Article

Cell Systems

# A vast evolutionarily transient translatome contributes to phenotype and fitness

## Highlights

- *S. cerevisiae* translates over 18,000 unannotated open reading frames
- Only a few translated unannotated sequences are conserved genes
- Almost all unannotated coding sequences are likely evolutionarily short lived
- Evolutionarily transient coding sequences provide fitness benefits

## Authors

Aaron Wacholder, Saurin Bipin Parikh, Nelson Castilho Coelho, Omer Acar, Carly Houghton, Lin Chou, Anne-Ruxandra Carvunis

## Correspondence

anc201@pitt.edu

## In brief

Ribosome profiling experiments show widespread translation of unannotated sequences, but it is unknown how many encode biologically significant proteins. We identify thousands of unannotated translated sequences in yeast and find that almost none are evolutionarily conserved. Nevertheless, we detect unannotated proteins by microscopy and find that some provide fitness benefits.

Wacholder et al., 2023, Cell Systems *14*, 363–381
May 17, 2023 
https://doi.org/10.1016/j.cels.2023.04.002

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

# A vast evolutionarily transient translatome contributes to phenotype and fitness

Aaron Wacholder,[1,2] Saurin Bipin Parikh,[1,2,3] Nelson Castilho Coelho,[1,2] Omer Acar,[1,2,4] Carly Houghton,[1,2,4] Lin Chou,[1,2,3] and Anne-Ruxandra Carvunis[1,2,5,*]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[2]Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[3]Integrative Systems Biology Program, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[4]Joint CMU-Pitt PhD Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA 15213, USA
[5]Lead contact
*Correspondence: anc201@pitt.edu
https://doi.org/10.1016/j.cels.2023.04.002

## SUMMARY

Translation is the process by which ribosomes synthesize proteins. Ribosome profiling recently revealed that many short sequences previously thought to be noncoding are pervasively translated. To identify protein-coding genes in this noncanonical translatome, we combine an integrative framework for extremely sensitive ribosome profiling analysis, iRibo, with high-powered selection inferences tailored for short sequences. We construct a reference translatome for *Saccharomyces cerevisiae* comprising 5,400 canonical and almost 19,000 noncanonical translated elements. Only 14 noncanonical elements were evolving under detectable purifying selection. A representative subset of translated elements lacking signatures of selection demonstrated involvement in processes including DNA repair, stress response, and post-transcriptional regulation. Our results suggest that most translated elements are not conserved protein-coding genes and contribute to genotype-phenotype relationships through fast-evolving molecular mechanisms.

## 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.[1–4] These experiments demonstrate that genomes encode not only the "canonical translatome," consisting of the open reading frames (ORFs) identified as protein-coding genes in genome databases such as RefSeq[5] but also a large "noncanonical translatome" consisting of ORFs that are not annotated as genes. Despite the lack of annotation, large-scale studies find that many noncanonical ORFs (nORFs) are translated to express stable proteins and show evidence of association with cellular phenotypes.[6–10] In addition, a handful of previously unannotated coding sequences, identified by RNA-seq or ribo-seq experiments, have now been characterized in depth, revealing that they play key roles in biological pathways and can increase organism fitness.[11–15] However, 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.[16–19] Alternatively, thousands of missing protein-coding genes that contribute to phenotype and fitness could be hidden in the noncanonical translatome.

A common and powerful approach to identifying biologically relevant genomic sequences is to look for evidence of selection.[20–22] Many canonical genes were annotated on the basis of such evidence,[23,24] and this approach has also been applied to nORFs detected by ribo-seq.[25–28] However, in the case of noncanonical translation, the evolutionary analysis is often limited by a lack of sufficient statistical power to confidently detect selection. Most nORFs are much shorter than canonical genes,[7,12,29] thus having fewer genetic variants that can be analyzed for evolutionary inference. As a result, short coding sequences are sometimes missed by genome-wide evolutionary analyses despite long-term evolutionary conservation.[13,30] It is especially challenging to detect selection among nORFs that are evolutionarily novel, as a short evolutionary history also provides less time for enough genetic variants to accumulate the signatures that allow for statistically distinguishing selective from neutral evolution.[31] Several young genes of recent *de novo* origin (i.e., coding genes that evolved from previously non-genic sequences) have been discovered from within the noncanonical translatome.[3,32,33]

In addition to the challenges short ORF length poses for the detection of selection, it also poses challenges for an unequivocal detection of translation in the first place. Microproteins are often missed by most proteomics techniques, although specialized methods have had some success.[9,10,34–36] In ribo-seq data, the most robust evidence of translation comes from a pattern of triplet periodicity in reads corresponding to the progression of

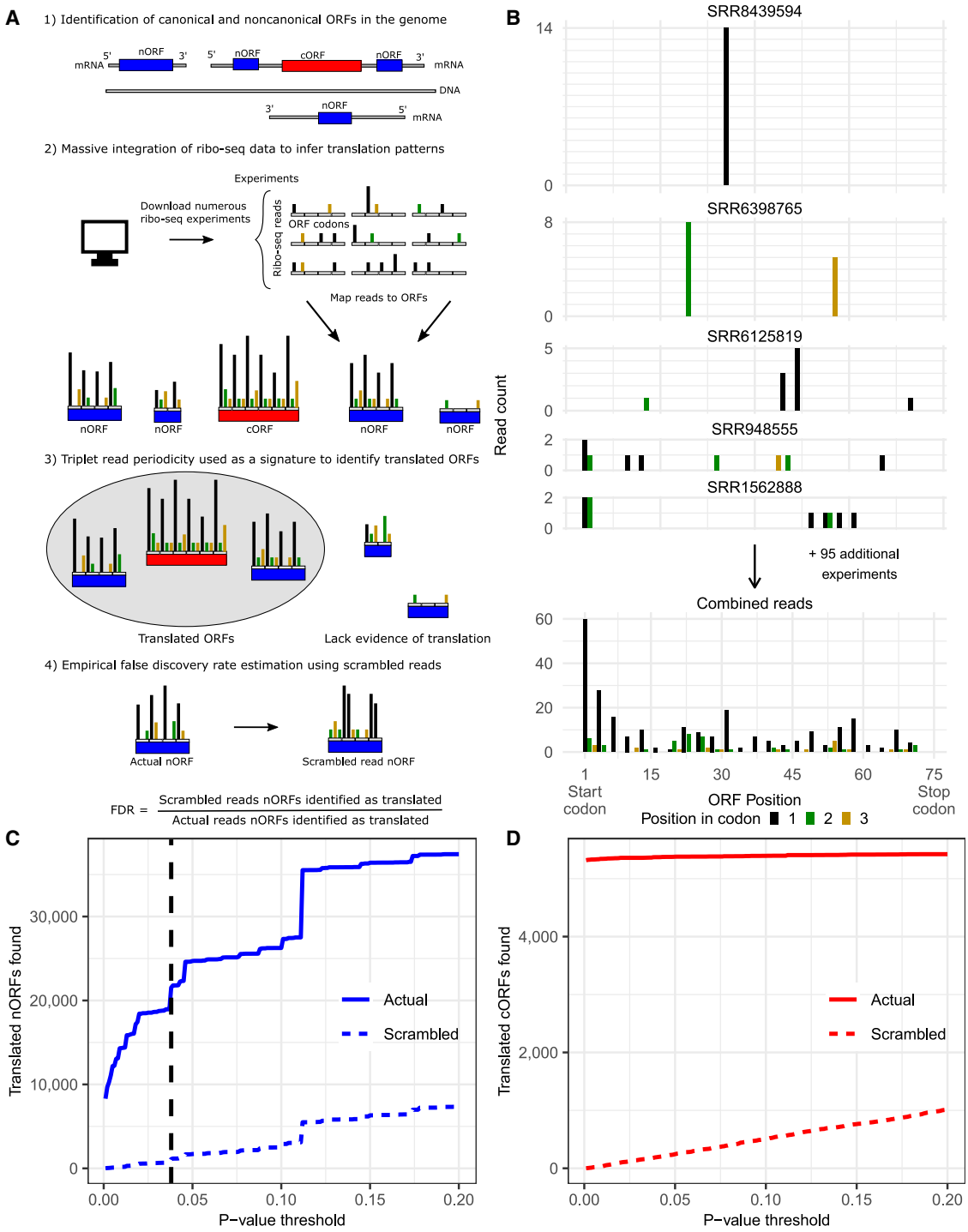



**Figure 1. The iRibo framework enables the detection of thousands of noncanonical translated sequences**

(A) The iRibo framework. (1) Candidate ORFs, both canonical (cORFs; red) and noncanonical ORFs (nORFs; blue), are identified in the genome. (2) Reads aggregated from published datasets are then mapped to these ORFs. (3) Translation is inferred from triplet periodicity of reads. (4) The false discovery rate is estimated by scrambling the ribo-seq reads of each ORF and then assessing periodicity in this scrambled set.

(B) iRibo identifies translated ORFs that are undetectable in any single experiment. Mapped ribo-seq reads across an example nORF located on chromosome II, 604,674–604,748. The top five graphs correspond to five individual experiments with reads mapping to the ORF, whereas the bottom graph includes all reads integrated across all experiments. Reads are colored according to their position on the codon.

*(legend continued on next page)*

the ribosome across codons.[6,37,38] Ribo-seq analysis methods are less capable of detecting translation of short ORFs, as they contain fewer positions to use to establish periodicity.[39] The low expression levels of some nORFs further increases the difficulties in identification.[3,27] Perhaps as a result of these limitations, less than half of the nORFs detected as translated in humans are reproducible across studies.[31]

Here, we designed an approach to increase sensitivity in the detection of both translation and selection among nORFs. We address the challenges in detecting translation through the development of a ribo-seq analysis framework (iRibo) that identifies signatures of translation with high sensitivity and high specificity by integrating data across hundreds of experiments from many published studies. This facilitates the detection of 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 *Saccharomyces cerevisiae* and characterize the biological relevance of nORFs. Using iRibo, we identified ~19,000 nORFs translated at high confidence and established the dependence of noncanonical translation on both genomic context and environmental conditions. Using genomic data both within strains of *S. cerevisiae* and across budding yeast species,[40,41] we identified a handful of undiscovered conserved genes within the yeast noncanonical translatome. However, we find that most of the yeast noncanonical translatome is evolutionarily young and of *de novo* origin, having emerged recently from the noncoding sequence. These young ORFs differ greatly from conserved genes in their length, amino acid composition, and expression level and show no signs of purifying selection. Nevertheless, we report experimental evidence based on fluorescent protein tagging and conditional loss-of-function fitness measurements showing that translation of evolutionarily young nORFs can generate stable protein products and affect cellular phenotypes. We thus propose that much of the noncanonical translatome is composed of neither translational noise nor conserved genes but rather of a distinct class of evolutionarily short-lived coding sequences with important biological implications. This "transient translatome" is larger than, and categorically distinct from, the conserved translatome made mostly of canonical protein-coding genes that have been studied for decades.

## RESULTS

### An integrative approach to defining the translatome

We designed iRibo to detect translation events with high sensitivity and high specificity. High sensitivity is achieved through the integration of ribo-seq data across hundreds of diverse experiments, which provides sufficient read depth for the detection of translated ORFs that are short or weakly expressed. High specificity is achieved through the use of three-nucleotide periodicity as the sole basis for translation inference. Three-nucleotide periodicity corresponds to the progression of the ribosome codon by codon across a transcript, a dynamic unique to translation. Three-nucleotide periodicity is therefore robust against the false inference of translation from other sources of ribo-seq reads.[37,38,42] High specificity is further achieved by controlling confidence levels using an empirical false discovery rate (FDR) approach that relies on minimal modeling assumptions. iRibo consists of four components (Figure 1A). First, a set of "candidate" ORFs that could potentially be translated are identified in the genome. Second, reads from multiple ribo-seq experiments are pooled and mapped to these ORFs. Third, the translation status of each candidate ORF is assessed based on whether the reads mapping to the ORF exhibit a pattern of triplet nucleotide periodicity according to a binomial test. Finally, a list of translated ORFs is constructed with a specified FDR, derived from applying the same translation calling method on a negative control set constructed to exhibit no genuine signatures of translation.

iRibo can be applied to a set of ribo-seq experiments conducted under a single environmental condition to identify ORFs that are translated under that condition. Alternatively, iRibo can be deployed on a broader set of ribo-seq experiments conducted in many different contexts to construct a reference translatome consisting of all elements within a genome with sufficient evidence of translation.

We used iRibo to identify translated ORFs across the *S. cerevisiae* genome (Figure S1). First, we constructed the set of candidate ORFs by 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 was classified either as a canonical ORF (cORF), if it was annotated as "verified," "uncharacterized," or "transposable element" in the *Saccharomyces* Genome Database (SGD)[43] or as an nORF, if it was annotated as "dubious," "pseudogene," or was unannotated. We excluded nORFs that overlap cORFs on the same strand. This process generated a list of 179,441 candidate ORFs: 173,868 nORFs and 5,573 cORFs. We assessed the translation status for candidate ORFs using data from 412 ribo-seq experiments across 42 studies (Tables S1 and S2).

As expected, integrating data from many experiments allowed for the identification of translated ORFs that would otherwise have too few reads in any individual experiment (Figure 1B). Setting a confidence threshold to ensure a 5% FDR among nORFs, we identified 18,953 nORFs (Figure 1C) as translated along with 5,364 cORFs (Figure 1D), for a total of 24,317 ORFs making up the yeast reference translatome (Table S3). This corresponds to an identification rate of 99% for verified cORFs, 77% for uncharacterized cORFs, 37% for dubious nORFs, and only 11% for unannotated nORFs (Figure 2A). Despite the low

(C) iRibo identifies 18,953 translated nORFs at a 5% false discovery rate. The number of nORFs found to be translated using iRibo (y axis) at a range of p value thresholds (x axis) is shown as a solid blue line. Translation calls for a negative control set, constructed by scrambling the actual ribo-seq reads for each nORF, is also plotted (dashed blue line). The dashed vertical line indicates a false discovery rate of 5% among nORFs.
(D) iRibo identifies 5,364 cORFs. 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.

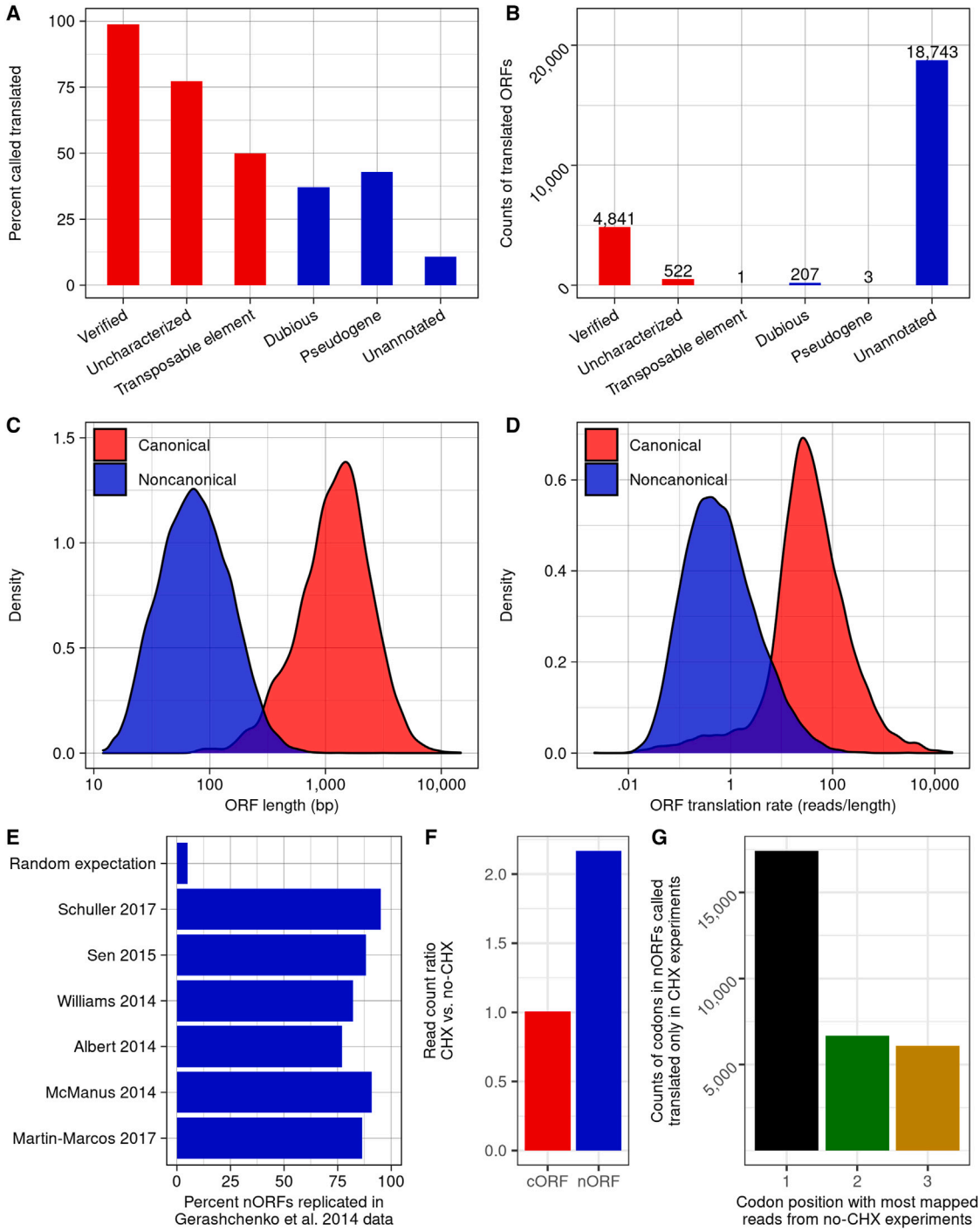



**Figure 2. The noncanonical yeast translatome is larger than the canonical**

(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. The total number of ORFs in each class are: 4,895 verified, 676 uncharacterized, 559 dubious, 7 pseudogene, and 173,302 unannotated.

(B) The number of ORFs of each annotation class that are detected as translated using iRibo.

(C) ORF length distributions for translated cORFs (N = 5,364) and nORFs (N = 18,953).

(D) Distribution of translation rate (in-frame ribo-seq reads per base) for translated cORFs (N = 5,364) and nORFs (N = 18,953).

(E) For six large studies, the proportion of nORFs (N = 18,953) identified using reads from that study that are replicated using reads from the largest study, Gerashchenko and Gladyshev.[44] Random expectation is the proportion that would be expected to replicate by chance.

*(legend continued on next page)*

rate of identified translation, unannotated nORFs make up a large majority of translated sequences (Figure 2B). In general, translated cORFs are much longer (Figure 2C) and translated at much higher rates (Figure 2D) than translated nORFs.

To assess replicability in translation calls for nORFs, we applied iRibo separately to each of the largest individual studies by read count. We then counted, among the nORFs that could be inferred to be translated using only the reads in each study, how many were also found in the largest study, Gerashchenko and Gladyshev.[44] For all studies, at least 75% of detected ORFs were also detected in the largest study (Figure 2E). In general, translation rates among ORFs were highly correlated among independent studies (Figure S2). These observations demonstrate that noncanonical translation patterns are highly reproducible, suggesting that they are driven by regulated biological processes rather than technical artifacts or stochastic ribosome errors.

Many ribo-seq experiments use the translation elongation inhibitor cycloheximide (CHX). This drug is known to influence ribo-seq results in several ways.[44–46] We therefore wished to specifically examine whether the size of the noncanonical translatome we identified could have been artificially inflated by CHX usage. To this aim, we compared translation signatures from experiments with (N = 139) and without (N = 170) CHX, randomly sampling the same number of reads from both groups of experiments. We observed a large enrichment in ribo-seq read counts among nORFs with CHX treatment ($p < 10^{-10}$, Fisher's exact test, Figure 2F), resulting in 56% more nORFs identified as translated ($p < 10^{-10}$, Fisher's exact test). This enrichment may be due to an accumulation of reads in the first 50 codons of ORFs with CHX treatment, which has a greater relative impact on shorter ORFs (Figure S3). The nORFs identified as translated only with CHX treatment nevertheless displayed a strong collective signal of triplet periodicity (i.e., preferential mapping to the first position in the codon) in experiments without CHX treatment when reads were aggregated across all such nORFs (Figure 2G). These results indicate that CHX treatment aids the detection of translation events that also occur but are more difficult to detect without CHX.

## Noncanonical translation patterns depend on genomic and environmental context

We examined to what extent the 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), nORFs antisense to a cORF and located entirely within the bounds of that cORF (antisense full overlap), and nORFs overlapping the boundaries of a cORF on the opposite strand (antisense partial overlap) (Figure 3A). In addition, for nORFs sharing a transcript with an RNA gene, the nORF was classified based on the type of the RNA gene. The transcripts used for these classifications were derived from the TIF-seq (transcript isoform sequencing) data collected by Pelechano et al.,[47] which provide the transcript start and end sites.

Most nonoverlapping translated nORFs were independent (6,373, 52%), and around 47% shared a transcript with a cORF, including 3,512 uORFs and 2,278 dORFs, whereas 1.5% (186) shared a transcript with an annotated RNA gene (Figure 3B). Among antisense nORFs, 73% (4,844) overlapped fully with the opposite-strand gene, whereas 27% (1,760) overlapped partially.

We next calculated the frequency at which candidate nORFs were identified as translated for each genomic context (Figure 3C); for purposes of comparison, we considered only those nORFs fully contained within a TIF-seq transcript. Consistent with previous research,[49] uORFs were translated at significantly higher rates than other classes, with 30% of the considered uORFs found to be translated compared with only 17% of dORFs ($p < 10^{-10}$, Fisher's exact test) and 20% of independent nORFs ($p < 10^{-10}$, Fisher's exact test). nORFs antisense to cORFs and only partially overlapping them were translated at the lowest rate of any context, with a rate of 10% compared with 26% for fully overlapping antisense nORFs ($p < 10^{-10}$, Fisher's exact test).

The amino acid frequencies of the proteins expressed from translated nORFs differ greatly from those of cORFs and depend on the genomic context ($p < 10^{-10}$ for any comparison between cORF amino acid frequencies and nORF frequencies in a given context, chi-square test; Figure 3D). The translation products of nORFs present a large excess of cysteine, phenylalanine, isoleucine, arginine, and tyrosine and deficiency in alanine, asparagine, glutamic acid, and glycine relative to cORFs. Aside from arginine, the amino acids with large excess in nORFs relative to cORFs are all hydrophobic. Amino acid frequencies of nORFs appear to largely reflect underlying DNA sequence composition biases that differ between distinct genomic contexts. Indeed, within each genomic context, amino acid frequencies of translated nORF are generally similar (with less than a 15% difference in frequency) to that of length- and context-matched nORFs that lack evidence of translation, although they do show significant differences ($p < 10^{-10}$ for all contexts, chi-square tests; Figure 3E). The largest differences include a large excess of methionine residues and a deficiency in tryptophan and glycine residues among translated nORFs compared with the untranslated control group.

In addition to the genomic context, we assessed 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) (Table S3). Previous research has reported an increase in detected noncanonical translation events relative to canonical translation events in response to starvation.[1,3] Consistent with these results, more nORFs were identified as translated in minimal than in rich media at equal read counts (Figure 3F). Furthermore, 2,968 nORFs were supported by a significantly higher number of in-frame reads in minimal media than rich media, whereas the converse was true for only 1,265 nORFs

(F) Ratio of total ribo-seq read counts mapping to cORFs or nORFs in experiments with vs. without CHX treatment. Note that the same number of total reads (n = 178,264,204) is sampled from each condition.
(G) Among nORFs identified as translated by iRibo only in the CHX condition (N = 5,944), all codons (N = 30,169) among these nORFs are classed based on which of the three positions in the codon (if any) have the most reads from experiments without CHX.

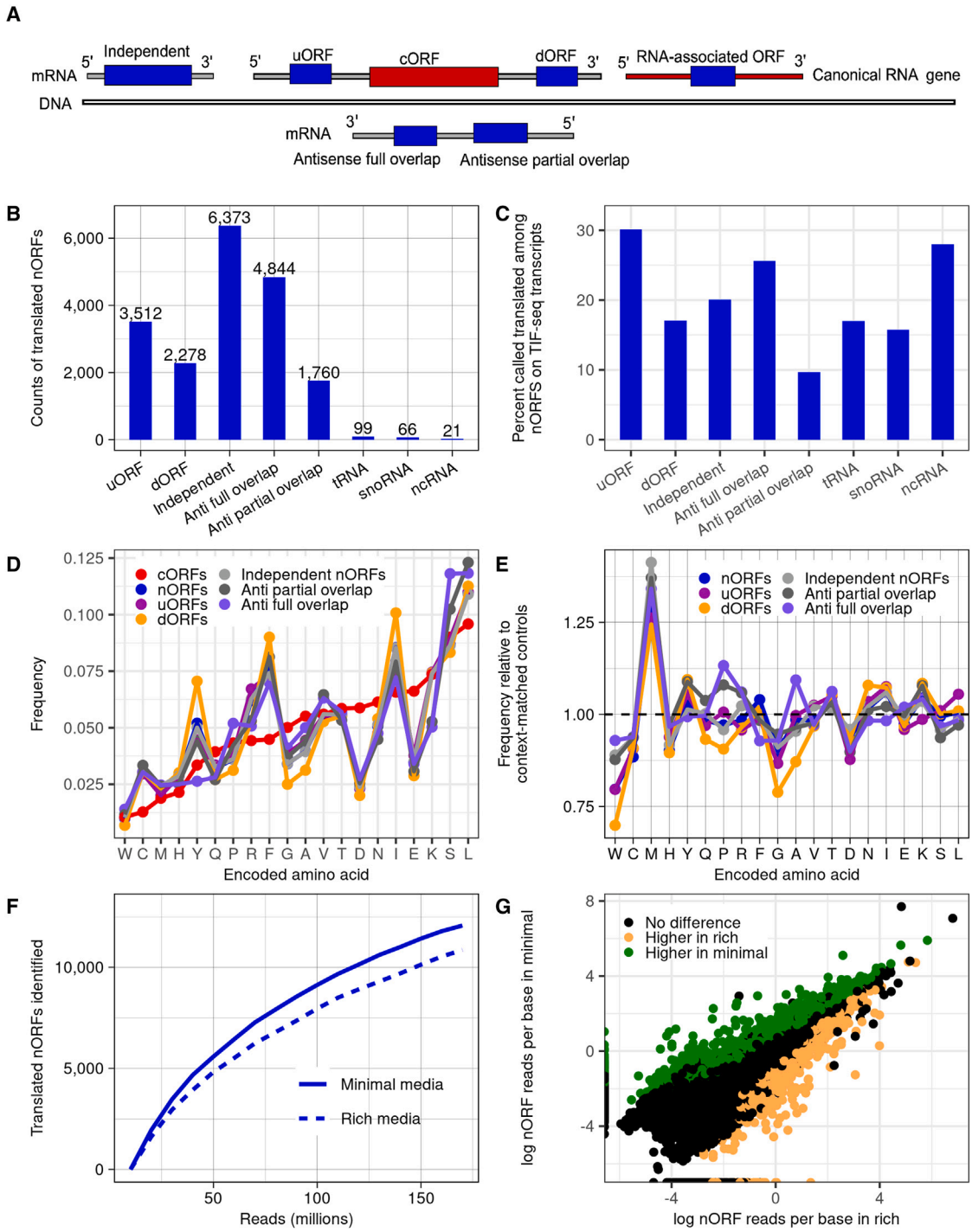



**Figure 3. Noncanonical translation patterns depend on both genomic and environmental context**

(A) Potential genomic contexts for nORFs in relation to nearby canonical genes. Transcripts are defined from published TIF-seq data.[47]

(B) Counts of translated nORFs (N = 18,953) identified by iRibo in each considered genomic context, determined by which elements share a transcript with the nORF and its position within the transcript. For nORFs that share a transcript with RNA genes, the annotation of the RNA gene is specified.

(C) Proportion of nORFs detected as translated by iRibo in each genomic context considered, among nORFs completely covered by a TIF-seq transcript (N = 15,572).

(D) Amino acid composition of translated nORFs differs from that of translated cORFs and depends on the genomic context. Amino acid frequencies among predicted protein products of translated nORFs in each genomic context and of cORFs. The start codon methionine is excluded from frequency estimates.

(E) Amino acid composition of translated nORFs is similar to that of context-matched controls. For each genomic context, the amino acid frequency of translated nORFs relative to that of length-matched untranslated nORFs in that same context. The start codon methionine is excluded from frequency estimates.

*(legend continued on next page)*

(5% FDR, Fisher's exact test with the Benjamini-Hochberg procedure[48]; Figure 3G). These results suggest that starvation conditions may increase noncanonical translation or alternatively that noncanonical translation is less affected by the general translation inhibition that occurs in starvation conditions.[50] Either way, these results support the hypothesis that nORF translation is regulated in response to changing environments.

## Two translatomes, transient and conserved

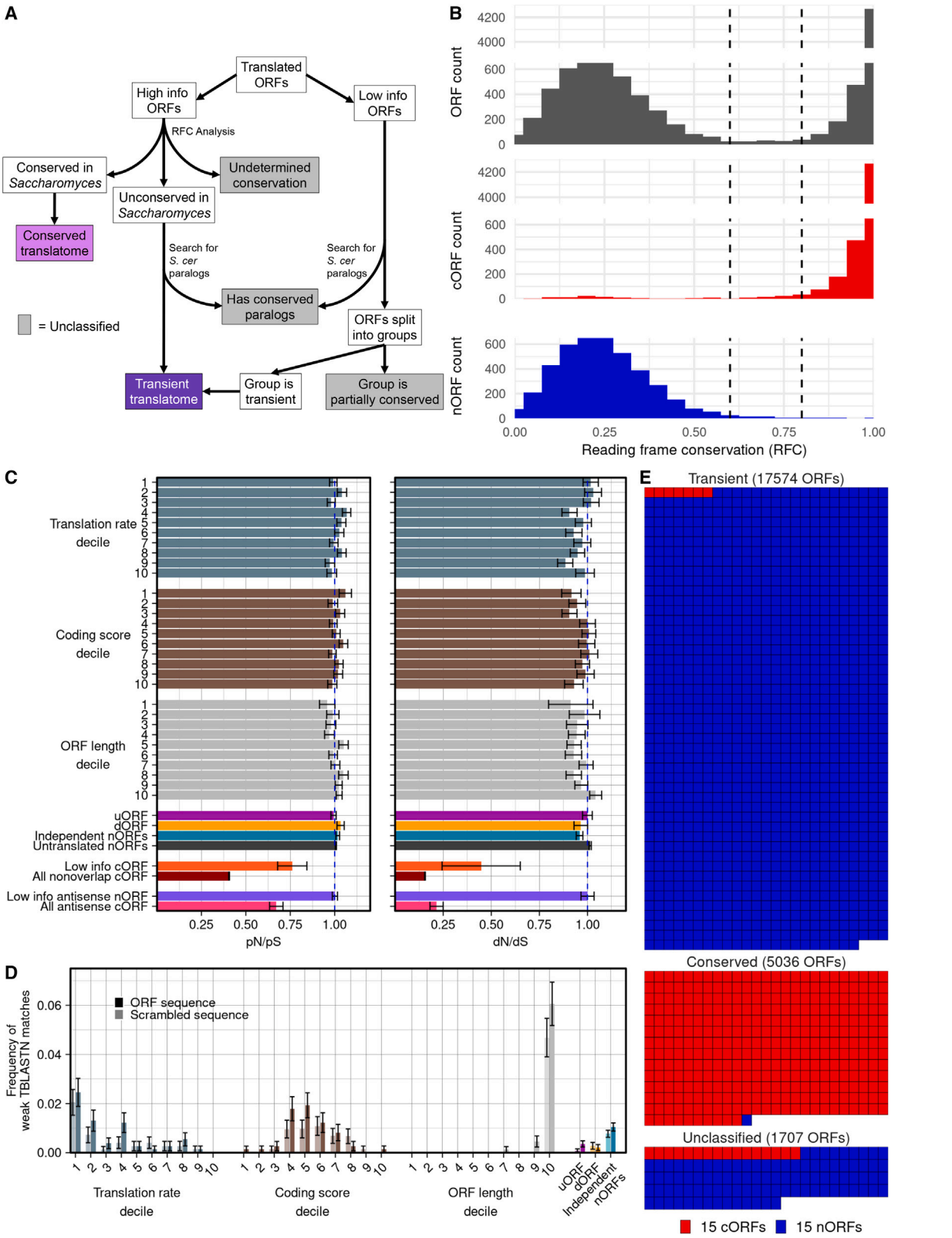
Given the large numbers of nORFs translated in the yeast genome, we next sought to assess the biological relevance of this translation by determining the extent to which these nORFs are evolving under selection. We assessed selection acting on nORFs, and cORFs for the purpose of comparison, across three evolutionary scales. At the population level, we analyzed 1,011 distinct *S. cerevisiae* isolates sequenced by Peter et al.[40] At the species level, we compared *S. cerevisiae* ORFs with their orthologs in the *Saccharomyces* genus, a taxon consisting of *S. cerevisiae* and its close relatives.[51] To detect long-term evolutionary conservation, we looked for homologs of *S. cerevisiae* ORFs among 332 budding yeast genomes (excluding *Saccharomyces*) in the subphylum *Saccharomycotina* collected by Shen et al.[41] 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 the 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 investigated the remaining "low-information" ORFs in groups to quantify collective evidence of selection (Figure 4A). Group-level analysis increases the power to detect the presence of selection but does not enable the identification of the individual ORFs under selection. The high-information set consisted of the ORFs that (1) have homologous DNA sequences in at least four other *Saccharomyces* species and (2) have a median count of nucleotide differences between the *S. cerevisiae* ORF and its orthologs of at least 20. We found that these criteria are sufficient to distinguish ORFs evolving under strong purifying selection (Figure S4). Under this definition, 9,440 translated ORFs that do not overlap a different cORF (henceforth "nonoverlapping ORFs," including 4,248 nORFs and 5,192 cORFs) and 3,022 ORFs that overlap a cORF on the opposite strand ("antisense ORFs," including 2,962 nORFs and 60 cORFs) were placed in the high-information set.

We attempted to detect purifying selection in the high-information set within the *Saccharomyces* genus and within the *Saccharomycotina* subphylum. For the *Saccharomyces* analysis, we adapted reading frame conservation (RFC), a sensitive approach developed by Kellis et al.[20] to distinguish ORFs evolving under selection from other ORFs in the yeast genome. RFC is an index ranging from 0 to 1 that indicates how well the reading frame is conserved between an ORF in a given species (here, *S. cerevisiae*) and its orthologous sequences in related species (other species in the *Saccharomyces* genus). An RFC value of 1 indicates perfect agreement of the reading frame, such that all bases that make up the first nucleotide in a codon in the *S. cerevisiae* ORF also make up the first nucleotide in a codon in each orthologous ORF. An RFC value of 0 indicates that all bases in the *S. cerevisiae* ORF align to bases with a different within-codon position in orthologous ORFs or that the aligned bases exist outside of any ORF. We found a bimodal distribution of RFC among nonoverlapping ORFs in the yeast translatome, considering cORFs and nORFs together: 53.3% have RFC above 0.8 and 45.5% have RFC less than 0.6, with only 1.2% of ORFs intermediate between these values (Figure 4B). The bimodal distribution of RFC among translated ORFs is similar to the bimodal distribution observed among all candidate ORFs, regardless of the translation status (Figure S5A), as observed previously by Kellis et al.[20] The modes of distribution largely correspond to annotation status, with 96.7% of cORFs having an RFC $>$ 0.8 and 98.5% of nORFs having an RFC $<$ 0.6. This pattern holds when evaluated only in the last 100 bp of ORFs, suggesting that it is not affected by the potential incorrect inference of nORF start positions (Figure S5B). The clean separation between well-conserved and poorly conserved ORFs indicates that most high-information ORFs can be straightforwardly classified into one of the two groups, and thus, nearly all high-information nonoverlapping nORFs can be placed in the poorly conserved class. High RFC among antisense ORFs does not demonstrate selection on the ORF itself, as it might be caused by selective constraints on the opposite-strand gene, but low RFC still indicates a lack of purifying selection. A majority of antisense translated nORFs (64.1%) have an RFC $<$ 0.6, indicating that most are not preserved by selection across the genus (Figure S5C). Overall, we find no evidence for purifying selection acting on nORFs on a large scale.

In light of the general correspondence between annotation and conservation, the exceptions are of interest: 110 cORFs had an RFC $<$ 0.6, and 13 nonoverlapping unannotated nORFs had an RFC $>$ 0.8. To further assess conservation among these two sets of ORFs, we performed a BLAST analysis (using both BLASTP and TBLASTN with default parameters) to search for homologs of each ORF among the budding yeast genomes assembled by Shen et al.[41] Among the 110 cORFs with low RFC, 101 also had no detected homology to other *S. cerevisiae* genes or

(F) More nORFs are identified as translated in minimal than rich media. Number of translated nORFs identified (y axis) for experiments on yeast grown in either minimal (SD, solid line) or rich media (YPD, dashed line) at a range of read depths (x axis). For each read depth, reads are sampled at random from experiments in each condition.

(G) For each nORF called translated by iRibo in minimal media (SD), rich media (YPD), or both (N = 15,563), the log reads per base in each condition is indicated. Total read count in each condition was held constant (n = 170 million) by randomly sampling reads until the target count was reached. nORFs with significantly more reads in one condition than the other are colored, green for SD, and brown for YPD. Lists of nORFs with significantly different translation rates were obtained as follows: p values for differential translation of each nORF were calculated from Fisher's exact test on in-frame ribo-seq reads mapping to the ORF in each condition and a 5% FDR was set using the Benjamini-Hochberg approach.[48] An nORF had to be detected as translated in a condition by iRibo to be identified as more highly translated in that condition.




*(legend on next page)*

any budding yeast genome outside of *Saccharomyces*, indicating that these are likely annotated ORFs of the recent *de novo* origin. For the 13 nORFs with high RFC, several additional lines of evidence suggest that these are indeed evolving under purifying selection (Table S4). For nine of the thirteen, we identified a homolog among budding yeast genomes outside of the *Saccharomyces* genus by either a BLASTP or TBLASTN search. The existence of a homolog in a distantly related species indicates that the ORF existed in the common ancestor of *S. cerevisiae* and that distant species, implying long-term preservation of the ORF by purifying selection in both lineages. We also performed the pN/pS analysis for each ORF on *S. cerevisiae* isolates and dN/dS analysis for each ORF among the *Saccharomyces* genus species (Table S4). A pN/pS or dN/dS ratio significantly below 1 indicates purifying selection on the ORF amino acid sequence among *S. cerevisiae* strains or among *Saccharomyces* genus species, respectively, whereas a ratio above 1 indicates positive selection. By these measures, two ORFs showed significant evidence of purifying selection by pN/pS and three by dN/dS (Table S4). Thus, a small number of nORFs appear to have beneficial biological roles preserved by selection.

We next assessed selection among the full set of nORFs (both high and low information) at the subphylum scale, searching for the addition nORFs that exhibited long-term conservation and thus purifying selection. Toward this end, we searched for distant homologs of all translated nonoverlapping *S. cerevisiae* nORFs using TBLASTN against budding yeast genomes in the *Saccharomycotina* subphylum, excluding species in the *Saccharomyces* genus. After excluding matches that appeared non-genic or pseudo-genic (Figure S6), we identified a single additional nORF with both distant TBLASTN matches and recent signatures of purifying selection (dN/dS = 0.5, p = 0.039 for the test of difference from 1.0): YBR012C, annotated as dubious on SGD. Thus, combining the 13 nORFs that appeared conserved by RFC analysis and the single additional nORF found using TBLASTN, we identified 14 translated nORFs that show evidence of preservation by purifying selection (Table S4).

To analyze collective evidence of selection among low-information ORFs, we first divided low-information nonoverlapping nORFs (7,855 nORFs, after excluding those with homology to conserved *S. cerevisiae* cORFs) according to properties that we expected to be potentially associated with selection: rate of translation (as measured by ribo-seq reads mapped to the first position within codons divided by the length of the ORF), coding score[28,52] (a measure of sequence similarity to annotated coding sequences), ORF length, and genomic context. For each group, we calculated the pN/pS ratio among 1,011 *S. cerevisiae* isolates[40] and the dN/dS ratio based on alignments of the *S. cerevisiae* ORFs with their orthologous DNA sequence in *S. paradoxus*. We also analyzed low-information nonoverlapping cORFs (22 cORFs) in the same manner. For low-information antisense nORFs (3,642 nORFs; only 2 cORFs fell in this category and were not analyzed), we calculated the pN/pS and dN/dS ratios restricted to substitutions that were synonymous on the opposite-strand cORF.[53,54] Unlike the RFC, dN/dS, and pN/pS analyses conducted above on individual high-information ORFs, these analyses were conducted by aggregating substitutions among all low-information ORFs in each group to assess evidence for selection (i.e., a ratio significantly different from 1) within the group as a whole. We expected that if low-information nORFs were evolving under selection, then more highly translated ORFs, longer ORFs, and ORFs with coding scores more similar to conserved genes would be enriched in biologically relevant nORFs and thus show stronger signatures of selection. Low-information nonoverlapping cORFs did show collective pN/pS and dN/dS ratios significantly below 1, indicating that some ORFs in this group are evolving under purifying selection (Figure 4C; Table S5). By contrast, for all groups of low-information nORFs examined, we observed no significant difference in the pN/pS or dN/dS ratio from 1, providing no evidence for either purifying or positive selection (Figure 4C; Table S5).

Finally, we assessed collective evidence of long-term evolutionary conservation in each group. To do this, we calculated the frequency of weak TBLASTN matches (e values between $10^{-4}$ and 0.05, above our threshold for homology detection at the individual level) of ORFs in each group to the *Saccharomycotina* subphylum genomes outside of *Saccharomyces* compared

**Figure 4. Two distinct translatomes: transient and conserved**

(A) Selection inference analyses conducted on low-information and high-information ORFs to classify them as evolutionarily conserved, transient, or unclassified.

(B) A bimodal distribution of reading frame conservation (RFC) among high-information translated ORFs. The distribution of RFC (x axis), indicating how well reading frame of the ORF is conserved in the *Saccharomyces* genus, is shown for all translated high-information ORFs (top, N = 9,440), only cORFs (middle, N = 5,192), and only nORFs (bottom, N = 4,248). See STAR Methods for details. Dashed lines separate RFC < 0.6 and RFC > 0.8, the thresholds used to distinguish ORFs preserved or not preserved by selection.

(C) No evidence of purifying selection acting on low-information nORFs. pN/pS, and dN/dS ratios are shown for each group of ORFs. Low-information nonoverlapping nORFs that lack a conserved homolog (N = 7,855) are divided into deciles of translation rate (in-frame ribo-seq reads per base), coding score, or ORF length and into three genomic contexts. Untranslated nORFs (N = 60,113) are the set of all nonoverlapping nORFs in the genome not called as translated by iRibo. Low-information nonoverlapping cORFs (N = 22) are assembled into a single group, with the set of all nonoverlapping cORFs (N = 5,364) shown for comparison. Low-information antisense nORFs (N = 6,604) were also assembled into a single group, with the set of all antisense cORFs (N = 62) shown for comparison. pN/pS is calculated from variation at each ORF codon among 1,011 *S. cerevisiae* isolates.[40] dN/dS is calculated among all codons that share the same frame between *S. cerevisiae* ORFs and aligned orthologous ORFs in *S. paradoxus*. Note that the displayed pN/pS and dN/dS values are not averages of these ratios among ORFs. Rather, synonymous and nonsynonymous variants among all ORFs in each class are counted, and a single ratio is calculated from the summed counts. Error bars indicate standard errors estimated from bootstrapping. The dashed blue line indicates a ratio of one, the expected ratio under neutral evolution.

(D) No evidence of distant homology for low-information nORFs. The frequency of nORFs with weak TBLASTN matches ($10^{-4}$ < e value < 0.05, N = 49) in each group of nORFs (dark bars) and negative controls (light bars, N = 49) consisting of the sequences of the nORFs of each group randomly scrambled. Error bars indicate standard errors estimated from bootstrapping.

(E) ORFs that are translated yet evolutionarily transient make up 72% of the yeast reference translatome. The components of the translatome (transient, conserved, unclassified) are represented with area proportional to frequency. Each box represents sets of 15 ORFs.

with a negative control set consisting of scrambled sequences of the ORFs in each group. Applying this strategy to the full set of 362 nonoverlapping cORFs that lacked TBLASTN matches outside *Saccharomyces* at the e value < $10^{-4}$ level, we found a large excess of weak matches relative to controls ($p = 0.0001$, Fisher's exact test; Figure S7), 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 nonoverlapping nORF group and scrambled controls (Figure 4D) nor among nonoverlapping nORFs overall ($p > 0.05$, Fisher's exact test). 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 without distant homology. However, our TBLASTN, dN/dS, and pN/pS analyses altogether 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 neither evidence of selection nor homology to conserved ORFs in our high-information and low-information sets as the "transient translatome." The transient translatome designation indicates membership in a set of ORFs that are expected to exist in the genome for only a short time on an evolutionary scale, although we cannot be certain that any particular translated ORF that currently exists in the yeast genome will be rapidly lost. The transient translatome includes 4,051 nonoverlapping and 1,923 antisense nORFs identified as not preserved by selection using RFC analyses and having no conserved homologs, along with 86 nonoverlapping and 15 antisense cORFs (total 101) matching the same criteria. Also included are 7,855 nonoverlapping and 3,644 antisense 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 17,574 ORFs that are translated yet likely evolutionarily transient makes up 72% of the yeast reference translatome (Figure 4E).

## Transient cORFs are representative of the transient translatome overall

By general theory and practice in evolutionary genomics, the lack of a selective signal suggests that the transient translatome does not meaningfully contribute to fitness.[55] Nevertheless, 101 cORFs belong to the transient set, suggesting that some transient ORFs do have phenotypes. To assess whether these cORFs are representative of the transient translatome overall, we compared their evolutionary and sequence properties with those of transient dubious nORFs (annotated but presumed nonfunctional) and transient unannotated nORFs. We found transient cORFs, transient dubious nORFs, and transient unannotated nORF to all have pN/pS ratios indistinguishable from 1.0 (Figure S8A), providing no evidence for purifying selection. Similarly, the average nucleotide diversity (mean number of nucleotide differences per site between pairs of isolates) of transient cORFs was indistinguishable from that of transient nORFs or untranslated controls and much higher than that of conserved cORFs (Figure S8B). In addition, no class of transient ORFs showed differences from each other in RFC between *S. cerevisiae* and *S. paradoxus* (Figure S8C), the rate of translation (Figure S8D), or the coding score (Figure S8E).

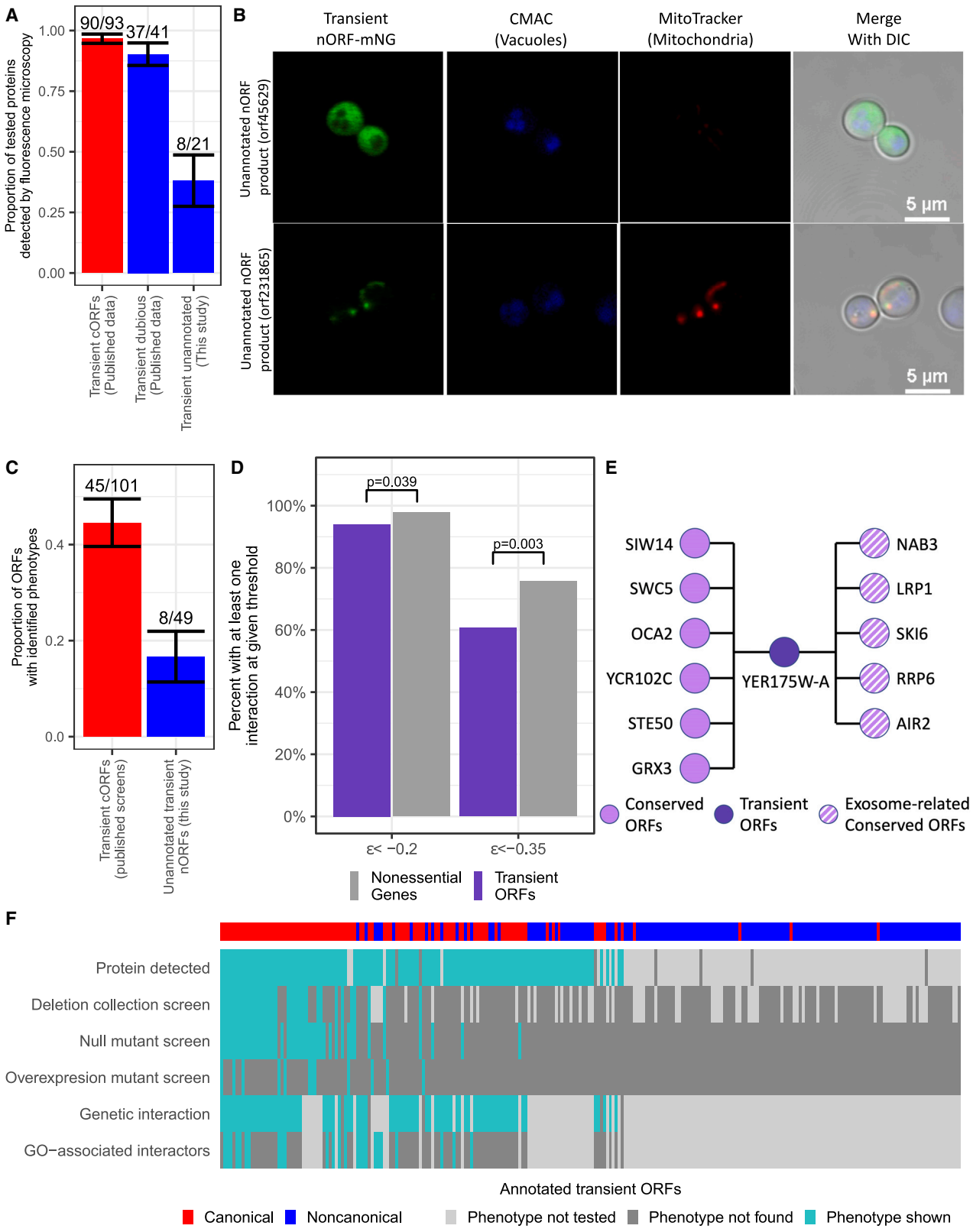
The only distinguishing property between classes of transient ORFs was their length: annotated transient cORFs and transient dubious nORFs are much longer on average than unannotated transient nORFs (Figure S8F). This is a consequence of the history of annotation of the *S. cerevisiae* genome, where a length threshold of 300 nt was set for the annotation of ORFs.[56,57] 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). Given that transient nORFs resemble transient cORFs in all respects besides length, we hypothesized that numerous never-studied transient nORFs are just as likely to have phenotypes as transient cORFs.

## Transient ORFs are detected in the cell and mediate diverse phenotypes

To gain further insights into the potential biological roles of transient ORFs, we examined published reports about annotated ORFs (transient cORFs and transient dubious nORFs) in the *S. cerevisiae* experimental literature and performed additional experiments to investigate transient unannotated nORFs. We examined whether transient ORF products could be detected experimentally, whether they affect phenotypes, and whether they interact with specific biological pathways.

We first assessed whether the proteins encoded by transient ORFs can be detected in the cell. We examined the CYCLoPs database,[58,59] the C-SWAT tagging library,[60] and the YeastRGB database,[61] which contain collections of fluorescently tagged proteins expressed from their native promoters and terminators, including both cORFs and dubious nORFs. Together, these studies detected the expression of a fluorescent protein product for 90 of the 93 (97%) transient cORFs tested, along with 37 of the 41 (90%) transient dubious nORFs tested (Figure 5A). For comparisons, we C-terminally tagged 21 highly expressed unannotated transient nORFs with mNeonGreen at their endogenous locus and examined their expression using microscopy. We detected 8 of the 21 tagged nORF proteins (38%) (Figures 5A, 5B, and S9). Thus, the translation of tagged proteins can be detected for both annotated and unannotated transient ORFs.

We next examined the evidence that transient ORFs affect phenotype. Five transient cORFs have been studied in depth. Two of these, *MDF1*[62] and YBR196C-A,[63] have been previously described as having emerged *de novo* from non-genic sequences. *MDF1* inhibits the mating pathway in favor of vegetative growth,[62,64] and YBR196C-A is an endoplasmic reticulum-located transmembrane protein whose expression is beneficial under nutrient limitations.[65] The remaining three have been experimentally characterized, although their evolutionary properties were not analyzed in the corresponding studies: *HUR1* plays a role in non-homologous end-joining DNA repair[66], YPR096C regulates the translation of *PGM2*[67], and *ICS3* is involved in copper homeostasis.[68] These cases demonstrate that some transient ORFs do the phenotypes.




*(legend on next page)*

To determine whether transient cORFs that are not well described also affect phenotype, we examined all literature listed as associated with the ORF on SGD. Many of these transient cORFs have direct evidence of the phenotype (Table S6). Of the 101 transient cORFs, 45 were reported to have deletion mutant phenotypes (i.e., a change in phenotype observed when the ORF is deleted) and 12 to have overexpression phenotypes. Overall, we found phenotypes reported in the literature for 50 of the 101 transient cORFs (50%).

As unannotated transient nORFs have not been systematically investigated for phenotype, we sought to experimentally determine whether these ORFs too might have deletion mutant phenotypes. We thus conducted a deletion mutant screen of 49 unannotated transient nORFs selected for high translation rate and to avoid intersecting cORFs, annotated ncRNAs, or promoters (200 bp upstream of canonical genes). We fully deleted the nORF using homologous recombination, and each strain was assayed for colony growth in seven conditions. Eight nORF deletion mutant strains showed deleterious phenotypes in at least one condition at a 5% FDR (Figure 5C; Table S7). Thus, the loss of transient nORFs, as with cORFs, can affect the phenotype despite the lack of evolutionary conservation.

To begin to understand the specific biological processes in which transient ORFs might be involved, we leveraged the large yeast genetic interaction network assembled in the study of Costanzo et al.[69] This dataset includes 75 nonoverlapping transient cORFs and 9 nonoverlapping dubious transient nORFs. Genetic interaction strength, $\varepsilon$, measures the difference between the observed fitness of a strain in which two genes are deleted and the expected fitness, given the fitness of the two single-gene deletion strains; a negative value of high magnitude suggests that the two mutated genes are involved in related processes. Of the 84 transient ORFs in the dataset, 79 (94%) have at least one negative genetic interaction at the high-stringency cutoff defined by Costanzo et al.[69] ($\varepsilon < -0.2$ and p value $< 0.05$) and 51 (61%) have synthetic lethal interactions ($\varepsilon < -0.35$ and p value $< 0.05$) as defined in that study (Figure 5D). This was only a slightly lower rate than that for conserved nonessential ORFs, 98% of which had negative interactions at the high-stringency cutoff and 76% of which had synthetic lethal interactions. At the high-stringency threshold, 27 transient ORFs were found to interact with groups of related genes enriched in specific gene ontology (GO) terms (5% FDR; Table S8). 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 5E), indicating likely involvement in CUT degradation or a closely related post-transcriptional regulation pathway. Other enrichments included diverse processes such as "mating projection tip" or "Golgi sub-compartment." In contrast, when we applied the GO enrichment analysis to the full set of genes that interact with any transient ORF, no significant enrichment was observed. These results suggest that transient ORFs in general do not participate in one shared biological process but rather are involved in a wide variety of cellular processes.

Overall, we uncovered evidence that 131 of the 250 (53%) annotated transient ORFs have at least one indicator of biological relevance (detection of a protein product, a reported phenotype in a screen, or a genetic interaction in the Costanzo et al.[69] network) (Figure 5F). In addition, we demonstrate that unannotated transient ORFs encode proteins that can be detected in the cell (38% of those tested in this study) and influence cellular fitness when deleted (17% of those tested in this study). Given that this class has received almost no study compared with the great number of experiments that have been conducted on cORFs, the number of transient ORFs with biological relevance may be substantially larger than that which has been annotated.

A limitation of much of the experimental evidence available on deletion mutant phenotypes is that most deletion mutants and genetic interaction screens are based on a full gene replacement strategy in which the entire ORF is lost, leaving the possibility that some deletion phenotypes could be caused by the loss of a ncRNA or a DNA regulatory element located at the same position as the ORF rather than the loss of the ORF translation (Figure 6A). To examine this possibility, we constructed a set of strains where the ORF start codon ATG was replaced with an AAG codon while keeping the rest of the ORF intact. This set included three transient cORFs that have previously been characterized on the basis of overexpression or full deletion mutants, *ICS3*,[68] YPR096C,[67] and YBR196C-A,[65] along with four transient nORFs that showed strong deleterious phenotypes in our full ORF deletion screen

**Figure 5. Transient nORFs and cORFs can be detected in the cell and exhibit phenotypes**

(A) Transient ORFs are detected by fluorescent microscopy. For cORFs or dubious nORFs, the proportion of proteins expressed by transient ORFs detected in the C-SWAT,[60] CYCLoPs,[59] or YeastRGB[61] microscopy datasets out of those tested. For unannotated transient nORFs, the proportion detected by mNeonGreen tagging in this study. Error bars indicate standard error of the proportion.

(B) Tagged unannotated transient nORFs show varied subcellular localizations. Microscopy images of unannotated transient nORFs taken at 100×. Left panel shows the expression of the nORFs tagged with mNeonGreen, middle panels the dyes CMAC blue and MitoTracker red for mitochondria and vacuoles identification, respectively, and the right panel the merge of all the above channels with DIC. Top panel show the nORF (orf45629) with a cytosolic expression and the bottom panel the nORF (orf231865) with expression localizing to the mitochondria. Each image is representative of around 100 individual cells. Scale bars, 5 μm.

(C) Loss of transient nORFs can affect phenotype despite lack of evolutionary conservation. The proportion of deletion mutants with reported loss-of-function phenotypes in two groups: transient cORFs in published deletion mutant screens, and transient nORFs assayed in this study. Reported phenotypes in published data were taken from the literature associated with each ORF on SGD. In this study, deleterious deletion mutant phenotypes were identified from a high-throughput colony fitness screen in six stress conditions using a 5% FDR threshold.

(D) Transient ORFs engage in epistatic relationships. The percent of transient ORFs (N = 84) and nonessential genes (N = 4,681) with at least one genetic interaction at the given threshold are shown. Differences between groups were tested using Fisher's exact test.

(E) Genetic interactions of the transient ORF YER175W-A. Five interactors are related to exosome (striped circles).

(F) Presence of phenotypes among annotated transient ORFs (N = 250). "Protein detected" indicates that the ORF product was found in either the C-SWAT or CYCLoPs database. Phenotypes of deletion collection, deletion, and overexpression screens were taken from reported findings in the yeast experimental literature (Table S6). "Genetic interaction" indicates a statistically significant genetic interaction with $\varepsilon < -0.2$, and "GO-associated interactors" indicates a GO enrichment was found among significant interactors at 5% FDR.

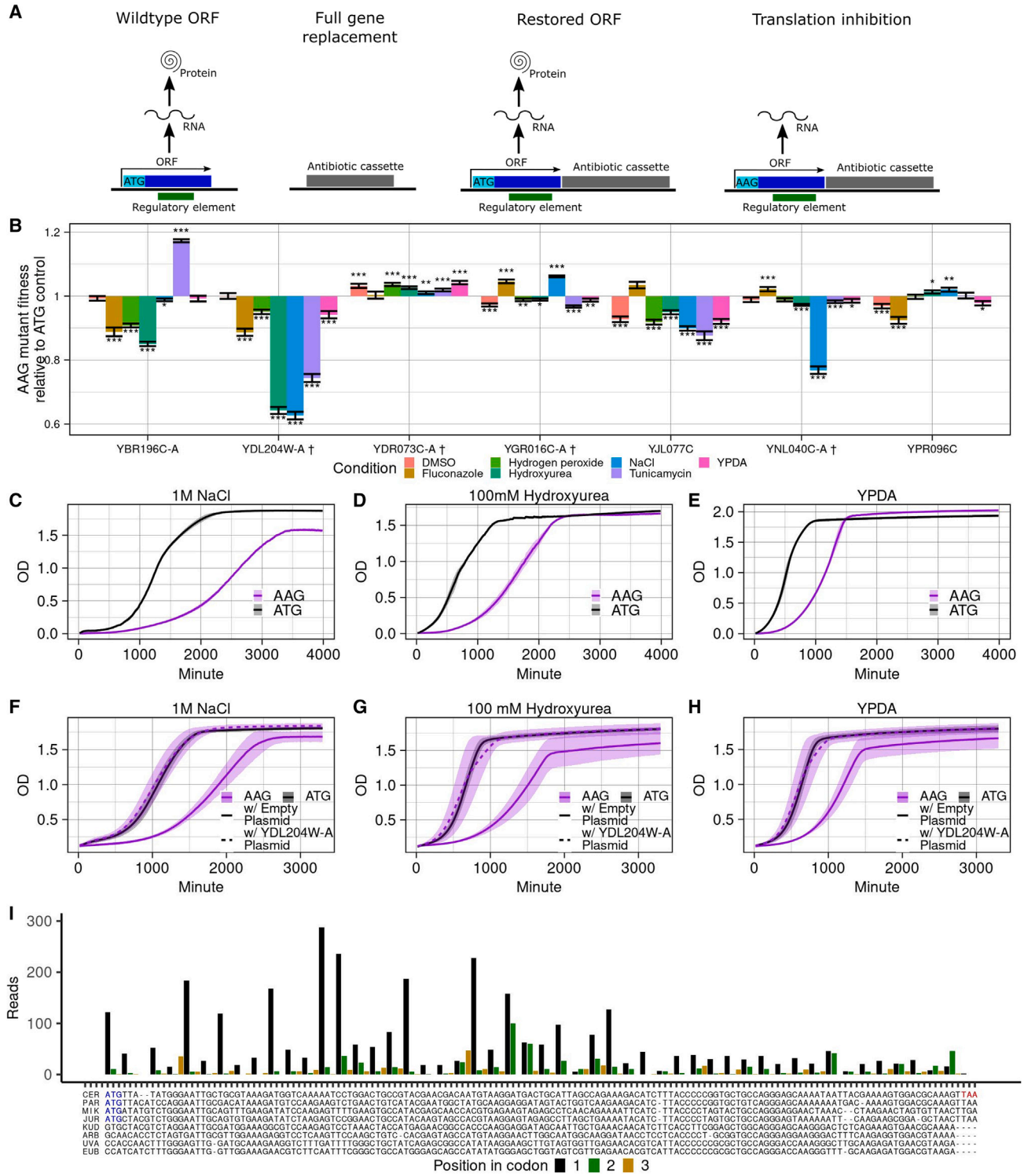



**Figure 6. Translation inhibition of transient ORFs causes strong phenotypes**

(A) A two-step strategy for inhibiting nORF translation. An ORF may overlap a DNA regulatory element or an RNA with a noncoding function (wild-type ORF), both of which are disrupted in a gene replacement strategy in addition to the loss of translation (full gene replacement). This creates ambiguity in interpreting comparisons between deletion mutants and wild-type strains. Following a deletion screen using gene replacement, we used a second round of homologous recombination to restore either the full ORF (restored ORF) or an ORF with its start codon mutated from ATG to AAG (translation inhibition). As these mutants differ only by this single base, the specific effects of translation inhibition can be inferred.

*(legend continued on next page)*

(Table S9; *HUR1* and *MDF1* were not tested because they overlap other cORFs). Each deletion strain was tested in seven environmental conditions. The single nucleotide ATG→AAG mutation caused a significantly reduced colony size for all three transient cORFs tested and for three of the four transient nORFs tested in at least one condition (Figure 6B). We gave these three nORFs systematic names YDL204W-A, YGR016C-A, and YNL040C-A. The remaining nORF, YDR073C-A, showed a weak beneficial phenotype from the ATG→AAG mutation in some conditions, as did two other nORFs, YGR016C-A and YNL040C-A. The largest growth reductions were observed from disabling translation in YDL204W-A: this strain reached only 64% of wild-type growth in hydroxyurea and 63% in high salt concentrations, with a smaller reduction to 94% growth in rich media (YPDA). These growth defects were also observed in a liquid growth setting (Figures 6C–6E). To confirm that these phenotypes were caused by loss of the YDL204W-A protein rather than *cis* effects at the locus, we expressed the intact YDL204W-A ORF from a plasmid in the ATG→AAG mutant strain. Plasmid expression of the ORF fully restored the wild-type phenotype in the mutant strains (Figures 6F–6H), providing further evidence that blocking YDL204W-A translation causes a loss-of-function phenotype mediated by loss of the encoded protein.

In our translation dataset, YDL204W-A has a translation rate at the top percentile among transient ORFs (Figure 6I), higher than 10% of cORFs. Comparing its sequence with the homologous region of other *Saccharomyces* genus species, only *S. paradoxus* and *S. mikatae* have a homologous start codon, but a 2 bp insertion in *S. cerevisiae* results in a frameshift such that little of the ORF is shared in any other species (Figure 6I); thus, this ORF has a RFC score of only 0.2 (Table 1). The other transient ORFs with phenotypes induced by an ATG→AAG mutation also showed no signs of selection (Table 1). Thus, our results exemplify the potential for unannotated coding sequences with no evident evolutionary conservation to affect cellular phenotypes and fitness.

## DISCUSSION

Since the advent of ribo-seq, 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 have remained elusive. To facilitate the study of this noncanonical translatome, we developed iRibo, a framework for integrating ribo-seq data to sensitively detect ORF translation across a variety of environmental conditions. The iRibo framework can be applied to any species and set of candidate ORFs of interest. Here, we deployed iRibo to map a high-confidence yeast reference translatome almost 5 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 designed iRibo to be highly sensitive at detecting patterns of triplet periodicity through the genome, but there are some limitations to our strategy. We focused exclusively on ORFs with AUG start codons and therefore missed the non-AUG codons that are sometimes used as starts.[71] Similarly, we did not consider ORFs overlapping canonical genes in a different frame on the same strand, although some such nORFs are known to be translated.[72,73] Finally, candidate ORFs were selected as the longest ORF in any reading frame, which means the true boundaries of identified ORFs could be shorter than described. We expect these limitations to cause an underestimation of the number of translated nORFs, suggesting that the true count is even larger than that identified here.

We used the iRibo yeast reference translatome to address a fundamental question: to what extent does the noncanonical translatome 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,[34] although it was not annotated on the SGD at the time of our analysis. Thus, even a genome as well studied as *S. cerevisiae*'s contains undiscovered conserved genes, likely missed in prior analyses due to difficulties in analyzing ORFs of a short length. These 14 nORFs are, however, the exception: the great majority of translated nORF show no signatures of selection, comprising a large pool of evolutionarily transient translated sequences.

The yeast genome thus encodes two translatomes, one conserved and one transient. The conserved translatome consists of coding sequences that are preserved by strong purifying selection and usually have a long evolutionary history. They tend to be relatively long, well expressed, and with sequence properties highly distinct from noncoding sequences. The transient translatome, by contrast, is evolutionarily young, of recent *de novo* origin from previously noncoding sequences and still similar to noncoding sequences in nucleotide composition. Evolving in the absence of strong purifying selection, transient

(B) Inhibiting translation of transient ORFs triggers colony growth phenotypes. The fitness of AAG mutants (translation inhibition) is shown for seven transient ORFs under stress conditions. Fitness is estimated by comparing colony size between AAG mutants (n = 24 per strain and condition) and ATG controls (n = 24) using the LI detector pipeline.[70] A permutation test is used to test for a difference in fitness between the AAG mutant and AAG control, with significant differences indicated as follows: *p < 0.05 **p < 0.01 ***p < 0.001. Error bars indicate standard errors. A cross symbol after the ORF names indicates unannotated nORFs assigned systematic names in this study.

(C–E) Deleterious impact of inhibiting translation of transient nORF YDL204W-A in a liquid growth assay. Liquid growth curve of a strain in which YDL204W-A translation is inhibited by mutating its start codon (AAG) and a strain with the initial codon as ATG in: 1 M NaCl (C), 100 mM hydroxyurea (D), and YPDA (E), with three technical replicates for each strain. The shaded area covers 1 SD from the mean OD value among replicates.

(F–H) Expression from plasmid restores wild-type growth to YDL204W-A start codon mutants. Liquid growth curves of an attempted rescue of the YDL204W-A AAG mutant by expressing intact YDL204W-A from a plasmid. The AAG start codon mutants were transformed with either an empty plasmid or a plasmid expressing the intact ORF; the ATG controls were transformed with an empty plasmid. All strains were then assayed for growth in liquid media in either 1 M NaCl (F), 100 mM hydroxyurea (G), or YPDA (H) with three technical replicates each. The shaded area covers 1 SD from the mean OD value among replicates.

(I) YDL204W-A is translated and not conserved. Top: ribosome profiling reads mapped by iRibo to YDL204W-A show triplet periodicity. Bottom: alignment of the YDL204W-A ORF against homologous DNA in the *Saccharomyces* genus.

**Table 1. Evolutionary properties of transient ORFs with phenotypes induced by inhibiting translation**

| ORF name | Reading frame conservation | pN/pS (p value) | TBLASTN matches |
|---|---|---|---|
| YBR196C-A | 0.29 | 1.34 (0.65) | 0 |
| YDL204W-A[a] | 0.20 | 1.25 (0.83) | 0 |
| YGR016C-A[a] | 0.29 | 0.66 (0.36) | 0 |
| YJL077C | 0.21 | 0.74 (0.19) | 0 |
| YNL040C-A[a] | 0.38 | 0.97 (1.00) | 0 |
| YPR096C | 0.20 | 1.39 (0.47) | 0 |

The pN/pS ratio is obtained from nucleotide variation in the ORF among the 1,011 *S. cerevisiae* strains assembled by Peter et al.[40] TBLASTN was run for each ORF against genomes in the subphylum *Saccharomycotina*, excluding the genus *Saccharomyces*, with an e value threshold of $10^{-4}$.
[a]We assigned this unannotated ORF a systematic name based on SGD conventions.

translated ORFs appear to be frequently lost to disrupting mutations, only to be replaced by other transient translated ORFs following translation-enabling mutations. Despite these profound differences, transient translated ORFs, like conserved ones, can affect the phenotype and fitness of the organism. Several well-characterized coding sequences unique to *S. cerevisiae*, such as *HUR1*[66] and *MDF1*,[62] play key roles in biological processes by encoding lineage-specific proteins that physically interact with conserved proteins. In addition, around 100 transient ORFs are annotated as coding genes and have therefore been extensively screened; a majority express stable proteins, and many have known loss-of-function phenotypes. Their genetic interaction patterns suggest involvement in a wide array of specialized cellular processes. Our experiments revealed that disabling the start codons of unannotated transient translated ORFs can cause a large fitness reduction in stress conditions. The strength of the fitness reduction observed was highly dependent on the stressor applied in the environment, suggesting again specialized cellular roles. In some cases, disabling the start codon resulted in growth increases, perhaps indicating that disabling translation saved the cell energy.

Our work adds to the growing research on the roles noncanonical coding play across many species, including humans.[7,74] We note that "noncanonical" is not a coherent biological category, as it simply indicates the class of sequences that have not been annotated in genome databases. We demonstrate that the division between "canonical" and "noncanonical" translation in *S. cerevisiae* corresponds largely, but not perfectly, to a biological division between transient and conserved. It is this biological division that is fundamental: the 101 yeast cORFs classified as transient have sequence and evolutionary properties nearly identical to noncanonical transient ORFs, except for sequence length, and should be placed in the same category. We can thus reclassify the translatome according to biology rather than annotation history.

It is perhaps unexpected 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.[32] For example, Xie et al.[75] identified a mouse protein contributing to the 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. 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[76]; it is plausible that translated sequences also show such a division. There are several explanations for why translated ORFs may lack detectable signatures of selection. Most transient ORFs are expressed at much lower levels than canonical genes and therefore may have minimal effects on the phenotype. For those that do have large and beneficial effects in some environmental conditions, these may be balanced by deleterious effects in other conditions. Moreover, selection may occur, and be biologically relevant, below the limits of detectability for the genomic approaches we used. Our findings do not imply an absence of selective forces in shaping the patterns of noncanonical translation. Rather, the particular selective environment favoring the expression of these sequences may be too short lived to detect selection using traditional comparative genomics approaches. Previous research, such as the proto-gene model of *de novo* gene birth,[3] has proposed that recently emerged translated ORFs serve as an intermediary between noncoding sequences and mature genes. Our results add to the evidence that these ORFs provide many potential phenotypes from which selection could preserve beneficial ones for the long term.[65] Still, the observation that even ORFs with phenotypes lack evidence of conservation at the population level suggests that there are filters that prevent the vast majority of recently emerged translated ORFs, even those with beneficial phenotypes, from evolving into mature genes that are preserved over long evolutionary time. The primary influence of the great majority of *de novo* ORFs is in their biological activity over their short lifespans.

The yeast reference translatome resource we constructed with iRibo is meant to facilitate community efforts to decipher the specific physiological implications of transient translated ORFs. Our proof-of-concept analyses of subcellular localization, genetic interactions, and ATG→AAG mutants suggest involvement in diverse cellular processes and pathways. We note that some transient translatome phenotypes may be mediated by a protein product, by the process of translation itself, or both. Translation of both uORFs[77] and dORFs[78] can affect the expression of nearby genes. Translation also plays a major role in the regulation of RNA metabolism through the nonsense-mediated decay pathway.[79,80] Dissection of the molecular mechanisms mediating transient translatome phenotypes is an exciting area for future research.

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 elements from most canonical genes. Nevertheless, as with conserved genes, understanding the biology of transient ORFs is necessary for understanding the relationship between genotypes and phenotypes.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Yeast strains
- METHOD DETAILS
  - Defining candidate ORFs
  - Yeast ribo-seq dataset collection and read mapping
  - Translation calling
  - Estimating translation rates across genomic contexts
  - Identifying ORF homologous sequences
  - Division of ORFs into sets
  - Reading frame conservation
  - Detecting distant homology among *S. cerevisiae* ORFs
  - Tests of selection using the dN/dS and pN/pS ratios
  - Classification of ORFs into transient and conserved
  - Coding score calculation
  - Analysis of published microscopy studies
  - Literature analysis of transient ORFs
  - Genetic interaction analysis
  - Creation of yeast strains
  - Screening strategy for fitness estimation
  - Liquid growth assay
  - Microscopy
- QUANTIFICATION AND STATISTICAL ANALYSIS

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.04.002.

## ACKNOWLEDGMENTS

We thank Dr. Emmanuel Doram Levy's team at the Weizmann Institute of Science for sharing the fluorescence intensity data displayed in YeastRGB. We thank Dr. Benjamin Dubreuil for the helpful discussion over YeastRGB data. We thank Dr. Allyson O'Donnell for her help in microscopy image acquisition. We thank Drs. Craig Kaplan and Nikolaos Vakirlis for helpful discussions of an earlier preprinted version of this manuscript. This work was supported by funds provided by the Searle Scholars Program to A.-R.C., the National Science Foundation grant MCB-2144349 to A.-R.C., and the National Institute of General Medical Sciences of the National Institutes of Health grants R00GM108865 and DP2GM137422 (awarded to A.-R.C.).

## AUTHOR CONTRIBUTIONS

Conceptualization, A.W. and A.-R.C.; methodology, A.W., A.-R.C., S.B.P., N.C.C., and O.A.; investigation, A.W., N.C.C., S.B.P., O.A., C.H., and L.C.; writing – original draft, A.W., S.B.P., O.A., and N.C.C.; writing – review & editing, A.W., A.-R.C., S.B.P., N.C.C., O.A., C.H., and L.C.; supervision, A.-R.C.

## DECLARATION OF INTERESTS

A.-R.C. is a member of the Scientific Advisory Board for Flagship Labs 69, Inc. (ProFound Therapeutics).

Received: August 18, 2022
Revised: January 30, 2023
Accepted: April 6, 2023
Published: May 9, 2023

## REFERENCES

1. Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. *8*, 1365–1379. https://doi.org/10.1016/j.celrep.2014.07.045.
2. Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. *15*, 205–213. https://doi.org/10.1038/nrg3645.
3. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and *de novo* gene birth. Nature *487*, 370–374. https://doi.org/10.1038/nature11184.
4. Wilson, B.A., and Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol. Evol. *3*, 1245–1252. https://doi.org/10.1093/gbe/evr099.
5. Pruitt, K.D., and Maglott, D.R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. *29*, 137–140. https://doi.org/10.1093/nar/29.1.137.
6. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018). Improved Ribo-seq enables identification of cryptic translation events. Nat. Methods *15*, 363–366. https://doi.org/10.1038/nmeth.4631.
7. Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146. https://doi.org/10.1126/science.aay0262.
8. Prensner, J.R., Enache, O.M., Luria, V., Krug, K., Clauser, K.R., Dempster, J.M., Karger, A., Wang, L., Stumbraite, K., Wang, V.M., et al. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat. Biotechnol. *39*, 697–704. https://doi.org/10.1038/s41587-020-00806-2.
9. van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The translational landscape of the human heart. Cell *178*, 242–260.e29. https://doi.org/10.1016/j.cell.2019.05.010.
10. Laumont, C.M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P., et al. (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat. Commun. *7*, 10238. https://doi.org/10.1038/ncomms10238.
11. Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A.G., Khan, O.M., Brewer, J.R., Skadow, M.H., Duizer, C., et al. (2018). The translation of non-canonical open reading frames controls mucosal immunity. Nature *564*, 434–438. https://doi.org/10.1038/s41586-018-0794-7.
12. Makarewich, C.A., and Olson, E.N. (2017). Mining for micropeptides. Trends Cell Biol. *27*, 685–696. https://doi.org/10.1016/j.tcb.2017.04.006.
13. Anderson, D.M., Anderson, K.M., Chang, C.-L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell *160*, 595–606. https://doi.org/10.1016/j.cell.2015.01.009.
14. Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228–232. https://doi.org/10.1038/nature21034.
15. Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S.E., Hildenbrand, C., Rom, J., Aulmann, S., Sinn, H.-P., Vandesompele, J.,

et al. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. Oncogene *37*, 4750–4768. https://doi.org/10.1038/s41388-018-0281-5.

16. Housman, G., and Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochim. Biophys. Acta *1859*, 31–40. https://doi.org/10.1016/j.bbagrm.2015.07.017.
17. Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105. https://doi.org/10.1038/nsmb0207-103.
18. Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. *19*, 208. https://doi.org/10.1186/s13059-018-1590-2.
19. Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565. https://doi.org/10.1101/gr.6036807.
20. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254. https://doi.org/10.1038/nature01644.
21. Ward, L.D., and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science *337*, 1675–1678. https://doi.org/10.1126/science.1225057.
22. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA *111*, 6131–6138. https://doi.org/10.1073/pnas.1318948111.
23. Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates, J.R., Lockhart, D.J., and Winzeler, E.A. (2002). Parallel identification of new genes in Saccharomyces cerevisiae. Genome Res. *12*, 1210–1220. https://doi.org/10.1101/gr.226802.
24. Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casarégola, S., de Montigny, J., Gaillardin, C., Lépingle, A., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. FEBS Lett. *487*, 31–36. https://doi.org/10.1016/S0014-5793(00)02275-4.
25. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993. https://doi.org/10.1002/embj.201488411.
26. Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., and Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics *14*, 648. https://doi.org/10.1186/1471-2164-14-648.
27. Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A.K., Nielly-Thibault, L., Namy, O., and Landry, C.R. (2019). 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. https://doi.org/10.1101/gr.239822.118.
28. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J.L., Messeguer, X., and Albà, M.M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat. Ecol. Evol. *2*, 890–896. https://doi.org/10.1038/s41559-018-0506-6.
29. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2016). sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *44*, D324–D329. https://doi.org/10.1093/nar/gkv1175.
30. Lee, J., Wacholder, A., and Carvunis, A.-R. (2021). Evolutionary characterization of the short protein SPAAR. Genes *12*, 1864. https://doi.org/10.3390/genes12121864.
31. Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Calvet, F., Jungreis, I., Gonzalez, J.M., Magrane, M., Martinez, T.F., Schulz, J.F., et al. (2022). Standardized annotation of translated open reading frames. Nat. Biotechnol. *40*, 994–999. https://doi.org/10.1038/s41587-022-01369-0.
32. Van Oss, S.B., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet. *15*, e1008160. https://doi.org/10.1371/journal.pgen.1008160.
33. Blevins, W.R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J.L., Espinar, L., Díez, J., Carey, L.B., and Albà, M.M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. Nat. Commun. *12*, 604. https://doi.org/10.1038/s41467-021-20911-3.
34. Yagoub, D., Tay, A.P., Chen, Z., Hamey, J.J., Cai, C., Chia, S.Z., Hart-Smith, G., and Wilkins, M.R. (2015). 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. https://doi.org/10.1021/acs.jproteome.5b00734.
35. Lu, S., Zhang, J., Lian, X., Sun, L., Meng, K., Chen, Y., Sun, Z., Yin, X., Li, Y., Zhao, J., et al. (2019). A hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res. *47*, 8111–8125. https://doi.org/10.1093/nar/gkz646.
36. Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Le Knisbacher, B.A., P.M., et al. (2020). Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. https://doi.org/10.1101/2020.02.12.945840.
37. Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., and Dieterich, C. (2017). Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res. *45*, 2960–2972. https://doi.org/10.1093/nar/gkw1350.
38. Ji, Z. (2018). RibORF: identifying genome-wide Translated Open Reading Frames Using Ribosome Profiling. Curr. Protoc. Mol. Biol. *124*, e67. https://doi.org/10.1002/cpmb.67.
39. Calviello, L., and Ohler, U. (2017). Beyond Read-Counts: ribo-seq data analysis to understand the functions of the transcriptome. Trends Genet. *33*, 728–744. https://doi.org/10.1016/j.tig.2017.08.003.
40. Peter, J., De Chiara, M.D., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344. https://doi.org/10.1038/s41586-018-0030-5.
41. Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. Cell *175*, 1533–1545.e20. https://doi.org/10.1016/j.cell.2018.10.023.
42. Choudhary, S., Li, W., and D Smith, A. (2020). Accurate detection of short and long active ORFs using Ribo-seq data. Bioinformatics *36*, 2053–2059. https://doi.org/10.1093/bioinformatics/btz878.
43. Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., et al. (1998). SGD: saccharomyces genome database. Nucleic Acids Res. *26*, 73–79. https://doi.org/10.1093/nar/26.1.73.
44. Gerashchenko, M.V., and Gladyshev, V.N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res. *42*, e134. https://doi.org/10.1093/nar/gku671.
45. Santos, D.A., Shi, L., Tu, B.P., and Weissman, J.S. (2019). Cycloheximide can distort measurements of mRNA levels and translation efficiency. Nucleic Acids Res. *47*, 4974–4985. https://doi.org/10.1093/nar/gkz205.
46. Duncan, C.D.S., and Mata, J. (2017). Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. Sci. Rep. *7*, 10331. https://doi.org/10.1038/s41598-017-10650-1.
47. Pelechano, V., Wei, W., Jakob, P., and Steinmetz, L.M. (2014). Genome-wide identification of transcript start and end sites by transcript isoform sequencing. Nat. Protoc. *9*, 1740–1759. https://doi.org/10.1038/nprot.2014.121.
48. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

49. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife *4*, e08890. https://doi.org/10.7554/eLife.08890.

50. Moro, S.G., Hermans, C., Ruiz-Orera, J., and Albà, M.M. (2021). Impact of uORFs in mediating regulation of translation in stress conditions. BMC Mol. Cell Biol. *22*, 29. https://doi.org/10.1186/s12860-021-00363-9.

51. Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and Hittinger, C.T. (2011). The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda) *1*, 11–25. https://doi.org/10.1534/g3.111.000273.

52. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M.M. (2015). Origins of de novo genes in human and chimpanzee. PLoS Genet. *11*, e1005721. https://doi.org/10.1371/journal.pgen.1005721.

53. Firth, A.E. (2014). Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. *42*, 12425–12439. https://doi.org/10.1093/nar/gku981.

54. Sealfon, R.S., Lin, M.F., Jungreis, I., Wolf, M.Y., Kellis, M., and Sabeti, P.C. (2015). FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. *16*, 38. https://doi.org/10.1186/s13059-015-0603-7.

55. Hardison, R.C. (2003). Comparative genomics. PLoS Biol. *1*, E58. https://doi.org/10.1371/journal.pbio.0000058.

56. Dujon, B. (1996). The yeast genome project: what did we learn? Trends Genet. *12*, 263–270. https://doi.org/10.1016/0168-9525(96)10027-5.

57. Dujon, B., Alexandraki, D., André, B., Ansorge, W., Baladron, V., Ballesta, J.P.G., Banrevi, A., Bolle, P.A., Bolotin-Fukuhara, M., Bossier, P., et al. (1994). Complete DNA sequence of yeast chromosome XI. Nature *369*, 371–378. https://doi.org/10.1038/369371a0.

58. Chong, Y.T., Koh, J.L.Y., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., and Andrews, B.J. (2015). Yeast proteome dynamics from single cell imaging and automated analysis. Cell *161*, 1413–1424. https://doi.org/10.1016/j.cell.2015.04.051.

59. Koh, J.L.Y., Chong, Y.T., Friesen, H., Moses, A., Boone, C., Andrews, B.J., and Moffat, J. (2015). CYCLoPs: A comprehensive database constructed from automated analysis of protein abundance and subcellular localization patterns in Saccharomyces cerevisiae. G3 (Bethesda) *5*, 1223–1232. https://doi.org/10.1534/g3.115.017830.

60. Meurer, M., Duan, Y., Sass, E., Kats, I., Herbst, K., Buchmuller, B.C., Dederer, V., Huber, F., Kirrmaier, D., Štefl, M., et al. (2018). Genome-wide C-SWAT library for high-throughput yeast genome tagging. Nat. Methods *15*, 598–600. https://doi.org/10.1038/s41592-018-0045-8.

61. Dubreuil, B., Sass, E., Nadav, Y., Heidenreich, M., Georgeson, J.M., Weill, U., Duan, Y., Meurer, M., Schuldiner, M., Knop, M., et al. (2019). YeastRGB: comparing the abundance and localization of yeast proteins across cells and libraries. Nucleic Acids Res. *47*, D1245–D1249. https://doi.org/10.1093/nar/gky941.

62. Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., and Wang, W. (2010). 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. https://doi.org/10.1038/cr.2010.31.

63. Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A molecular portrait of de novo genes in yeasts. Mol. Biol. Evol. *35*, 631–645. https://doi.org/10.1093/molbev/msx315.

64. Li, D., Yan, Z., Lu, L., Jiang, H., and Wang, W. (2014). Pleiotropy of the de novo-originated gene MDF1. Sci. Rep. *4*, 7280. https://doi.org/10.1038/srep07280.

65. Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S.B., Wacholder, A., Medetgul-Ernar, K., Bowman, R.W., Hines, C.P., Iannotta, J., et al. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat. Commun. *11*, 781. https://doi.org/10.1038/s41467-020-14500-z.

66. Omidi, K., Jessulat, M., Hooshyar, M., Burnside, D., Schoenrock, A., Kazmirchuk, T., Hajikarimlou, M., Daniel, M., Moteshareie, H., Bhojoo, U., et al. (2018). Uncharacterized ORF HUR1 influences the efficiency of non-homologous end-joining repair in Saccharomyces cerevisiae. Gene *639*, 128–136. https://doi.org/10.1016/j.gene.2017.10.003.

67. Hajikarimlou, M., Moteshareie, H., Omidi, K., Hooshyar, M., Shaikho, S., Kazmirchuk, T., Burnside, D., Takallou, S., Zare, N., Jagadeesan, S.K., et al. (2020). Sensitivity of yeast to lithium chloride connects the activity of YTA6 and YPR096C to translation of structured mRNAs. PLoS One *15*, e0235033. https://doi.org/10.1371/journal.pone.0235033.

68. Alesso, C.A., Discola, K.F., and Monteiro, G. (2015). The gene ICS3 from the yeast Saccharomyces cerevisiae is involved in copper homeostasis dependent on extracellular pH. Fungal Genet. Biol. *82*, 43–50. https://doi.org/10.1016/j.fgb.2015.06.007.

69. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420. https://doi.org/10.1126/science.aaf1420.

70. Parikh, S.B., Castilho Coelho, N., and Carvunis, A.R. (2021). LI Detector: a framework for sensitive colony-based screens regardless of the distribution of fitness effects. G3 (Bethesda) *11*, jkaa068. https://doi.org/10.1093/g3journal/jkaa068.

71. Kearse, M.G., and Wilusz, J.E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. *31*, 1717–1731. https://doi.org/10.1101/gad.305250.117.

72. Loughran, G., Zhdanov, A.V., Mikhaylova, M.S., Rozov, F.N., Datskevich, P.N., Kovalchuk, S.I., Serebryakova, M.V., Kiniry, S.J., Michel, A.M., O'Connor, P.B.F., et al. (2020). Unusually efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. Proc. Natl. Acad. Sci. USA *117*, 24936–24946. https://doi.org/10.1073/pnas.2001433117.

73. McVeigh, A., Fasano, A., Scott, D.A., Jelacic, S., Moseley, S.L., Robertson, D.C., and Savarino, S.J. (2000). IS1414, an Escherichia coliInsertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. Infect. Immun. *68*, 5710–5715. https://doi.org/10.1128/IAI.68.10.5710-5715.2000.

74. Wright, B.W., Yi, Z., Weissman, J.S., and Chen, J. (2022). The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol. *32*, 243–258. https://doi.org/10.1016/j.tcb.2021.10.010.

75. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife *8*, e44392. https://doi.org/10.7554/eLife.44392.

76. Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science *317*, 815–819. https://doi.org/10.1126/science.1140748.

77. Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip. Rev. RNA *5*, 765–778. https://doi.org/10.1002/wrna.1245.

78. Wu, Q., Wright, M., Gogol, M.M., Bradford, W.D., Zhang, N., and Bazzini, A.A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J. *39*, e104763. https://doi.org/10.15252/embj.2020104763.

79. Andjus, S., Szachnowski, U., Vogt, N., Hatin, I., Papadopoulos, C., Lopes, A., Namy, O., Wery, M., and Morillon, A. (2022). Translation is a key determinant controlling the fate of cytoplasmic long non-coding RNAs. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.25.493276.

80. Wery, M., Descrimes, M., Vogt, N., Dallongeville, A.-S., Gautheret, D., and Morillon, A. (2016). Nonsense-mediated decay restricts LncRNA levels in yeast unless blocked by double-stranded RNA structure. Mol. Cell *61*, 379–392. https://doi.org/10.1016/j.molcel.2015.12.020.

81. Dunham, M.J., Dunham, M.J., Gartenberg, M.R., and Brown, G.W. (2015). Methods in Yeast Genetics and Genomics: a Cold Spring Harbor Laboratory Course Manual (Cold Spring Harbor Laboratory Press).

82. Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019.

83. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2011). The European nucleotide archive. Nucleic Acids Res. *39*, D28–D31. https://doi.org/10.1093/nar/gkq967.

84. Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341. https://doi.org/10.1038/nature07743.

85. Liti, G., Nguyen Ba, A.N.N., Blythe, M., Müller, C.A., Bergström, A., Cubillos, F.A., Dafhnis-Calas, F., Khoshraftar, S., Malla, S., Mehta, N., et al. (2013). High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. BMC Genomics *14*, 69. https://doi.org/10.1186/1471-2164-14-69.

86. Naseeb, S., Alsammar, H., Burgis, T., Donaldson, I., Knyazev, N., Knight, C., and Delneri, D. (2018). Whole Genome Sequencing, de novo Assembly and phenotypic Profiling for the New Budding Yeast Species Saccharomyces jurei. G3 (Bethesda) *8*, 2967–2977. https://doi.org/10.1534/g3.118.200476.

87. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

88. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

89. Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In Multiple Sequence Alignment Methods Methods in Molecular Biology, D.J. Russell, ed. (Humana Press), pp. 155–170. https://doi.org/10.1007/978-1-62703-646-7_10.

90. Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. eLife *3*, e03523. https://doi.org/10.7554/eLife.03523.

91. Usaj, M., Tan, Y., Wang, W., VanderSluis, B., Zou, A., Myers, C.L., Costanzo, M., Andrews, B., and Boone, C. (2017). TheCellMap.org: A web-accessible database for visualizing and mining the global yeast genetic interaction network. CellMap.org. G3 (Bethesda) *7*, 1539–1549. https://doi.org/10.1534/g3.117.040220.

92. Bauer, S., Grossmann, S., Vingron, M., and Robinson, P.N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinform. Oxf. Engl. *24*, 1650–1651. https://doi.org/10.1093/bioinformatics/btn250.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Chemicals, peptides, and recombinant proteins** | | |
| Yeast Extract | BD Difco | DF0127179 |
| Peptone | BD Difco | DF0118170 |
| G-418 | RPI | G64000-1.0 |
| D(+) Glucose | Thermo Fisher | AAA168280E |
| Hygromycin B | RPI | H75020-1.0 |
| CellTracker Blue CMAC Dye | Invitrogen | C2110 |
| MitoTracker Red CMXRos | Invitrogen | M7512 |
| Tunicamycin | Sigma | SML1287-1ML |
| Fluconazole | Sigma | PHR1160-1G |
| Sodium Chloride | Spectrum | S1240-1KG |
| Hydroxyurea | Thermo Scientific | A10831.14 |
| Hydrogen Peroxide | Fisher Scientific | H323-500 |
| DMSO | Amresco | 0231-500ML |
| Poly(ethylene-glycol) 3350 | Sigma | P4338-500G |
| ssDNA | Life Technologies | 15632011 |
| Lithium Acetate dihydrate | Sigma | L4158-100G |
| **Deposited data** | | |
| Deletion screen colony growth images | This paper | Figshare: https://doi.org/10.6084/m9.figshare.21741434.v1 |
| Ribosome profiling analysis results | This paper | Figshare: https://doi.org/10.6084/m9.figshare.22312729.v1 |
| C-SWAT collection | Meurer et al.[60] | Table S2 |
| YeastRGB collection | Dubreuil et al.[61] | Yeastrgb.org |
| CYCLoPs collection | Ko et al.[59] | https://thecellvision.org/cyclops/ |
| *Saccharomyces cerevisiae* S288C reference genome sequence R64.2.1 | Saccharomyces genome database | http://sgd-archive.yeastgenome.org/sequence/ |
| *S. paradoxus* genome | Liti et al.[70] | http://www.saccharomycessensustricto.org/ |
| *S. arboricolus* genome | Liti et al.[81] | GCF_000292725.1 |
| *S. jurei* genome | Naseeb et al.[82] | GCA_900290405.1 |
| *S. mikatae*, *S. bayanus var. uvarum*, *S. bayanus var. bayanus*, and *S. kudriavzevii* genome | Scannell et al.[51] | http://www.saccharomycessensustricto.org/ |
| TIF-seq data | Pelechano et al.[47] | GEO: GSE39128 |
| *S. cerevisiae* strain genomes | Peter et al.[40] | http://1002genomes.u-strasbg.fr/files/ |
| Budding yeast genomes | Shen et al.[41] | https://y1000plus.wei.wisc.edu/data |
| **Experimental models: Organisms/strains** | | |
| *Saccharomyces cerevisiae*: BY4741 | Dharmacon | YSC1048 |
| *Saccharomyces cerevisiae*: BY4741, deletion collection | Dharmacon | YSC1053 |
| *Saccharomyces cerevisiae*: BY4741, ORF::KanMx (mini collection with the 49 nORFs and 3 cORFs deleted) | This paper | Wacholder 2023, deletion collection, see Table S10 |
| *Saccharomyces cerevisiae*: BY4741, ORF-mNG:HYG (mini collection with the selected ORFs tagged with mNeonGreen) | This paper | Wacholder 2023, mNG collection, see Table S10 |
| *Saccharomyces cerevisiae*: BY4741, YDL204W-A(wt):HYG | This paper | yARC0602 |
| *Saccharomyces cerevisiae*: BY4741, YDL204W-A(ATG→AAG):HYG | This paper | yARC0601 |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| *Saccharomyces cerevisiae*: BY4741, YBR196C-A(wt):HYG | This paper | yARC0604 |
| *Saccharomyces cerevisiae*: BY4741, YBR196C-A(ATG→AAG):HYG | This paper | yARC0603 |
| *Saccharomyces cerevisiae*: BY4741, YDR073C-A(wt):HYG | This paper | yARC0606 |
| *Saccharomyces cerevisiae*: BY4741, YDR073C-A(ATG→AAG):HYG | This paper | yARC0605 |
| *Saccharomyces cerevisiae*: BY4741, YGR016C-A(wt):HYG | This paper | yARC0608 |
| *Saccharomyces cerevisiae*: BY4741, YGR016C-A(ATG→AAG):HYG | This paper | yARC0607 |
| *Saccharomyces cerevisiae*: BY4741, YJL077C(wt):HYG | This paper | yARC0610 |
| *Saccharomyces cerevisiae*: BY4741, YJL077C(ATG→AAG):HYG | This paper | yARC0609 |
| *Saccharomyces cerevisiae*: BY4741, YNL040C-A(wt):HYG | This paper | yARC0612 |
| *Saccharomyces cerevisiae*: BY4741, YNL040C-A(ATG→AAG):HYG | This paper | yARC0611 |
| *Saccharomyces cerevisiae*: BY4741, YPR096C(wt):HYG | This paper | yARC0614 |
| *Saccharomyces cerevisiae*: BY4741, YPR096C(ATG→AAG):HYG | This paper | yARC0613 |
| *Saccharomyces cerevisiae*: BY4741, YDL204W-A(wt):HYG, pAG-GPD-ccdB1-KanMx | This paper | yARC0831 |
| *Saccharomyces cerevisiae*: BY4741, YDL204W-A(ATG→AAG):HYG, pAG-GPD-ccdB1-KanMx | This paper | yARC0842 |
| *Saccharomyces cerevisiae*: BY4741, YDL204W-A(ATG→AAG):HYG, pAG-GPD-YDL204W-A-KanMx | This paper | yARC0848 |
| Recombinant DNA | | |
| Plasmid: pAG-GPD-ccdB1-KanMx | This paper | pARC0112 |
| Plasmid: pAG-GPD-YDL204W-A-KanMx | This paper | pARC0300 |
| Software and algorithms | | |
| Code for analyses conducted | This paper | Zenodo: https://doi.org/10.5281/zenodo.7474228 |
| Code for analyzing images of colonies on plates | This paper | Zenodo: https://doi.org/10.5281/zenodo.7760846 |
| R version 4.12 | R | https://www.r-project.org/ |
| BLAST 2.9.0+ | National Library of Medicine | https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi |
| Ontologizer 2.0 | Bauer et al.[83] | http://ontologizer.de/ |
| Water | EMBOSS | https://www.ebi.ac.uk/Tools/psa/emboss_water/ |
| MUSCLE 3.8.31 | Edgar[84] | https://www.drive5.com/muscle/ |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Anne-Ruxandra Carvunis (anc201@pitt.edu).

### Materials availability

All materials will be made available on request to the lead contact.

### Data and code availability

- Ribosome profiling analysis results are publicly available at Figshare. Plate images for colony growth assays are available at Figshare and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Yeast strains

All strains used in this study are derived from *S. cerevisiae* BY4741 (Dharmacon, YSC1048). Cells were stored at -80° C and routinely cultured in YPDA at 30° C with shaking or in YPDA agar plates at 30° C. The parental strain and all derivatives produced in this study are listed in Table S10. The lithium acetate method[81] was used to create new strains and selection was performed on appropriate selection plates. For genomic integration, the inserts were PCR amplified from plasmids or GBlocks.

## METHOD DETAILS

### Defining candidate ORFs

To identify a set of translated ORFs, a set of candidate ORFs was constructed for which translation status could be inferred using ribo-seq data. ORFs were identified on the R64.2.1 *Saccharomyces cerevisiae* genome assembly downloaded from SGD.[43] The initial set of candidates consisted of all possible single-exon reading frames starting with an ATG, ending with a canonical stop codon, and having at least one additional codon between the start and stop. Among all ORFs that shared a stop codon, all but the longest were discarded. An ORF was considered canonical if it shared a stop codon with an ORF annotated as ''verified'', ''uncharacterized'', or ''transposable element gene'' on SGD. All other ORFs that overlapped a canonical ORF on the same strand were removed (including pairs of overlapping canonical genes) while ORFs that overlapped cORFs on the opposite strand were classified as antisense ORFs.

### Yeast ribo-seq dataset collection and read mapping

A list of *S. cerevisiae* ribosome profiling (ribo-seq) studies was identified by conducting a broad literature search. For each study, all ribo-seq experiments were added to the dataset except those conducted on mutants designed to alter wildtype translation patterns. The full list of experiments and studies included is given in Tables S1 and S2, respectively. The fastq files associated with each experiment were downloaded from Sequence Read Archive[82] or the European Nucleotide Archive.[83] If adaptors were present in the fastq file, they were trimmed. 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.

In initial mapping, reads were assigned to the genomic position aligning with the first base of the read. It was necessary to remap the reads such that the position assigned to the read instead corresponded to the first amino acid in the P-site of the translating ribosome, as in previous ribo-seq analyses,[37] so that the triplet periodic signal indicative of active translation overlaps precisely the bounds of translated ORFs. This was done by shifting all reads by the same number of positions, with the number determined separately for each read length and each experiment. To determine this number, a metagene profile was constructed: the number of reads in each of the -20 to +20 positions relative to the start codon was counted, accumulated over all annotated genes on Saccharomyces Genome Database (SGD).[43] As there should be many more reads on the start codon of annotated genes than the sequence immediately upstream of these genes, the first attempt was to remap the first position with read count above a threshold to the first amino acid on the start codons, which then requires all other reads to shift by the same amount. The threshold selected was 5% of the total reads within 20 bases of the annotated start codons. The attempted shift was accepted if the expected triplet periodic pattern was obtained; i.e., there were more remapped reads on the first base of the codons of annotated genes than on the second or third base. Otherwise, a second shift was attempted from the next position exceeding the read count threshold, and so on until both criteria were met.

For quality control, presence of triplet periodicity was then tested for each read length in each experiment. The number of reads mapping (after remapping) to the first, second, and third position of each codon was counted 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 for a given experiment 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.

### Translation calling

The iRibo program can be applied to any set of ribo-seq experiments to identify a set of ORFs with evidence of translation among those experiments. To construct a reference translatome, translation was inferred using ribo-seq data from the full set of experiments we collected that passed quality control (Table S4). Separately, iRibo was also run on specific subsets of the full collection, including: experiments with or without the drug cycloheximide, experiments only on cells grown in YPD; only on cells grown on SD; and only on cells grown in YPD without cycloheximide (Table S4). iRibo was also run separately for each individual study, generating lists of translated ORFs within each study.

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

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

### Estimating translation rates across genomic contexts

All nORFs were partitioned into genomic contexts, with nonoverlapping nORFs classified by the relation between the nORF and any cORF located on the same transcript and antisense nORFs classified by partial or complete overlap of the opposite strand gene. The transcripts reported in Pelechano et al.[47] based on TIF-seq data were used for this analysis. An nORF was considered antisense if it overlapped an ORF annotated as "verified", "uncharacterized", "transposable element" or "blocked" on SGD on the opposite strand and nonoverlapping otherwise (ORFs overlapping annotated genes on the same strand were excluded from analysis, as described above). A nonoverlapping nORF was considered to share a transcript with a cORF or annotated non-coding RNA if any transcript fully contained both the nORF and the cORF or annotated RNA sequence; the ORF was then further classified as being either a uORF or dORF based on whether it was upstream or downstream of the cORF or RNA. If an nORF shared a transcript with both its upstream and downstream neighboring cORFs, it was classified according to the cORF that was closer.

### Identifying ORF homologous sequences

Genomes were obtained from seven relatives of *S. cerevisiae* within the *Saccharomyces* genus: *S. paradoxus* from Liti et al.,[84] *S. arboricolus* from Liti et al.,[85] *S. jurei* from Naseeb et al.,[86] and *S. mikatae*, *S. bayanus var. uvarum*, *S. bayanus var. bayanus*, and *S. kudriavzevii* from Scannell et al.[51] Alignments were constructed between each *S. cerevisiae* ORF and its homologs in each *Saccharomyces* relative using synteny information. To identify anchor genes for syntenic blocks, BLASTP was run for each annotated ORF in *S. cerevisiae* against each ORF in the comparison species. Identified homolog pairs with e-value $< 10^{-7}$ were selected as potential anchors. For each ORF in the *S. cerevisiae* genome, the upstream anchor $G_0$ and downstream anchor $G_1$ were selected that minimized the sum of the distance between the anchors in *S. cerevisiae* and the distance between the anchors in the comparison species; this sum was required to be less than 60 kb. The sequence between and including $G_0$ and $G_1$ were then extracted from both the *S. cerevisiae* genome and the comparison species and a pairwise alignment of the syntenic region was generated using MUSCLE 3.8.31.[87]

To confirm that the ORF was matched to genuinely homologous DNA, the alignment of the *S. cerevisiae* ORF along with its 50 bp flanking regions was extracted from the full syntenic alignment. The extracted region was then realigned using the Smith-Waterman algorithm[88] with a match bonus of 5, a mismatch penalty of 4, and a gap penalty of 4. To test homology, 1000 alignments were constructed 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. Homology was accepted as confirmed if the p-value was less than 1%, and alignments were excluded from analysis if homology was not confirmed.

If a syntenic alignment could not be constructed for a particular *S. cerevisiae* ORF and comparison species (because homology failed or there were no appropriate anchors), BLAST was attempted as an alternative method of finding the homologous DNA sequence. For these ORF sequences, BLASTN was run against the genome of the comparison species. For each reciprocal best matching pair with e-value $< 10^{-4}$, the matched sequences in both species were extracted, together with a 1000 bp flanking region in both ends, and aligned using MUSCLE.[87] DNA homology was then tested using Smith-Waterman alignment as described above.

### Division of ORFs into sets

Evolutionary analysis of ORFs was done separately for those ORFs for which there existed substantial information to test selection ("high information ORFs") and those for which less information was available ("low information ORFs"). To be placed in the high information set, the ORF had to meet a homology criterion and a diversity criterion. The homology criterion required that DNA homology was confirmed in either a synteny or BLAST-based pairwise alignment with at least four other species in the *Saccharomyces* genus. For the diversity criterion, the number of single nucleotide differences (excluding gaps) was counted between the *S. cerevisiae* ORF and all its aligned sequence with confirmed homology among *Saccharomyces* genomes. The diversity criterion was satisfied if the median count of differences exceeded 20.

### Reading frame conservation

Reading frame conservation is a measure of conservation of codon structure developed by Kellis et al.[20] and used here with some modifications. Calculation of reading frame conservation was done on a pairwise alignment of a genomic region containing the *S. cerevisiae* ORF (either a syntenic block between conserved genes or the 1000 bp flanking region around a BLAST hit). All single-exon ORFs (ATG to stop codon) in the comparison species were identified across this region. For each ORF in the comparison species, the reading frame conservation was 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 (including start and stop codons). Positions that align to gaps or are outside the range of the *S. cerevisiae* ORF are always considered to be not in the same codon position and do not add to the numerator. 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, an index of reading frame conservation (RFC) was defined equal to the average reading frame conservation of the *S. cerevisiae* ORF against all species in the *Saccharomyces* genus for which homologous DNA could be identified.

### Detecting distant homology among *S. cerevisiae* ORFs

The genomes of 332 budding yeasts were taken from Shen et al.[41] We applied TBLASTN and BLASTP for each *S. cerevisiae* translated ORF against each genome in this dataset (excluding the *Saccharomyces* genus). Default settings were used except for setting an e-value threshold of 0.1; results were then filtered by a stricter e-value threshold as described in each analysis. The BLASTP analysis was run against the list of protein coding genes used in Shen et al.[41] while the TBLASTN analysis was run against each entire genome. In the TBLASTN analysis, scrambled sequences of each *S. cerevisiae* ORF were also included as queries to serve as a negative control.

### Tests of selection using the dN/dS and pN/pS ratios

Variant call file data for 1011 *S. cerevisiae* isolates was taken from Peter et al.[40] For each ORF, nucleotide diversity was estimated from the full set of isolates. Nucleotide diversity was estimated as the mean number of differences per site in the ORF between any pair of isolates. To calculate dN/dS, the consensus sequence among all isolates was determined. At each position in the consensus, the three possible nucleotide variations were recorded as possible polymorphisms and distinguished by polymorphism type (12 possible combinations of consensus and variant nucleotide) and whether they would result in a synonymous or nonsynonymous difference from the consensus. If at least one isolate had the polymorphism, the polymorphism was also recorded as observed. All possible and observed polymorphisms were counted among all considered ORFs.

The pN/pS ratio was calculated in a similar manner to Ruiz-Orera et al.[28] and could be applied to either a single ORF or a group of ORFs. For each ORF under consideration, the consensus sequence among all isolates was determined. At each position in the consensus, the three possible nucleotide variations were recorded as possible polymorphisms and distinguished by polymorphism type (12 possible combinations of consensus and variant nucleotide) and whether they would result in a synonymous or nonsynonymous difference from the consensus. If at least one isolate had the polymorphism, the polymorphism was also recorded as observed. All possible and observed polymorphisms were counted among all considered ORFs.

Consider a variant $X \rightarrow Y$ where $X$ is the consensus at a site and $Y$ is a possible variant. The probability of observing variant $Y$ at a position with consensus $X$, $p_{x \rightarrow y}$ was estimated as the observed count of X→Y variant sites divided by the possible count of X→Y variant sites. Under neutrality, the expected count of either synonymous or nonsynonymous X→Y variant sites is then the product of $p_{x \rightarrow y}$ and the number of possible synonymous or nonsynonymous X→Y variant sites. In this manner the expected and observed counts of synonymous and nonsynonymous variants were calculated. The pN/pS ratio is then estimated as:

$$\omega = \frac{nonsyn_{obs}/nonsyn_{exp}}{syn_{obs}/syn_{exp}}$$

Under neutrality, then, the expected count of X→Y nonsynonymous variant sites is the number of possible such variant sites times the expected probability of this variant. In this manner the expected and observed counts of all synonymous variant types were calculated. To test for deviation from neutrality, we used a chi-squared test with one degree of freedom to compare observed vs. expected counts of synonymous and nonsynonymous variants. Standard errors for the pN/pS ratio in group analyses were estimated by bootstrapping: the ORFs in the group were resampled with replacement 1000 times and the pN/pS ratio was calculated each time. The standard error was then estimated as the sample standard deviation among the 1000 pN/pS ratios.

The dN/dS ratio was calculated based on differences in the pairwise ORF alignments *S. cerevisiae* and its closest relative *S. paradoxus*. Each S. cerevisiae ORF was associated with an *S. paradoxus* ORF for which the pair had the highest reading frame conservation (or none if homology with *S. paradoxus* was not confirmed or the highest reading frame conservation was 0). Counts of differences were made only for codons that shared the same frame between these ORFs and with at most one nucleotide difference between the codons. For every eligible position in the *S. cerevisiae* ORF, each possible *S. paradoxus* difference was counted and distinguished by whether the difference was synonymous or nonsynonymous and by type (four *S. cerevisiae* nucleotides, each with three possible *S. paradoxus* differences). These observed and possible differences were then used to estimate the dN/dS ratio in the same way as described above for the pN/pS ratio.

Among nORFs with high RFC, the strong conservation in *Saccharomyces* permitted calculation of dN/dS over the entire *Saccharomyces* tree, and so this was done as an additional test of selection (as reported in Table 1). For this analysis, ancestral reconstruction of the *Saccharomyces* phylogeny was conducted using PRANK[89] with parameters -showanc -showevents -once -prunetree -keep. Ancestral reconstruction included all species in which DNA homology was confirmed. Codons were only used for counting substitutions if they shared frame conservation among all species. Observed and possible substitutions were counted across each branch and distinguished by substitution type and whether the substitutions were synonymous or nonsynonymous. Then, dN/dS was estimated in the same way as described for pN/pS above.

### Classification of ORFs into transient and conserved

All high-information nonoverlapping translated ORFs with RFC > 0.8 were classified as conserved (Figure 4A). An nORF was also classified as conserved if it overlapped no annotated feature on SGD, had TBLASTN matches with e-value < $10^{-4}$ with at least two species outside the *Saccharomyces* genus and showed at least one additional signature of purifying selection (RFC > 0.8 or a p-value < 0.05 in a test of neutrality using dN/dS or pN/pS) (Figure S6A).

Nonoverlapping ORFs were excluded from classification in the transient set if they showed homology to an ORF classified as conserved in *S. cerevisiae* (e-value < $10^{-4}$ using BLASTP) or to any sequence among budding yeasts outside *Saccharomyces*[41] (e-value < $10^{-4}$ using TBLASTN). Among remaining translated ORFs, all high-information ORFs with RFC < 0.6 were classified as transient. Low information ORFs were divided into groups and classified as transient if no group they belonged to showed evidence of selection in dN/dS analysis, pN/pS analysis, or weak homology matching analysis. Two low-information groups were cORFs and antisense nORFs. Low information nonoverlapping nORFs were each assigned to three groups corresponding to deciles of translation rate, coding score and ORF length. Analyses of dN/dS and pN/pS are described above. For weak homology detection, the number of ORFs with at least two weak TBLASTN matches (e-value < 0.05) to budding yeast genomes collected by Shen et al.[41] (excluding *Saccharomyces* species) was counted for both actual and scrambled ORF sequences. Selection was inferred if actual matches significantly ($p < 0.05$) exceeded scrambled matches using Fisher's exact test. Only ORFs that did not overlap any annotated feature on SGD were included in weak homology detection analysis.

### Coding score calculation

The coding score, described by Ruiz-Orera et al.,[90] is a measure of how close the hexamer (i.e., the nucleotide sequence of a pair of adjacent codons) frequency of an ORF is to the hexamer of coding vs. noncoding sequences. Higher scores indicate a more gene-like hexamer distribution. Coding hexamer frequencies were calculated among all ORFs annotated as "verified" or "uncharacterized" by Saccharomyces Genome Database.[43] Noncoding hexamer frequencies were calculated for all intergenic sequences (sequences in between verified or uncharacterized ORFs) in the *S. cerevisiae* genome. As intergenic sequence has no codon structure, hexamer frequencies for intergenic sequence were counted as if read in each possible coding frame. The score was then calculated as described in Ruiz-Orera et al.[90]

### Analysis of published microscopy studies

Published results were examined from fluorescent tagging experiments where the expression of ORFs was driven by native promoters and terminators. A list of ORFs detected in 15 GFP-tagged screens on wildtype strains in either normal conditions or with chemical treatment (hydroxyurea or rapamycin) were retrieved from the CYCLoPs database.[58,59] Lists of ORFs detected in the C-SWAT tagging library were taken from Meurer et al.[60] and from YeastRGB.[61] ORFs with fluorescent intensity below the reported detection threshold in each screen were filtered out. Transient ORFs that showed detectable translation products in at least one screen were considered as detected.

### Literature analysis of transient ORFs

For each transient translatome cORF, we examined all publications listed on SGD as "primary" or "additional" literature for the ORF. If the ORF had a phenotype in any listed publication, we noted the evidence for the phenotype (Table S6).

### Genetic interaction analysis

Single mutant fitness and genetic interaction data were downloaded from TheCellMap.org.[91] In this dataset, mutants of nonessential genes are full deletions and mutants of essential genes are temperature-sensitive alleles. Transient ORFs were all nonessential. Different temperature-sensitive alleles for the same essential gene were treated separately. We removed all genes or transient ORFs with a genomic overlap to another genetic element from our analyses as it is not possible to assign the observed phenotypes to either of the overlapping pairs.

We counted the number of transient ORF and nonessential genes that showed at least one genetic interaction with $\varepsilon < -.2$ and p-value < 0.05 (a negative genetic interaction) or $\varepsilon < -.35$ with a p-value<0.05 (a synthetic lethal interaction). We then divided this number by the total number of transient ORFs or nonessential genes in the Costanzo et al.[69] genetic interaction dataset to calculate the percentage showing at least one genetic interaction. We used Fisher's exact test to assess the significance of differences between percentages of nonessential genes and transient ORFs.

Gene ontology analysis of the interactors of each ORF was conducted with Ontologizer,[92] using Benjamini-Hochberg multiple testing correction and the term-for-term calculation method. The gene association file was downloaded from SGD. Gene ontology evidence codes relating to genetic interactions (IGI and HGI) were not used.

### Creation of yeast strains

Deletion mutant strains for 49 transient nORFs and 3 transient cORFs were created by using homologous recombination to replace the ORFs with a KanMX cassette. Transformations were done using the LiAc/PEG protocol[81] in the background BY4741 strain, and selected in media containing G-418. After an initial screen of these strains, a subset of the deletion strains that showed strong deleterious effects were transformed a second time, also using the LiAc/PEG protocol,[81] to replace the KanMx cassette with either an intact copy of the original ORF, or a mutant copy of the ORF with the start codon ATG and (in some cases) additional in-frame ATG codons mutated to AAG to prevent translation. This was accomplished by using homologous recombination to replace the KanMx cassette with a construct containing the intact or mutant ORF followed by a hygromycin resistance cassette. These constructs were synthesized by IDT (Integrated DNA Technologies). The resulting transformants were selected in agar plates containing hygromycin. All positive clones were sequenced to confirm presence of either the restored wildtype ORF or the ORF with a mutated start codon.

Strains containing an mNeonGreen tag for microscopy purposes were also made by homologous recombination using the LiAc/PEG protocol[81] in the BY4741 background. The mNeonGreen and hygromycin cassette sequences were amplified from a plasmid using primers containing homology to the 3′ of each ORF. The primers were designed to remove the STOP codon of each ORF and place the mNeonGreen in frame with the ORF, to be expressed under its native promoter. Positive clones were selected on agar plates containing hygromycin.

All strains were kept in glycerol stocks at −80 °C in 96 and 384-well format until used for screening. Strain genotypes are listed in Table S10.

### Screening strategy for fitness estimation

Both rounds of deletion screening were conducted at 1536 colony density, with 1 in 4 colonies on the plate being reference strains used to correct for spatial biases as described in Parikh et al.[70] In the initial deletion screen, each mutant strain was tested using 12 replicates; 72 replicates were tested per strain in the start codon mutant screen. Conditions tested were YPDA and YPDA+DMSO as unstressed conditions and five stress conditions: YPDA supplemented with 1M NaCl, 100mM Hydroxyurea, 0.6μM Tunicamycin, 25μg/ml Fluconazole, or 30mM Hydrogen peroxide ($H_2O_2$). Agar plates were incubated and imaged periodically until the colonies reached saturation. The plate handler Singer ROTOR (Singer Instrument Co. Ltd) was used to prepare all plates starting from glycerol stocks. Serial imaging of the plates was conducted using the spImager Automated Imaging System (S & P Robotics Inc., Ontario, Canada). The images were analyzed in bulk using a custom script made using functions from the MATLAB Colony Analyzer Toolkit[70] to provide colony size estimations. The LI Detector analytical pipeline[70] was used to correct for spatial biases in colony size and obtain colony fitness estimates. Strain fitness was estimated as the median of bias-corrected colony size among replicates of the strain at 40 hours in the initial screen and 90 hours in the start codon mutant screen. In the LI Detector pipeline,[70] sets of reference colonies are treated as if they were replicates of a mutant strain, with their median fitness calculated in order to construct an empirical null distribution of median fitness values to compare with estimated strain fitness. Strains were called as beneficial or deleterious using a 5% false discovery rate threshold based on this empirical null distribution. For any selected fitness threshold used to infer deleterious strains, the false discovery rate can be calculated as the proportion of null distribution fitness values below that threshold divided by the proportion of mutant strain fitness values below the threshold. Thus, fitness thresholds were selected such that a 5% FDR was obtained and strains with fitness below that threshold were inferred to be deleterious. In the same manner, a list of beneficial strains at 5% FDR was also selected.

### Liquid growth assay

For liquid growth assays, cells were first grown in liquid YPDA media overnight at 30°C in a 96-density microplate. These were then used to inoculate a new 96-density microplate with 150μl YPDA+ stress conditions (1M NaCl, 100mM Hydroxyurea)) using the Singer ROTOR (Singer Instrument Co. Ltd). This microplate was incubated at 30°C with constant double orbital shaking for a period of 72h on microplate reader Biotek Synergy H1 (Aligent Technology Inc.). Optical density readings at 600nm ($OD_{600}$) were taken every 15 minutes.

### Microscopy

The strains containing the ORFs tagged with mNeonGreen were imaged on a Nikon TiE2 inverted A1R confocal microscope. A first screening was done at high density in 96-well plates with a 40x water objective, to assess the success of the transformations. Plates were incubated with CellTracker Blue CMAC Dye (Invitrogen) and MitoTracker Red CMXRos Dye (Invitrogen) at least 10 min prior to imaging. Plates were then imaged in 4 channels (405, 488, 561, and DIC), and 3 fields of view were taken for each strain that contained

many cells. Strains that demonstrated visibly higher signal in the green channel (488nm) compared to a non-transformed background strain were selected to examine in single dishes under a 100X oil objective to more accurately evaluate sub-cellular localization. All strains were imaged in triplicate at high density and triplicate in dishes (once without CMAC and MitoTracker and two times with the dyes).

## QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed in R version 4.1.2. Details for each statistical test and analysis can be found in the results section and figure legends.