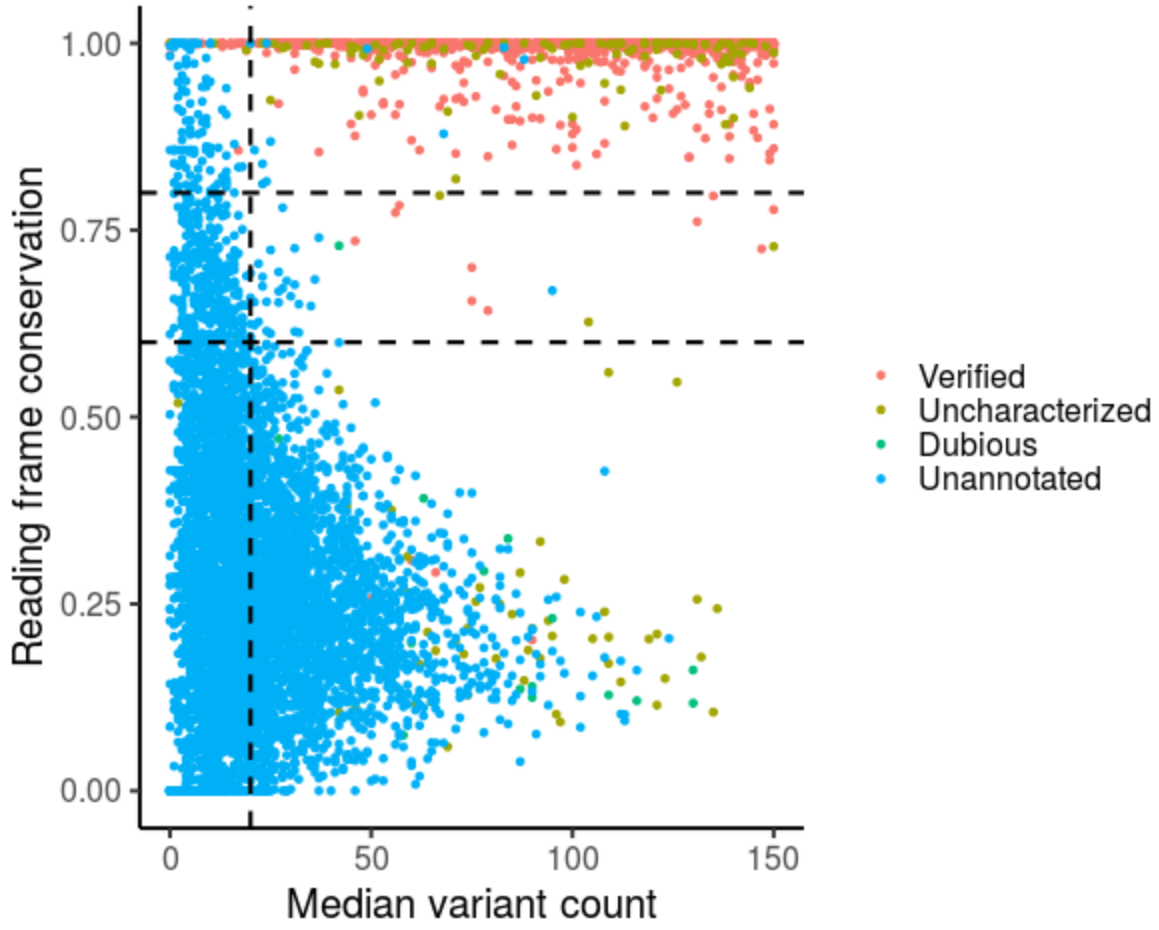
Rather, the particular selective environment favoring expression of these sequences may be too short-lived to detect selection using traditional comparative genomics approaches.

Our results indicate that the yeast noncanonical translatome is neither a major reservoir of conserved genes missed by annotation, nor mere "translational noise." Instead, many translated nORFs are evolutionarily novel and likely affect the biology, fitness and phenotype of the organism through species-specific molecular mechanisms. As transient ORFs differ greatly in their evolutionary and sequence properties from conserved ORFs, they should be understood as representing a distinct class of coding element from most canonical genes. Nevertheless, as with conserved genes, understanding the biology of transient ORFs is necessary for understanding the relationship between genotype and phenotypes.

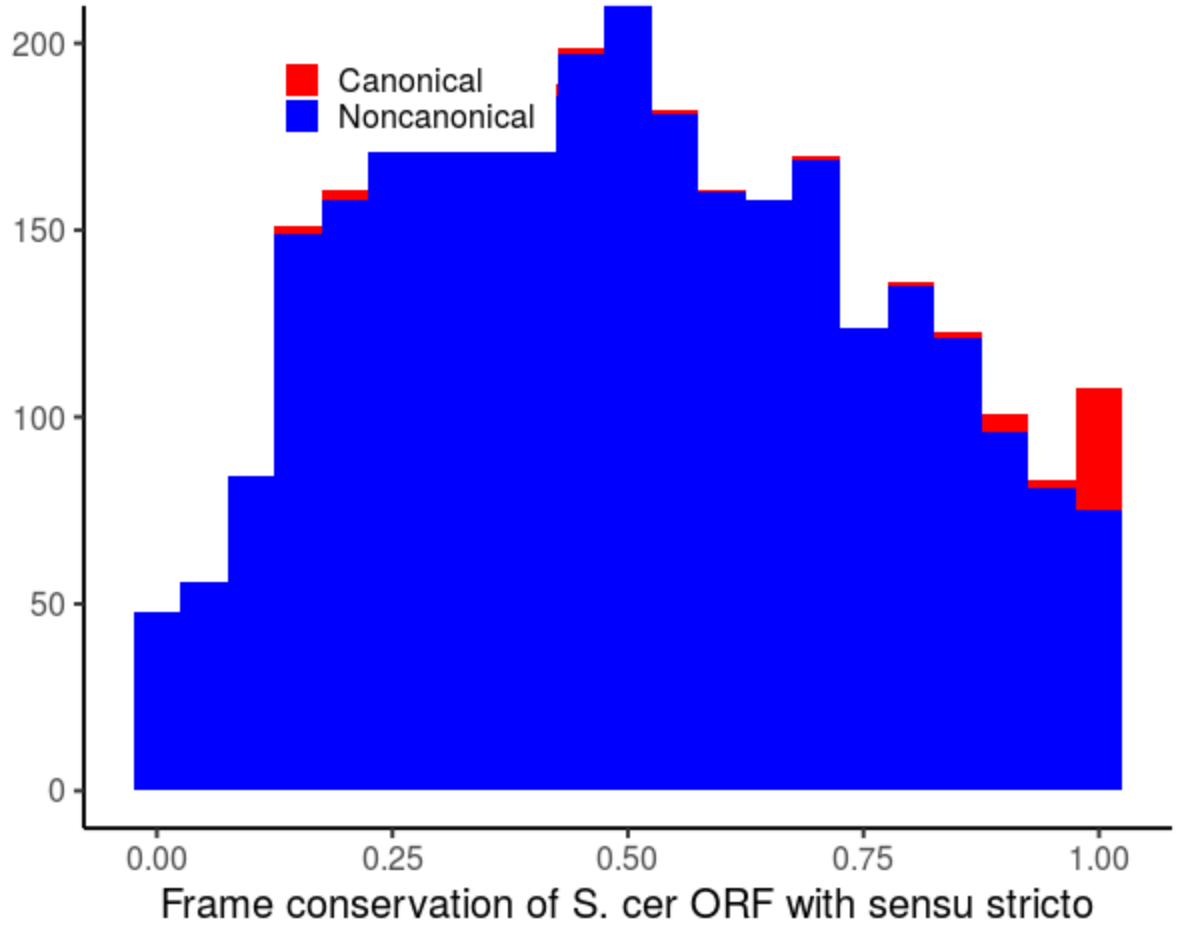
Supplementary Figures



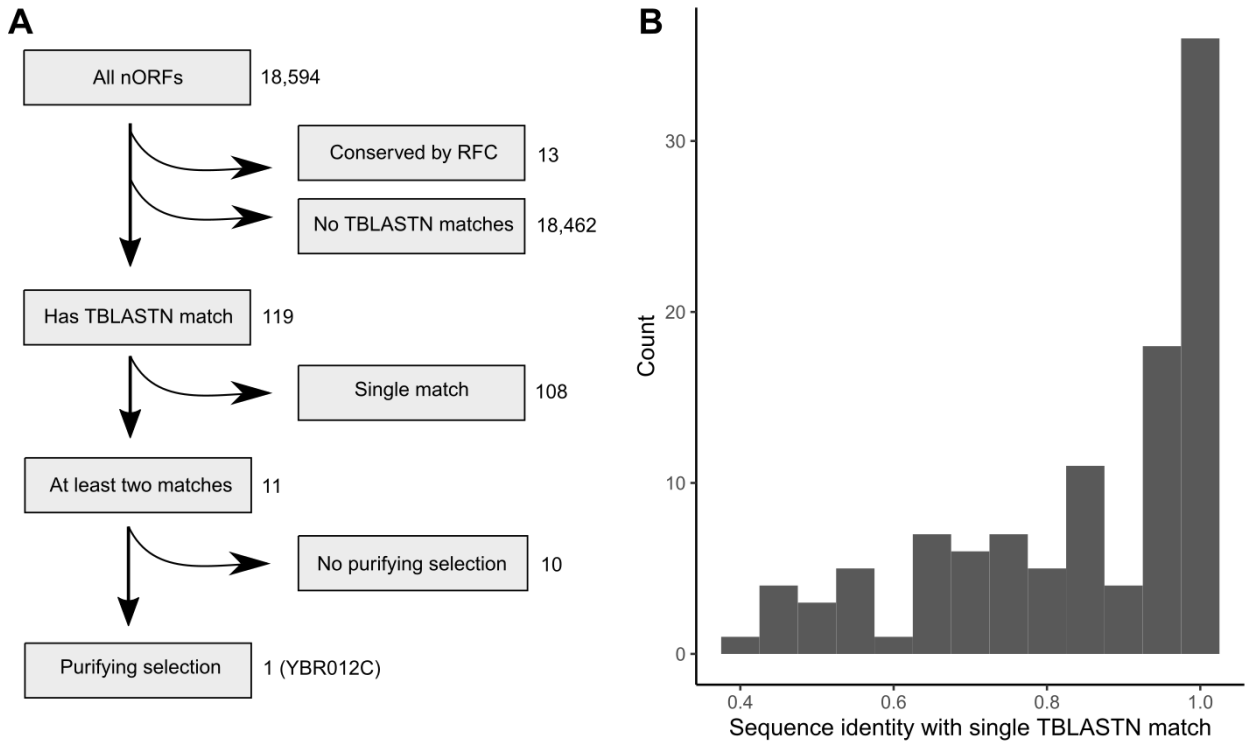
Supplementary Figure 1: Translation patterns in candidate ORFs show high replicability between studies. A) Pairwise correlation between ribo-seq coverage of all candidate ORFs between studies included in dataset. The set of 27 studies at the bottom left show high correlation among each other, while other studies show more distinct translation patterns. B) For each candidate ORF, the reads per base (considering only in-frame reads) are plotted for the two largest studies in our dataset.



Supplementary Figure 2: Nucleotide variation determines ability to distinguish conserved ORFs. Reading frame conservation for each nonoverlapping ORF is plotted against the median count of differences between the *S. cerevisiae* ORF and the aligned homologous sequence in each *Saccharomyces* relative. Colors indicate SGD annotation categories. The dashed lines separate distinct groups: to the right of the vertical line, there are two distinct populations divided by reading frame conservation, along with an intermediate region with few ORFs. For ORFs to the left of the vertical line, with few differences between species, there is no clear distinction between high-RFC and low-RFC populations



Supplementary Figure 3: Distribution of frame conservation among anti-sense ORFs. The distribution of frame conservation is plotted for translated cORFs and nORFs that are antisense to canonical genes, with canonical stacked atop noncanonical. In contrast to frame conservation among nonoverlapping ORFs, the distribution does not appear bimodal.



Supplementary Figure 4: Identification of conserved genes in the noncanonical translatome using TBLASTN. A) Process for identification of conserved nORFs evolving under purifying selection. Starting with the full list of nORFs, nORFs identified as conserved by RFC analysis are excluded, as these are already described in Table 2. The remaining nORFs with TBLASTN matches are divided into those only a single match among all compared species and those with at least two matches. Single matches were excluded, as these could be a result of contamination of genome sequencing data. The properties of the nORFs with multiple distant identified homologs were then examined for additional evidence of purifying selection (Supplementary Table 3). B) Among translated *S. cerevisiae* ORFs with a single TBLASTN hit among budding yeasts outside the *Saccharomyces* genus, the distribution of sequence identities with that match is plotted. The existence of only a single match together with the prevalence of high sequence identities (>80%) suggests that the matches may be the result of genomic contamination rather than genuine homology.

Methods

Yeast ribo-seq dataset collection and read mapping

We identified a list of *S. cerevisiae* ribosome profiling (ribo-seq) studies by conducting a broad literature search. For each study, all ribo-seq experiments were added to our dataset except those conducted on mutants designed to alter wildtype translation patterns. The full list of experiments and studies included is given in Supplementary Tables 1 and 2, respectively. The fastq files associated with each experiment were downloaded from Sequence Read Archive[53] or the European Nucleotide Archive[18]. Reads were filtered to exclude reads in which any base had a Phred score below 20. For each remaining read, the number of perfect matches in the *S. cerevisiae* genome were identified, and only unique perfect matches were kept.

It was next necessary to remap the reads such that the position assigned to the read corresponded to the P-site of the translating ribosome, as in previous ribo-seq analyses.[23] The aim of remapping is to shift all read positions such that the triplet periodic signal indicative of active translation overlaps precisely the translated ORF, with the first position of each codon being the highest point of the period. To accomplish this, reads in each experiment were grouped by read length. For each set of reads of a given length, we then counted the number of reads in each of the -50 to +50 positions relative to the start codon accumulated over all annotated genes on Saccharomyces Genome Database (SGD)[54]. The appropriate reading frame to map to is the one with the highest total read count. Within that frame, the start of translation can be identified using the knowledge that there are more reads on the translating ORF than the preceding region. We inferred that the first position in the correct frame with at least 5% of the total reads in the -50 to +50 region corresponds to the location of the p-site of the ribosome translating the start codon. All reads of the given read length were then offset such that this P-site matched the first position of the start codon.

For each read length in each experiment, we then tested whether the reads showed a pattern of strong triplet periodicity that would enable robust translation inference. We counted the number of reads mapping (after P-site remapping) to the first, second, and third position of each codon among annotated genes, requiring at least twice as many reads in the first position than each of the second and third. If a read length failed this test it was excluded from further analysis, and if all read lengths for an experiment failed the experiment itself was excluded. All read lengths from 25 to 35 nucleotides were tested.

Defining Candidate ORFs

To identify a set of translated ORFs, we first constructed a set of candidate ORFs for which translation status could be inferred using ribo-seq data. ORFs were identified on the R64.2.1 genome downloaded from SGD. The initial set of candidates consisted of all possible single-exon reading frames starting with an ATG and ending with a canonical start codon. Among all ORFs that shared a stop codon, all but the longest were discarded. All ORFs that overlapped a canonical gene (annotated as “verified”, “uncharacterized” or “transposable element gene” on SGD) on the same strand were also discarded (including pairs of overlapping canonical genes) unless the ORF shared a stop codon with the canonical gene and the canonical gene was single-exon. An ORF with the same stop codon as an annotated gene on SGD was considered to be that gene.

513 **Translation Calling**

514 In our full dataset of translated ORFs, translation was inferred using ribo-seq data from all experiments
515 that showed robust triplet periodicity among annotated genes (**Supplementary Table 3**). We also
516 generated lists of translated ORFs based only on experiments with or without the drug cycloheximide,
517 only on cells grown in YPD, only on cells grown on SD, and only on cells grown in YPD without
518 cycloheximide (**Supplementary Table 3**). In each case, mapped reads from all eligible experiments were
519 combined into a common pool.

520 Translation was assessed as follows: for each codon in each candidate ORF, the position within the
521 codon with the most reads was noted, if any. The number of times each codon position had the highest
522 read count across the ORF was then counted. We then used the binomial test to calculate a p-value for
523 the null hypothesis that all positions were equally likely, against the alternative that the first position
524 was favored. This p-value is an indicator of the strength of evidence for triplet periodicity favoring the
525 first codon position.

526 To estimate the false positive rate (FDR), we constructed a set of ORFs corresponding to the null
527 hypothesis. For each ORF, we scrambled the ribo-seq reads randomly position by position (not read by
528 read); e.g., if 10 reads mapped to the first base on the actual ORF, a random position in the scrambled
529 ORF was assigned 10 reads, and so on. In this way the read distribution across positions was maintained
530 but the spatial structure was eliminated. We then used the same binomial test on all scrambled ORFs.
531 For every p-value threshold, the FDR can then be calculated as the number of scrambled ORFs with p-
532 value below the threshold divided by the number of actual ORFs with p-values below the threshold. For
533 each list of translated ORFs, the p-value threshold was set to give a 5% FDR among noncanonical ORFs;
534 all ORFs below this threshold were then included in the translated set whether canonical or
535 noncanonical.

536 **Estimating translation rates across different genomic contexts**

537 We assessed the frequency at which nORFs were found to be translated in different genomic contexts,
538 defined by the relation between the nORF and any cORF (ORFs annotated as “verified” or
539 “uncharacterized” on SGD) located on the same transcript, if any. For this analysis, transcripts were
540 taken from the analysis of TIF-seq data in Pelachano et al.[55] An nORF was considered to share a
541 transcript with a cORF f any transcript fully contained both ORFs ; the ORF was then further classified as
542 being in either a uORF or dORF context based on whether it was upstream or downstream of the gene.

Noncanonical ORFs were classified as antisense to a noncanonical gene is they had any overlap on the opposite strand.

Identifying homologous sequences of the *S. cerevisiae* ORF in other *Saccharomyces* genus species

We obtained genomes and genome annotations from seven relatives of S. cerevisiae within the *Saccharomyces* genus: *S. paradoxus* from Liti et al. 2009[56], *S. arboricolus* from Liti et al. 2013[57], *S. jurei* from Naseeb et al. 2018[58], and *S. mikatae*, *S. bayanus var. uvarum*, *S. bayanus var. bayanus,* and *S. kudriavzevii* from Scannell et al. 2011.[34]

Syntenic blocks were constructed between the *S. cerevisiae* genome and the genome of each *Saccharomyces* relative in the following manner: for each gene G_0 in *S. cerevisiae* that had an annotated homolog in a given relative, the closest downstream gene G_1 was identified such that, in the relative, a homolog of G_1 was within 60 kb of a homolog of G_0. The sequence between and including the homologs of G_0 and G_1 were then extracted from the species and an alignment of the syntenic region was generated using MUSCLE 3.8.31.[59]

To confirm that the ORF was matched to a genuine homolog, we extracted the alignment of the S. cerevisiae ORF itself along with a 50 bp flanking region on both ends from the full syntenic alignment. We then realigned this extracted region using the Smith-Waterman algorithm with a match bonus of 5, a mismatch penalty of 4, and a gap penalty of 4. We ran 1000 alignments using the same score system in which the sequence of the comparison species was shuffled at random, reflecting a null hypothesis that the region was not homologous. The proportion of times the alignment of the real sequence scored better than the shuffled ones is a p-value indicating the strength of the null hypothesis against the alternative that the region is homologous. We considered homology confirmed if the p-value was less than 1%.

If a syntenic alignment could not be constructed or if homology of the ORF was not confirmed, we attempted to find the homologous ORF by BLAST as an alternative to the synteny approach. We performed BLASTn on all *S. cerevisiae* single-exon ORFs against all single-exon ORFs in the comparison species. For each reciprocal best matching pair with e-value < 10^{-4}, we took the sequences of the ORFs in both species, together with a 1000 bp flanking region in both ends, and aligned this in the same manner as the syntenic blocks. We then attempted to confirm homology using Smith-Waterman alignment as described above. As BLAST-based alignments offer less confidence than syntenic alignments, we marked all ORFs for which a homolog could be found only using BLAST (**Supplementary Table 3**).

Reading frame conservation

Reading frame conservation is a measure of conservation of codon structure developed by Kellis et al.[13] and used here with some variations. We begin with a pairwise alignment of a genomic region (either a syntenic block or the area around a BLAST hit) containing the *S. cerevisiae* ORF. We identify all ORFs (ATG to stop) in the comparison species across this region. For each ORF in the comparison species, the reading frame conservation is calculated by summing up all points in the alignment where the pair of aligned bases are in the same position within the codon (i.e., both are in either the first, second, or third position) and dividing by the length of the *S. cerevisiae* ORF in nucleotides. The ORF in the comparison species with the highest reading frame conservation is considered the best match, and the reading frame conservation of the S. cerevisiae ORF in relation to each other *Saccharomyces* species is defined as its reading frame conservation with its best match. In addition to the pairwise reading frame conservation of each *S. cerevisiae* ORF in relation to its homologs in all other species, we defined an index of reading frame conservation equal to the average reading frame conservation of the S. cerevisiae ORF against all species in the *Saccharomyces* genus.

Analysis of population data

Variant call file data for 101 *S. cerevisiae* isolates was taken from Peter et al.[28] For every ORF, we considered only isolates for which every position in the ORF was called in calculating nucleotide diversity and pN/pS ratios. To calculate pN/pS ratios, we first obtained expected variant frequencies for each possible majority allele (A, C, G, T) by counting the frequency of minor variants of each type at positions with that majority allele across the entire genome that does not overlap annotated coding sequence. This provides an expected frequency of nonsynonymous and synonymous variants for a given ORF open reading frame that can be obtained by summing the expected variant frequencies across each position in the ORF, as determined by its majority variant. These frequencies were then converted into an expected probability any given single nucleotide variant will be nonsynonymous rather than synonymous.

For testing the pN/ps ratio for any individual ORF, we tested for excess nonsynonymous variants using a binomial test, the nonsynonymous variant probability, and the count of nonsynonymous and synonymous variants. For testing pN/pS among classes of ORFs, we summed up counts of both observed and expected nonsynonymous and synonymous variants across the entire class of ORFs before using the same binomial test.

603 **Analysis of budding yeast genomes**

604 The genomes of 332 budding yeasts were taken from Shen et al. 2018[29]. We applied TBLASTN and
605 BLASTP for each *S. cerevisiae* translated ORF against each genome in this dataset (excluding the
606 Saccharomyces genus). Default settings were used except for setting an e-value threshold of .1; results
607 were then filtered by a stricter e-value threshold as described in each analysis.

608 **Coding Score**

609 The coding score, described by Ruiz-Orera et al. 2014[60], is a measure of how close the hexamer (i.e., the
610 nucleotide sequence of a pair of adjacent codons) frequency of an ORF is to the hexamer coding vs.
611 non-coding sequences. Hexamer frequencies were calculated among all sequences annotated as
612 "verified" or "uncharacterized" ORFs by Saccharomyces Genome Database. Hexamer frequencies were
613 also calculated among all intergenic sequence. As intergenic sequence has no codon structure, hexamer
614 frequencies for intergenic sequence were counted as if read in each possible coding frame. The score
615 was then calculated as described in Ruiz-Orera et al. 2014.

616 **Literature analysis of transient translatome ORFs**

617 For each annotated ORF, we examined all publications listed on SGD as "primary" or "additional"
618 literature for the ORF. If the ORF had a phenotypes in any listed publication, we noted the evidence for
619 the phenotype (**Supplementary Table 5**).

620 **Genetic interaction analysis**

621 Single mutant fitness and genetic interaction data were downloaded from TheCellMap.org[61]. In this
622 dataset, mutants of nonessential genes are full deletions and mutants of essential genes are
623 temperature-sensitive alleles. Transient ORFs were all nonessential. Different temperature-sensitive
624 alleles for the same essential gene were treated separately. For all analyses, we only included genetic
625 interactions with a p-value < 0.05.

626 For each transient ORF or nonessential gene, we calculated how many show at least one genetic
627 interaction value at Ɛ<-.2 or Ɛ<-.35. We then divided this number by the total number of transient ORFs
628 or nonessential genes in the Costanzo et al. 2016[46] genetic interaction dataset to calculate the
629 percentage showing at least one genetic interaction.

Interaction densities were calculated for each ORF by dividing the number of interactions at Ɛ<-.2 either with nonessential or essential genes to the total number of double mutants with nonessential or essential genes, respectively.

We created an unweighted-undirected network from the interactions at Ɛ<-.2 and calculated the degree of each transient ORF. This network was then used to create the subnetwork shown in Figure 7E.

Gene ontology analysis of the interactors of each ORF was conducted with Ontologizer,[62] using Benjamini-Hochberg multiple testing correction and the term-for-term calculation method. The gene association file was downloaded from SGD.

Competing interests

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

Funding

This work was supported by funds provided by the Searle Scholars Program to A.-R.C. and the National Institute of General Medical Sciences of the National Institutes of Health grants R00GM108865 (awarded to A.-R.C.).

References

1. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
2. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).
3. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
4. Erhard, F. *et al.* Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* **15**, 363–366 (2018).
5. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).

6. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* 1–8 (2021) doi:10.1038/s41587-020-00806-2.
7. Jackson, R. *et al.* The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434 (2018).
8. Makarewich, C. A. & Olson, E. N. Mining for Micropeptides. *Trends Cell Biol.* **27**, 685–696 (2017).
9. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**, 595–606 (2015).
10. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228–232 (2017).
11. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105 (2007).
12. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
13. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241 (2003).
14. Ward, L. D. & Kellis, M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science* **337**, 1675–1678 (2012).
15. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111**, 6131–6138 (2014).
16. Oshiro, G. *et al.* Parallel Identification of New Genes in Saccharomyces cerevisiae. *Genome Res.* **12**, 1210–1220 (2002).
17. Blandin, G. *et al.* Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett.* **487**, 31–36 (2000).
18. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28–D31 (2011).

19. Carvunis, A.-R. *et al.* Proto-genes and *de novo* gene birth. *Nature* **487**, 370–374 (2012).
20. Van Oss, S. B. & Carvunis, A.-R. De novo gene birth. *PLoS Genet.* **15**, (2019).
21. Laumont, C. M. *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238 (2016).
22. Yagoub, D. *et al.* Proteogenomic Discovery of a Small, Novel Protein in Yeast Reveals a Strategy for the Detection of Unannotated Short Open Reading Frames. *J. Proteome Res.* **14**, 5038–5047 (2015).
23. Malone, B. *et al.* Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* **45**, 2960–2972 (2017).
24. Ji, Z. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. *Curr. Protoc. Mol. Biol.* **124**, e67 (2018).
25. Calviello, L. & Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet.* **33**, 728–744 (2017).
26. Durand, É. *et al.* Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **29**, 932–943 (2019).
27. Mudge, J. M. *et al.* A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. *bioRxiv* 2021.06.10.447896 (2021) doi:10.1101/2021.06.10.447896.
28. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339 (2018).
29. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545.e20 (2018).
30. Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.* **42**, e134–e134 (2014).

31. Santos, D. A., Shi, L., Tu, B. P. & Weissman, J. S. Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Res.* **47**, 4974–4985 (2019).

32. Duncan, C. D. S. & Mata, J. Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Sci. Rep.* **7**, 10331 (2017).

33. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).

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

35. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genet.* **11**, e1005721 (2015).

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

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

38. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).

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

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

41. Alesso, C. A., Discola, K. F. & Monteiro, G. The gene ICS3 from the yeast Saccharomyces cerevisiae is involved in copper homeostasis dependent on extracellular pH. *Fungal Genet. Biol.* **82**, 43–50 (2015).
42. Giaever, G. & Nislow, C. The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics* **197**, 451–465 (2014).
43. Ho, B., Baryshnikova, A. & Brown, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **6**, 192-205.e3 (2018).
44. Chong, Y. T. *et al.* Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161**, 1413–1424 (2015).
45. Meurer, M. *et al.* Genome-wide C-SWAT library for high-throughput yeast genome tagging. *Nat. Methods* **15**, 598–600 (2018).
46. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
47. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
48. Dujon, B. *et al.* Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371–378 (1994).
49. Basrai, M. A., Hieter, P. & Boeke, J. D. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Res.* **7**, 768–771 (1997).
50. Schlesinger, D. & Elsässer, S. J. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J.* **n/a**,.
51. Xie, C. *et al.* A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife* **8**, (2019).
52. Borneman, A. R. *et al.* Divergence of Transcription Factor Binding Sites Across Related Yeast Species. *Science* **317**, 815–819 (2007).

53. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, on behalf of the I. N. S. D. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).

54. Cherry, J. M. *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).

55. Pelechano, V., Wei, W., Jakob, P. & Steinmetz, L. M. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nat. Protoc.* **9**, 1740–1759 (2014).

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

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

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

59. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

60. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, (2014).

61. Usaj, M. *et al.* TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. *G3 GenesGenomesGenetics* **7**, 1539–1549 (2017).

62. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinforma. Oxf. Engl.* **24**, 1650–1651 (2008).