`

A vast evolutionarily transient translatome contributes to phenotype and fitness

Aaron Wacholder[12], Saurin Bipin Parikh[123], Nelson Castilho Coelho[12], Omer Acar[124], Carly Houghton[124], Lin Chou[123], and Anne-Ruxandra Carvunis[125*]

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

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

`

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. Surprisingly, 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.

Keywords:

Noncanonical translation, ribosome profiling, de novo gene birth, protein evolution, evolutionary genomics, microproteins, smORFs, genome annotation

`

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 like RefSeq[5], but also a large "noncanonical translatome" consisting of ORFs that are not annotated as genes. Despite lack of annotation, large-scale studies find that many noncanonical ORFs are translated to express stable proteins and show evidence of association with cellular phenotypes.[6–10] Additionally, 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 are important to organism fitness.[11–15] Yet, these well-studied examples represent only a small fraction of the noncanonical translatome. Most noncanonical translation could simply be biologically insignificant "translational noise" resulting from the imperfect specificity of translation processes.[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 significant 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 noncanonical ORFs detected by ribo-seq.[25–28] However, in the case of noncanonical translation, evolutionary analysis is often limited by a lack of sufficient statistical power to confidently detect selection. Most noncanonical ORFs 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 noncanonical ORFs 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 nongenic sequences) have been discovered from within the noncanonical translatome.[3,32,33]

In addition to the challenges short ORF length poses for detection of selection, it also poses challenges for unequivocal detection of translation in the first place. Microproteins are often missed by most

`

62 proteomics techniques, though specialized methods have had some success.[9,10,34–36] In ribo-seq data, the
63 most robust evidence of translation comes from a pattern of triplet periodicity in reads corresponding to
64 the progression of the ribosome across codons.[6,37,38] Ribo-seq analysis methods are less capable of
65 detecting translation of short ORFs, as they contain fewer positions to use to establish periodicity.[39] The
66 low expression levels of some noncanonical ORFs further increases the difficulties in identification.[3,27]
67 Perhaps as a result of these limitations, less than half of the noncanonical ORFs detected as translated in
68 humans are reproducible across studies.[31]

69 Here, we designed an approach to increase sensitivity in detection of both translation and selection
70 among noncanonical ORFs. We address the challenges in detecting translation through the development
71 of a ribo-seq analysis framework (iRibo) that identifies signatures of translation with high sensitivity and
72 high specificity by integrating data across hundreds of experiments from many published studies. This
73 facilitates detection of sequences that are short or poorly expressed. We address the challenges in
74 detecting selection through a comparative genomics framework that analyzes translated sequences
75 collectively across evolutionary scales within- and between-species.

76 We applied our approach to define a "reference translatome" for the model organism *Saccharomyces*
77 *cerevisiae* and to characterize the biological significance of noncanonical ORFs. Using iRibo, we identified
78 ~19,000 noncanonical ORFs translated at high confidence and established the dependence of
79 noncanonical translation on both genomic context and environmental condition. Using genomic data
80 both within strains of *S. cerevisiae* and across budding yeast species[40,41], we identified a handful of
81 undiscovered conserved genes within the yeast noncanonical translatome. However, we find that most
82 of the yeast noncanonical translatome is evolutionarily young and of *de novo* origin, having emerged
83 recently from noncoding sequence. These young ORFs differ greatly from conserved genes in their
84 length, amino acid composition, and expression level, and show no signs of purifying selection.
85 Nevertheless, we report experimental evidence based on fluorescent protein tagging and conditional
86 loss-of-function fitness measurements showing that translation of evolutionarily young noncanonical
87 ORFs can generate stable protein products and affect cellular phenotypes. We thus propose that much
88 of the noncanonical translatome is composed of neither translational noise nor conserved genes, but
89 rather of a distinct class of evolutionarily short-lived coding sequences with important biological
90 implications. This "transient translatome" is larger than, and categorically distinct from, the conserved
91 translatome made mostly of canonical protein-coding genes that have been studied for decades.

92 **Results**

`

93 **An integrative approach to defining the translatome**

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

`

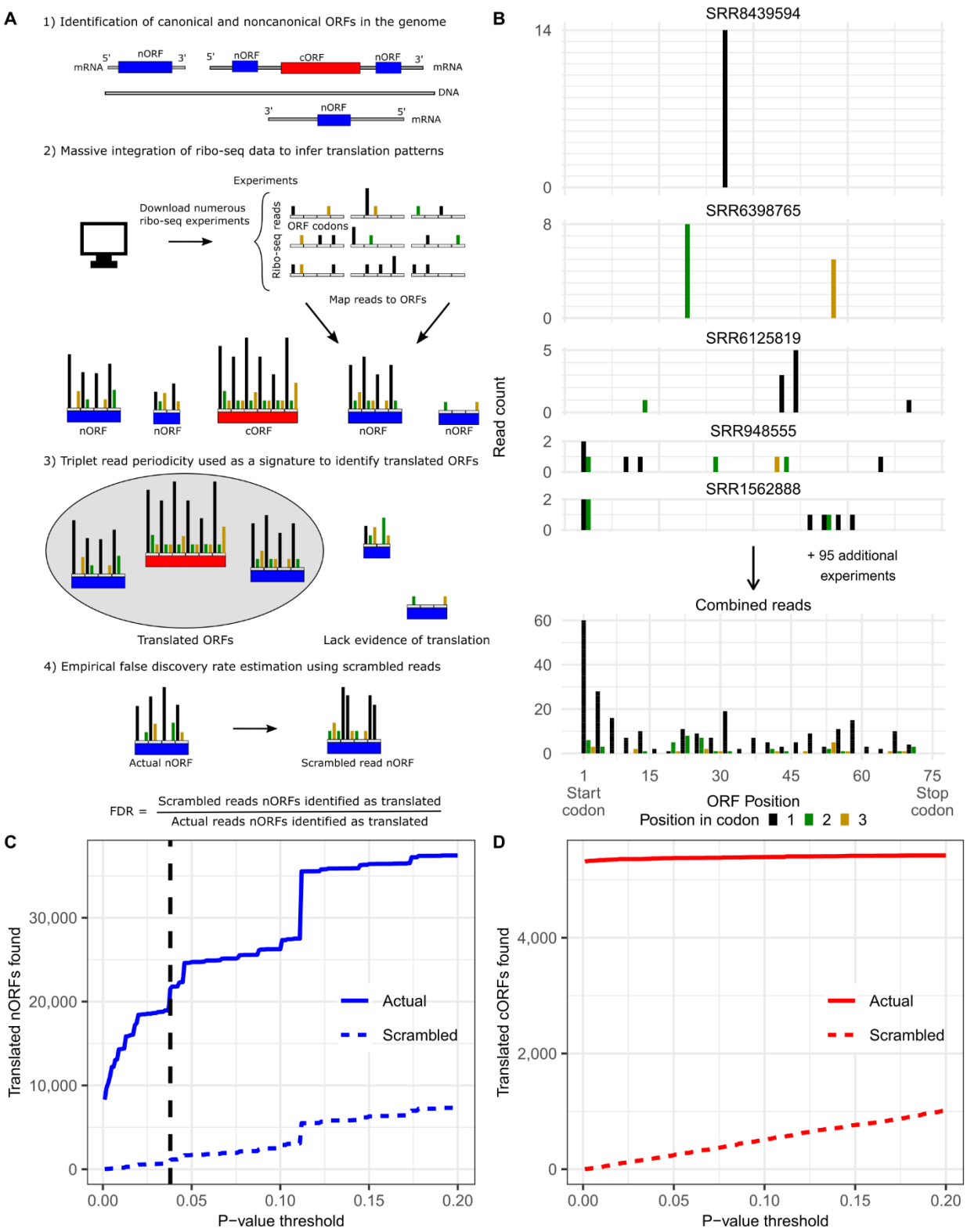


Figure 1: The iRibo framework enables detection of thousands of noncanonical translated sequences. A) The iRibo framework. 1) Candidate ORFs, both canonical (cORFs; red) and noncanonical (nORFs; blue), are identified in the genome. 2) Reads aggregated from published datasets are then mapped to these

`

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 (y-axis) across an example nORF located on chromosome II, 604674-604748 (x-axis). The top five graphs correspond to five individual experiments with reads mapping to the ORF while the bottom graph includes all reads integrated across all experiments. Reads are colored according to their position on the codon. **C)** iRibo identifies 18,953 translated nORFs at 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 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.

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.

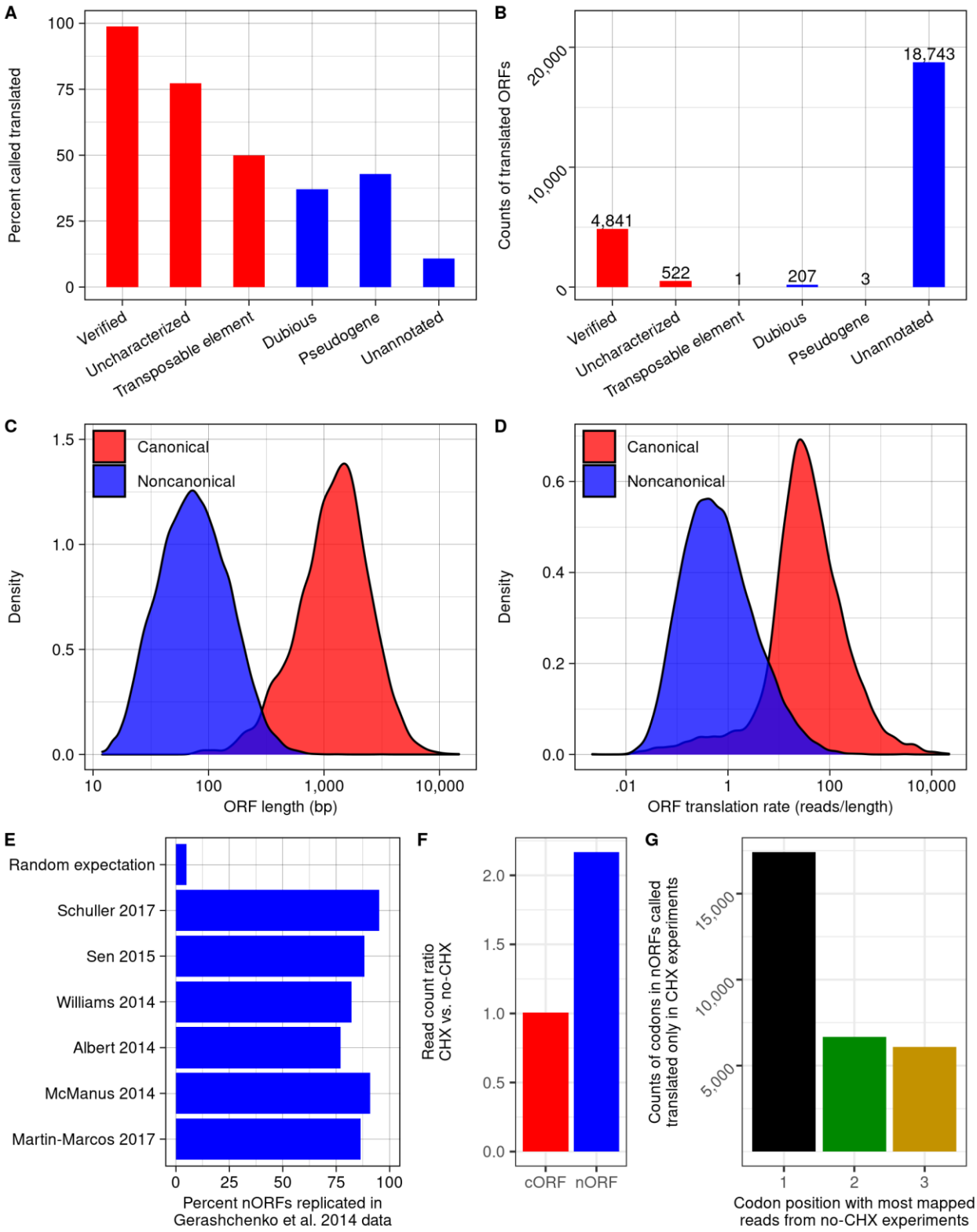
We used iRibo to identify translated ORFs across the *S. cerevisiae* genome (**Supplementary Figure 1**). First, we constructed the set of candidate ORFs by collecting all genomic sequences at least three codons in length that start with ATG and end with a stop codon in the same frame. For ORFs overlapping in the same frame, only the longest ORF was kept. Each candidate ORF was classified either as canonical (cORF), if it was annotated as "verified," "uncharacterized," or "transposable element" in the Saccharomyces Genome Database (SGD)[43] or as noncanonical (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 translation status for candidate ORFs using data from 412 ribo-seq experiments across 42 studies (**Supplementary Table 1, Supplementary Table 2**).

As expected, integrating data from many experiments allowed for 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% false discovery rate (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. 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 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.

`



150

151 **Figure 2: The noncanonical yeast translatome is larger than the canonical. A)** A majority of cORFs, and
152 a minority of nORFs, are translated. The percent of ORFs (y-axis) in each Saccharomyces Genome

`

153 Database annotation class that are detected as translated by iRibo, with canonical classes indicated in
154 red and noncanonical in blue. **B)** Unannotated nORFs make up a large majority of translated sequences.
155 The number of ORFs (y-axis) of each annotation class (x-axis) that are detected using iRibo. **C)** nORFs are
156 shorter than cORFs. ORF length distributions for translated cORFs and nORFs. **D)** nORFs are translated at
157 lower rates than cORFs. Distribution of translation rate (in-frame ribo-seq reads per base) for translated
158 cORFs and nORFs. **E)** Translation calls are highly reproducible. For six large studies (y-axis), the
159 proportion of nORFs identified using reads from that study that are replicated using reads from the
160 largest study, Gerashchenko et al. 2014[44] (x-axis). Random expectation is the proportion that would be
161 expected to replicate by chance. **F)** CHX facilitates detection of translated nORFs. Ratio of total ribo-seq
162 read counts mapping to cORFs or nORFs in experiments with vs. without CHX treatment. Note that the
163 same number of total reads are sampled from each condition. **G)** nORFs identified as translated only
164 with CHX nevertheless show preference for the first codon position in its absence. Among nORFs
165 identified as translated by iRibo only in the CHX condition, all codons among these nORFs are classed
166 based on which of the three positions in the codon have the most reads from experiments without CHX.

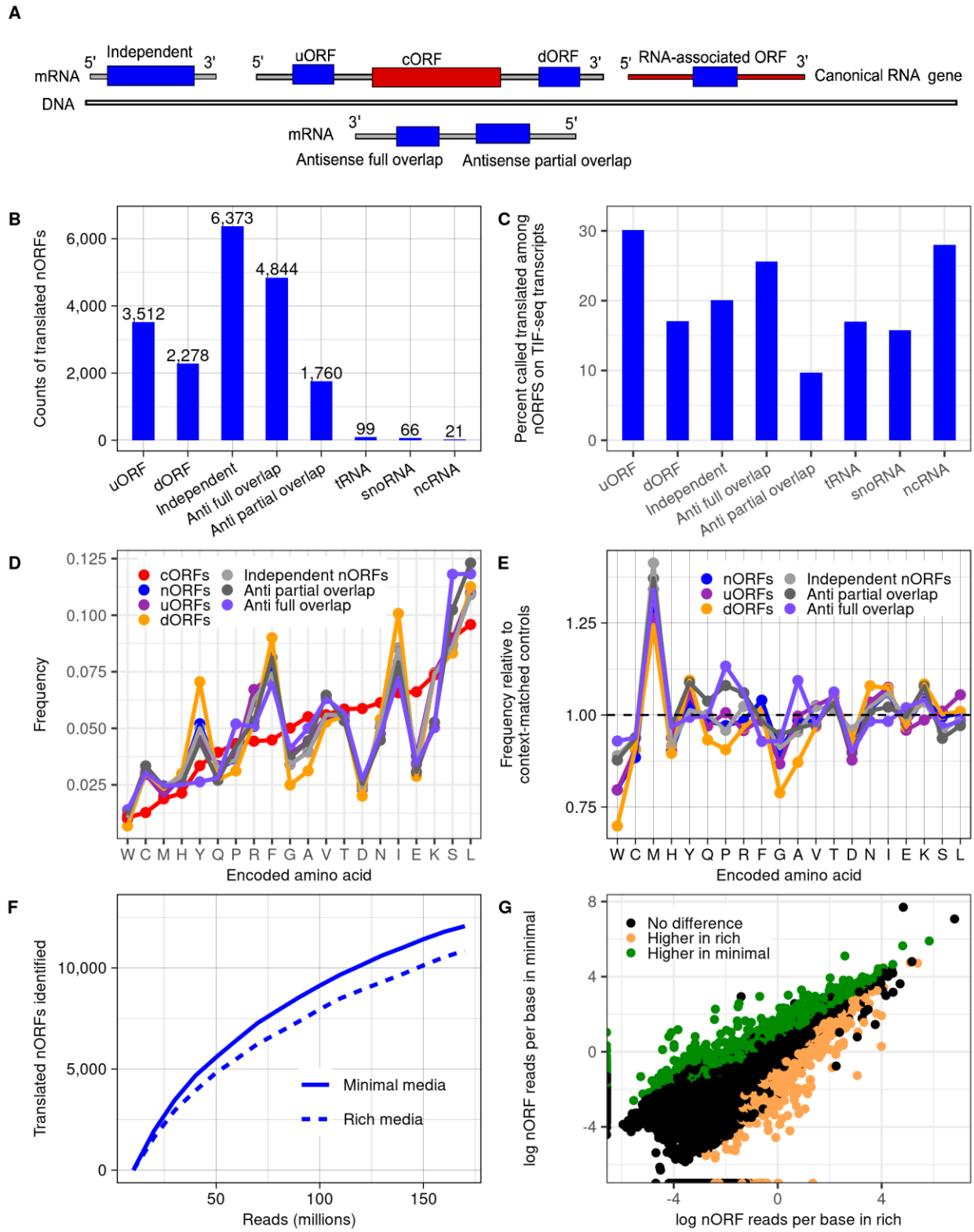
167 To assess replicability in translation calls for nORFs, we applied iRibo separately to each of the largest
168 individual studies by read count. We then counted, among the nORFs that could be inferred to be
169 translated using only the reads in each study, how many were also found in the largest study,
170 Gerashchenko and Gladyshev, 2014.[44] For all studies, at least 75% of detected ORFs were also detected
171 in the largest study (**Figure 2E**). In general, translation rates among ORFs were highly correlated among
172 independent studies (**Supplementary Figure 2**). These observations demonstrate that noncanonical
173 translation patterns are highly reproducible, suggesting that they are driven by regulated biological
174 processes rather than technical artifacts or stochastic ribosome errors.

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

187 **Noncanonical translation patterns depend on genomic and environmental context**

`

188 We examined to what extent translation of nORFs depends on genomic context. We classified nORFs as:
189 upstream nORFs (uORFs) located on the 5' untranslated regions of transcripts containing cORFs;
190 downstream nORFs (dORFs) located on the 3' untranslated regions of transcripts containing cORFs;
191 intergenic nORFs that do not share transcripts with cORFs (independent); nORFs antisense to a cORF
192 and located entirely within the bounds of that cORF (antisense full overlap); nORFs overlapping the
193 boundaries of a cORF on the opposite strand (antisense partial overlap) (**Figure 3A**). Additionally, for
194 nORFs sharing a transcript with an RNA gene, the nORF was classified based on the type of RNA gene.
195 The transcripts used for these classifications were derived from the TIF-seq data collected by Pelechano
196 et al. 2014[47], which provide transcript start and end sites.



197

198 **Figure 3: Noncanonical translation patterns depend on both genomic and environmental context. A)**
199 Potential genomic contexts for nORFs in relation to nearby canonical genes. Transcripts are defined from
200 published TIF-seq data[47]. **B)** Counts of translated nORFs identified by iRibo (y-axis) in each considered

genomic context (x-axis), 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 (y-axis) in each genomic context considered, among nORFs completely covered by a TIF-seq transcript (x-axis). **D)** Amino acid composition of translated nORFs differs from that of translated cORFs and depends on 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. **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, the log reads per base in each condition is indicated. Total read count in each condition was held constant by randomly sampling reads from YPD experiments until the read count in SD experiments was matched. 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.

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, while 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 while 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 prior research[49], uORFs were translated at significantly higher rates than other classes, with 30% of considered uORFs found to be translated compared to 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 to 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 genomic context ($p<10^{-10}$ for any comparison between cORF amino acid

`

239 frequencies and nORF frequencies in a given context, chi-square test; **Figure 3D**). The translation
240 products of nORFs present a large excess of cysteine, phenylalanine, isoleucine, arginine, and tyrosine
241 and deficiency in alanine, asparagine, glutamic acid, and glycine relative to cORFs. Notably, aside from
242 arginine, the amino acids with large excess in nORFs relative to cORFs are all hydrophobic. Amino acid
243 frequencies of nORFs appear to largely reflect underlying DNA sequence composition biases that differ
244 between the distinct genomic contexts. Indeed, within each genomic context, amino acid frequencies of
245 translated nORF are generally similar (with less than 15% difference in frequency) to that of length- and
246 context- matched nORFs that lack evidence of translation, though they do show significant differences
247 ($p < 10^{-10}$ for all contexts, chi-square test; **Figure 3E**). The most striking differences include a large excess
248 of methionine residues and a deficiency in tryptophan and glycine residues among translated nORFs
249 compared to the untranslated control group.

250 In addition to genomic context, we assessed how environmental context affects noncanonical
251 translation. To this aim, we leveraged the power of iRibo to construct separate datasets of nORFs found
252 translated in rich media (YPD) or in nutrient-limited minimal media (SD) (**Supplementary Table 3**).
253 Previous research has reported an increase in detected noncanonical translation events relative to
254 canonical translation events in response to starvation.[1,3] Consistent with these results, more nORFs were
255 identified as translated in minimal than in rich media at equal read counts (**Figure 3F**). Furthermore,
256 2968 nORFs were supported by a significantly higher number of in-frame reads in minimal media than
257 rich media while the converse was true for only 1265 nORFs (5% FDR, Fisher's exact test with Benjamini-
258 Hochberg procedure[48]; **Figure 3G**). These results suggest that starvation conditions may increase
259 noncanonical translation, or alternatively that noncanonical translation is less affected by the general
260 translation inhibition that occurs in starvation conditions.[50] Either way, these results support the
261 hypothesis that nORF translation is regulated in response to changing environments.

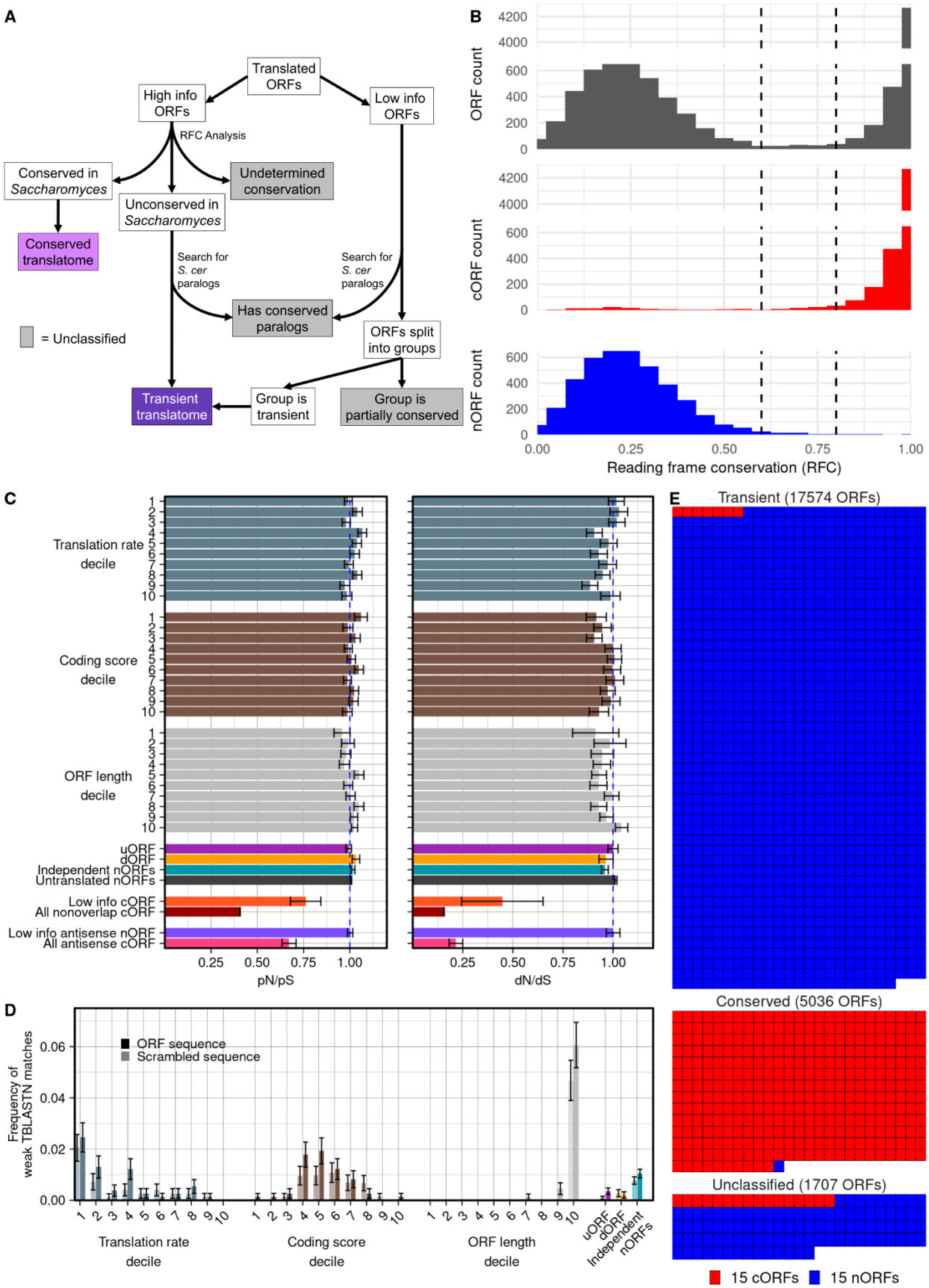
262 **Two translatomes, transient and conserved**

263 Given the large numbers of nORFs translated in the yeast genome, we next sought to assess the
264 biological significance of this translation by determining the extent to which these nORFs are evolving
265 under selection. We assessed selection acting on nORFs, as well as on cORFs for purpose of comparison,
266 across three evolutionary scales. At the population level, we analyzed 1011 distinct *S. cerevisiae* isolates
267 sequenced by Peter et al. 2018.[40] At the species level, we compared *S. cerevisiae* ORFs to their orthologs
268 in the *Saccharomyces* genus, a taxon consisting of *S. cerevisiae* and its close relatives.[51] To detect long
269 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. 2018.[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 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 power to detect the presence of selection but does not enable identification of the individual ORFs under selection. The "high information" set consisted of the ORFs that 1) have homologous DNA sequence 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 these criteria are sufficient to distinguish ORFs evolving under strong purifying selection (**Supplementary Figure 3**). 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.

`


A
Translated ORFs
High info ORFs
Low info ORFs
RFC Analysis
Conserved in Saccharomyces
Undetermined conservation
Unconserved in Saccharomyces
Conserved translatome
Search for S. cer paralogs
Search for S. cer paralogs
Has conserved paralogs
= Unclassified
ORFs split into groups
Transient translatome
Group is transient
Group is partially conserved
B
ORF count
cORF count
nORF count
Reading frame conservation (RFC)
C
Translation rate decile
Coding score decile
ORF length decile
uORF
dORF
Independent nORFs
Untranslated nORFs
Low info cORF
All nonoverlap cORF
Low info antisense nORF
All antisense cORF
pN/pS
dN/dS
D
Frequency of weak TBLASTN matches
ORF sequence
Scrambled sequence
Translation rate decile
Coding score decile
ORF length decile
uORF
dORF
Independent nORFs
E
Transient (17574 ORFs)
Conserved (5036 ORFs)
Unclassified (1707 ORFs)
15 cORFs
15 nORFs

287

`

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), only cORFs (middle) and only nORFs (bottom). See 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 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 are the set of all nORFs in the genome not called as translated by iRibo. Low information nonoverlapping cORFs are assembled into a single group, with the set of all nonoverlapping cORFs shown for comparison. Low information antisense nORFs were also assembled into a single group, with the set of all antisense cORFs shown for comparison. pN/pS is calculated from variation at each ORF codon among *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 < .05) in each group of nORFs (dark bars) and negative controls (light bars) consisting of the sequences of the nORFs of each group randomly scrambled. Error bars indicate standard errors estimated from bootstrapping. **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.

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. 2003[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 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 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,

`

328 with only 1.2% of ORFs intermediate between these values (**Figure 4B)**. The bimodal distribution of RFC
329 among translated ORFs is similar to the bimodal distribution observed among all candidate ORFs,
330 regardless of translation status (**Supplementary Figure 4A**), as observed previously by Kellis et al. 2003.[20]
331 The modes of the distribution largely correspond to annotation status, with 96.7% of cORFs having RFC >
332 0.8 and 98.5% of nORFs having RFC < 0.6. This pattern holds when evaluated only in the last 100 bp of
333 ORFs, suggesting that it is not affected by potential incorrect inference of nORF start positions
334 (**Supplementary Figure 4B**). The clean separation between well-conserved and poorly-conserved ORFs
335 indicate that most high-information ORFs can be straightforwardly classified into one of the two groups,
336 and thus nearly all high-information nonoverlapping nORFs can be placed in the poorly-conserved class.
337 High RFC among antisense ORFs does not demonstrate selection on the ORF itself, as it might be caused
338 by selective constraints on the opposite-strand gene, but low RFC still indicates lack of purifying
339 selection. A majority of antisense translated nORFs (64.1%) have RFC <0.6, indicating that most are not
340 preserved by selection across the genus (**Supplementary Figure 4C**). Overall, we find no evidence for
341 purifying selection acting on nORFs on a large scale.

342 In light of the general correspondence between annotation and conservation, the exceptions are of
343 interest: 110 cORFs had RFC < 0.6, and 13 nonoverlapping unannotated nORFs had RFC > 0.8. To further
344 assess conservation among these two sets of ORFs, we performed a BLAST analysis (using both BLASTP
345 and TBLASTN with default parameters) to search for homologs of each ORF among the budding yeast
346 genomes assembled by Shen et al. 2018.[41] Among the 110 cORFs with low RFC, 101 also had no detected
347 homology to other *S. cerevisiae* genes or any budding yeast genome outside of *Saccharomyces*,
348 indicating that these are likely annotated ORFs of recent *de novo* origin. For the 13 nORFs with high RFC,
349 several additional lines of evidence suggest that these are indeed evolving under purifying selection
350 (**Table 1)**. For nine of the thirteen, we identified a homolog among budding yeast genomes outside of
351 the *Saccharomyces* genus by either a BLASTP or TBLASTN search. The existence of a homolog in a
352 distantly related species indicates that the ORF existed in the common ancestor of *S. cerevisiae* and that
353 distant species, implying long-term preservation of the ORF by purifying selection in both lineages. We
354 also performed pN/pS analysis for each ORF on *S. cerevisiae* isolates and dN/dS analysis for each ORF
355 among the *Saccharomyces* genus species (**Table 1**). A pN/pS or dN/dS ratio significantly below 1
356 indicates purifying selection on the ORF amino acid sequence among *S. cerevisiae* strains or among
357 *Saccharomyces* genus species, respectively, while a ratio above 1 indicates positive selection. By these
358 measures, two ORFs showed significant evidence of purifying selection by pN/pS and three by dN/dS

`

359 (**Table 1**). Thus, a small number of nORFs appear to be evolving under selection, indicating significant
360 biological roles.

361 We next assessed selection among the full set of nORFs (both high and low information) at the
362 subphylum scale, searching for addition nORFs that exhibited long term conservation and thus purifying
363 selection. Towards this end, we searched for distant homologs of all translated nonoverlapping *S.*
364 *cerevisiae* nORFs using TBLASTN against budding yeast genomes in the *Saccharomycotina* subphylum,
365 excluding species in the *Saccharomyces* genus. After excluding matches that appeared non-genic or
366 pseudo-genic (**Supplementary Figure 5**) we identified a single additional nORF with both distant
367 TBLASTN matches and recent signatures of purifying selection (dN/dS = 0.5, p=.039 for test of difference
368 from 1.0): YBR012C, annotated as "dubious" on SGD. Thus, combining the 13 nORFs that appeared
369 conserved by RFC analysis and the single additional nORF found using TBLASTN, we identified 14
370 translated nORFs that show evidence of preservation by purifying selection (**Table 1**).

371 To analyze collective evidence of selection among "low information" ORFs, we first divided low
372 information nonoverlapping nORFs (7,855 nORFs, after excluding those with homology to conserved *S.*
373 *cerevisiae* cORFs) according to properties that we expected to be potentially associated with selection:
374 rate of translation (as measured by ribo-seq reads mapped to the first position within codons divided by
375 the length of the ORF), coding score[28,52] (a measure of sequence similarity to annotated coding
376 sequences), ORF length, and genomic context. For each group, we calculated the pN/pS ratio among
377 1,011 *S. cerevisiae* isolates[40] and the dN/dS ratio based on alignments of the *S. cerevisiae* ORFs with their
378 orthologous DNA sequence in *S. paradoxus*. We also analyzed low information nonoverlapping cORFs
379 (22 cORFs) in the same manner. For low information antisense nORFs (3642 nORFs; only 2 cORFS fell in
380 this category and were not analyzed), we calculated the pN/pS and dN/dS ratios restricted to
381 substitutions that were synonymous on the opposite-strand cORF.[53,54] Unlike the RFC, dN/dS and pN/pS
382 analyses conducted above on individual high information ORFs, these analyses were conducted by
383 aggregating substitutions among all low information ORFs in each group to assess evidence for selection
384 (i.e., a ratio significantly different from 1) within the group as a whole. We expected that, if low
385 information nORFs were evolving under selection, then more highly translated ORFs, longer ORFs, and
386 ORFs with coding scores more similar to conserved genes, would be enriched in biologically relevant
387 nORFs and thus show stronger signatures of selection. Low information nonoverlapping cORFs did show
388 collective pN/pS and dN/dS ratios significantly below 1, indicating that some ORFs in this group are
389 evolving under purifying selection (**Supplementary Table 4**, **Figure 4C**). In contrast, for all groups of low

`

390 information nORFs examined, we observed no significant difference in the pN/pS or dN/dS ratio from 1,
391 providing no evidence for either purifying or positive selection (**Supplementary Table 4, Figure 4C**).

392 Finally, we assessed collective evidence of long-term evolutionary conservation in each group. To do
393 this, we calculated the frequency of weak TBLASTN matches (e-values between 10^{-4} and .05, above our
394 threshold for homology detection at the individual level) of ORFs in each group to the *Saccharomycotina*
395 subphylum genomes outside of *Saccharomyces* as compared to a negative control set consisting of
396 scrambled sequences of the ORFs in each group. Applying this strategy to the full set of 362
397 nonoverlapping cORFs that lacked TBLASTN matches outside *Saccharomyces* at the e-value < 10^{-4} level,
398 we found a large excess of weak matches relative to controls (p=.0001, Fisher's exact test;
399 **Supplementary Figure 6**), demonstrating the ability of this approach to detect faint signals of homology
400 within a group of ORFs. However, we identified no significant difference in the frequency of weak
401 TBLASTN hits between any nonoverlapping nORF group and scrambled controls (**Figure 4D**), nor among
402 nonoverlapping nORFs overall (p>.05, Fisher's exact test). The lack of a significant result does not
403 exclude the possibility that a small subset of short conserved nORFs could be lost in the noise of a much
404 larger set of nORFs without distant homology. However, our TBLASTN, dN/dS and pN/pS analyses
405 altogether indicate that ORFs evolving under strong purifying selection are not a major component of
406 the yeast noncanonical translatome.

407 Overall, our analyses distinguish two distinct yeast translatomes: a conserved, mostly canonical
408 translatome with intact ORFs preserved by selection; and a mostly noncanonical translatome where
409 ORFs are not preserved over evolutionary time. This distinction is rooted in evolutionary evidence rather
410 than annotation history. We thus propose to group the translated ORFs that showed neither evidence of
411 selection nor homology to conserved ORFs in our high-information and low-information sets as the
412 "transient translatome." The "transient translatome" designation indicates membership in a set of ORFs
413 that are expected to exist in the genome for only a short time on an evolutionary scale, though we
414 cannot be certain that any particular translated ORF that currently exists in the yeast genome will be
415 rapidly lost. The transient translatome includes 4,051 nonoverlapping and 1,923 antisense nORFs
416 identified as not preserved by selection using RFC analyses and having no conserved homologs, along
417 with 86 nonoverlapping and 15 antisense cORFs (total 101) matching the same criteria. Also included are
418 7,855 nonoverlapping and 3,644 antisense nORFs that lack sufficient information to analyze at the
419 individual level but were found to show no selective signal in group-level analyses. Together, this set of

`

420 17,574 ORFs that are translated yet likely evolutionarily transient makes up 72% of the yeast reference
421 translatome (**Figure 4E**).

422 **Transient cORFs are representative of the transient translatome overall**

423 By general theory and practice in evolutionary genomics, the lack of selective signal suggests that the
424 transient translatome does not meaningfully contribute to fitness.[55] Surprisingly, however, 101 cORFs
425 belong to the transient set, suggesting that some transient ORFs have phenotypes. To assess whether
426 these cORFs are representative of the transient translatome overall, we compared their evolutionary
427 and sequence properties with those of transient "dubious" nORFs (annotated but presumed
428 nonfunctional) and transient unannotated nORFs. We found transient cORFs, transient dubious nORFs
429 and transient unannotated nORF to all have pN/pS ratios indistinguishable from 1.0 (**Figure 5A**),
430 providing no evidence for purifying selection. Similarly, the average nucleotide diversity (mean number
431 of nucleotide differences per site between pairs of isolates) of transient cORFs was indistinguishable
432 from that of transient nORFs or untranslated controls, and much higher than that of conserved cORFs
433 (**Figure 5B**). In addition, no class of transient ORFs showed differences from each other in RFC between
434 *S. cerevisiae* and *S. paradoxus* (**Figure 5C**), rate of translation (**Figure 5D**) or coding score (**Figure 5E)**.

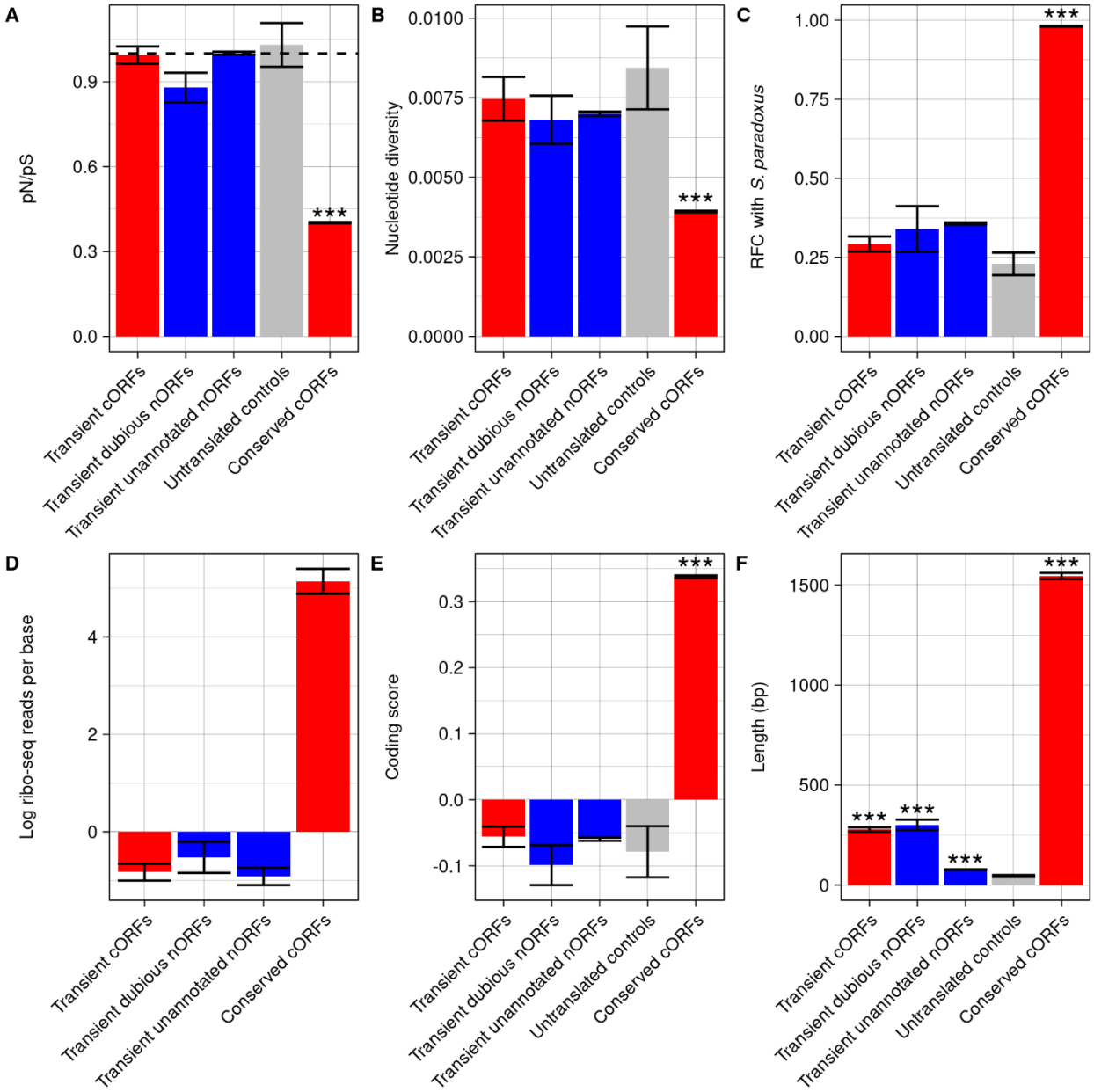


Figure 5: Canonical and noncanonical transient ORFs are similar except for their length. Properties of transient cORFs (n=86), transient dubious nORFs (n=25), transient unannotated nORFs (n=12,160), untranslated controls (n=100) and conserved cORFs (n=5130). Untranslated controls consist of nonoverlapping ORFs that would be grouped in the transient class (RFC <.6) but are not inferred to be translated based on ribo-seq evidence. Conserved cORFs are nonoverlapping cORFs with RFC >.8. All groups are restricted to nonoverlapping ORFs. Error bars represent standard error. Stars indicate significant differences from untranslated controls by permutation test: P-value <.001: ***. **A)** pN/pS values for each group among *S. cerevisiae* strains. **B)** Average nucleotide diversity (π) among each group. **C)** Average reading frame conservation between *S. cerevisiae* and *S. paradoxus* ORFs. **D)** Average ribo-seq reads per base (logged), considering only in-frame reads. Unannotated nORFs and untranslated controls are sampled to match the length distribution of transient cORFs. **E)** Coding scores for each group. **F)** ORF lengths in nucleotides for each group.

`

448

449 The only distinguishing property between classes of transient ORFs was their length: annotated
450 transient cORFs and transient "dubious" nORFs are much longer on average than unannotated transient
451 nORFs (**Figure 5F**). This is a consequence of the history of annotation of the *S. cerevisiae* genome, where
452 a length threshold of 300 nt was set for annotation of ORFs.[56,57] The sharp 300 nt threshold is still clearly
453 reflected in annotations. For example, genome annotations include 96% of nonoverlapping transient
454 ORFs in the 300-400 nt range (55/57), but only 4% in the 252-297 nt range (4/101). Given that transient
455 nORFs resemble transient cORFs in all respects besides length, we hypothesized that numerous never-
456 studied transient nORFs are just as likely to have phenotypes as transient cORFs.

457 **Transient ORFs are detected in the cell and mediate diverse phenotypes**

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

463 We first assessed whether the proteins encoded by transient ORFs can be detected in the cell. We
464 examined the CYCLoPs database[58,59], the C-SWAT tagging library[60], and the YeastRGB database[61], which
465 contain collections of fluorescently tagged proteins expressed from their native promoters and
466 terminators, including both cORFs and dubious nORFs. Together these studies detected expression of a
467 fluorescent protein product for 90 of 93 (97%) transient cORFs tested, along with 37 of 41 (90%)
468 transient dubious nORFs tested (**Figure 6A**). For comparison, we C-terminally tagged 21 highly expressed
469 unannotated transient nORFs with mNeonGreen at their endogenous locus and examined their
470 expression using microscopy. We detected 8 of 21 tagged nORF proteins (38%) (**Figure 6A-B,**
471 **Supplementary Figure 7**). Thus, translation of tagged proteins can be detected for both annotated and
472 unannotated transient ORFs.

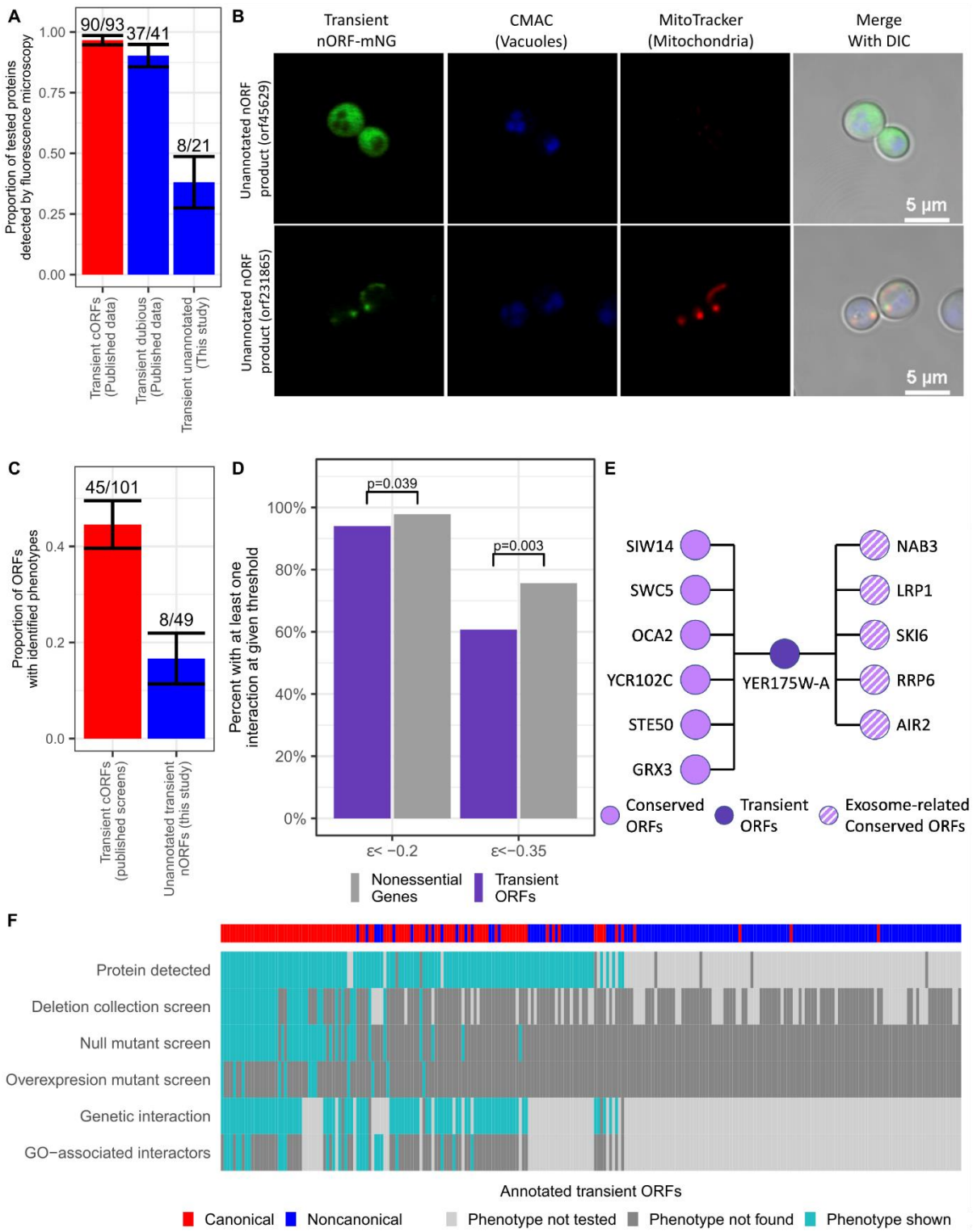


Figure 6: Transient nORFs and cORFs can be detected in the cell and exhibit phenotypes. A) Transient ORFs are detected by fluorescent microscopy. For cORFs or dubious nORFs, the proportion of proteins expressed by transient ORFs detected in the C-SWAT[60], CYCLoPs[59], or YeastRGB[61] microscopy datasets out of those tested. For unannotated transient nORFs, the proportion detected by mNeonGreen tagging

`

478 in this study. Error bars indicate standard error of the proportion. **B)** Tagged unannotated transient
479 nORFs show varied sub-cellular localizations. Microscopy images of unannotated transient nORFs taken
480 at 100X. Left panel shows the expression of the nORFs tagged with mNeonGreen, middle panels the
481 dyes CMAC Blue and MitoTracker Red for mitochondria and vacuoles identification, respectively, and
482 the right panel the merge all the above channels with DIC. Top panel show the nORF (orf45629) with a
483 cytosolic expression and the bottom panel the nORF (orf231865) with expression localizing to the
484 mitochondria. **C)** Loss of transient nORFs can affect phenotype despite lack of evolutionary conservation
485 The proportion of deletion mutants with reported loss-of-function phenotypes in two groups: transient
486 cORFs in published deletion mutant screens, and transient nORFs assayed in this study. Reported
487 phenotypes in published data was taken from literature associated with each ORF on SGD. In this study,
488 deleterious deletion mutant phenotypes were identified from a high-throughput colony fitness screen in
489 six stress conditions using a 5% FDR threshold. **D)** Transient ORFs engage in epistatic relationships. The
490 percent of transient ORFs and nonessential genes with at least one genetic interaction at given
491 threshold are shown. Differences between groups were tested using Fisher's exact test. **E)** Genetic
492 interactions of the transient ORF YER175W-A. Five interactors are related to exosome (striped circles). **F)**
493 Presence of phenotypes among annotated transient ORFs. "Protein detected" indicates that the ORF
494 product was found in either the C-SWAT or CYCLoPs database. Phenotypes of deletion collection,
495 deletion and overexpression screens were taken from reported findings in the yeast experimental
496 literature (**Supplementary Table 5**). "Genetic interaction" indicates a statistically significant genetic
497 interaction with $\varepsilon < -0.2$, and "GO-associated interactors" indicates a GO enrichment was found among
498 significant interactors at 5% FDR.

499

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

509 To determine whether transient cORFs that are not well described also affect phenotypes, we examined
510 all literature listed as associated with the ORF on SGD. Many of these transient cORFs have direct
511 evidence of phenotype (**Supplementary Table 5**). Of 101 transient cORFs, 45 were reported to have
512 deletion mutant phenotypes (i.e., a phenotype observed when the ORF is deleted) and 12 to have
513 overexpression phenotypes. Overall, we found phenotypes reported in the literature for 50 of 101
514 transient cORFs (50%).

`

515 As unannotated transient nORFs have not been systematically investigated for phenotype, we sought to
516 experimentally determine whether these ORFs too might have deletion mutant phenotypes. We thus
517 conducted a deletion mutant screen of 49 unannotated transient nORFs selected for high translation
518 rate and to avoid intersecting cORFs, annotated ncRNAs, or promoters (200 bp upstream of canonical
519 genes). We fully deleted the nORF using homologous recombination and each strain was assayed for
520 colony growth in seven conditions. Eight nORF deletion mutant strains showed deleterious phenotypes
521 in at least one condition at a 5% FDR (**Figure 6C**, **Supplementary Table 6**). Thus, loss of transient nORFs,
522 as with cORFs, can affect phenotype despite lack of evolutionary conservation.

523 To begin to understand the specific biological processes in which transient ORFs might be involved, we
524 leveraged the large yeast genetic interaction network assembled in Costanzo et al. 2016.[69] This dataset
525 includes 75 non-overlapping transient cORFs and 9 non-overlapping dubious transient nORFs. Genetic
526 interaction strength, Ɛ, measures the difference between the observed fitness of a strain in which two
527 genes are deleted and the expected fitness given the fitness of the two single gene deletion strains; a
528 negative value of high magnitude suggests that the two mutated genes are involved in related
529 processes. Of the 84 transient ORFs in the dataset, 79 (94%) have at least one negative genetic
530 interaction at the high-stringency cut-off defined by Costanzo et al.[69] (Ɛ<-0.2 and p-value<0.05) and 51
531 (61%) have synthetic lethal interactions (Ɛ<-0.35 and p-value<0.05) as defined in that study (**Figure 6D**).
532 This was only a slightly lower rate than for conserved non-essential ORFs, 98% of which had negative
533 interactions at the high stringency cut-off and 76% of which had synthetic lethal interactions. At the high
534 stringency threshold, 27 transient ORFs were found to interact with groups of related genes enriched in
535 specific gene ontology (GO) terms (5% FDR; **Supplementary Table 7**). For example, the interactors of
536 YER175W-A are associated with the GO category "cryptic unstable transcript (CUT) metabolic processes"
537 with high confidence, and five of its eleven interactors are components or co-factors of the exosome
538 (**Figure 6E**), indicating likely involvement in CUT degradation or a closely related post-transcriptional
539 regulation pathway. Other enrichments included diverse processes such as "mating projection tip" or
540 "Golgi sub-compartment". In contrast, when we applied GO enrichment analysis to the full set of genes
541 that interact with any transient ORF, no significant enrichment was observed. These results suggest that
542 transient ORFs in general do not participate in one shared biological process, but rather are involved in a
543 wide variety of cellular processes.

544 Overall, we uncovered evidence that 131 of 250 (53%) annotated transient ORFs have at least one
545 indicator of biological significance (detection of a protein product, a reported phenotype in a screen, or

`

546 a genetic interaction in the Costanzo et al. 2016[69] network) (**Figure 6F**). Additionally, we demonstrate
547 that unannotated transient ORFs encode proteins that can be detected in the cell (38% of tested in this
548 study) and influence cellular fitness when deleted (17% of tested in this study). Given that this class has
549 received almost no study compared to the great number of experiments that have been conducted on
550 cORFs, the number of transient ORFs with biological relevance may be substantially larger than that
551 which has been annotated.

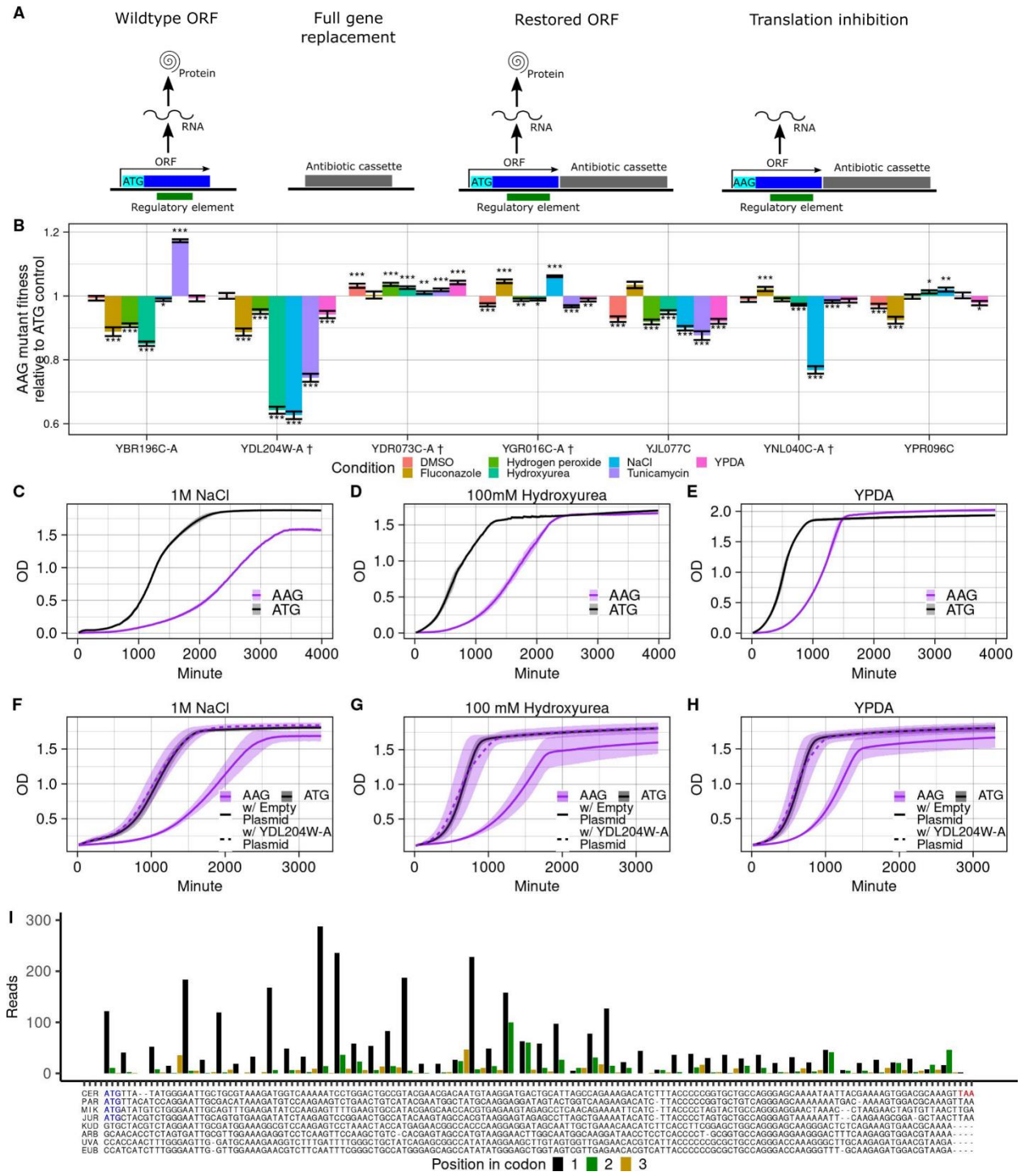
552 A limitation on much of the experimental evidence available on deletion mutant phenotypes is that
553 most deletion mutant and genetic interaction screens are based on a full gene replacement strategy in
554 which the entire ORF is lost, leaving the possibility that some deletion phenotypes could be caused by
555 loss of a ncRNA or a DNA regulatory element located at the same position as the ORF rather than loss of
556 the ORF translation (**Figure 7A**). To examine this possibility, we constructed a set of strains where the
557 ORF start codon ATG was replaced with an AAG codon while keeping the rest of the ORF intact. This set
558 included three transient cORFs that have previously been characterized on the basis of overexpression
559 or full deletion mutants, *ICS3*[68], YPR096C[67], and YBR196C-A[65], along with four transient nORFs that
560 showed strong deleterious phenotypes in our full ORF deletion screen (**Supplementary Table 8**; *HUR1*
561 and *MDF1* were not tested because they overlap other cORFs). Each deletion strain was tested in seven
562 environmental conditions. The single nucleotide ATG→AAG mutation caused significantly reduced
563 colony size for all three transient cORFs tested and for three of four transient nORFs tested in at least
564 one condition (**Figure 7B**). We gave these three nORFs systematic names YDL204W-A, YGR016C-A, and
565 YNL040C-A. The remaining nORF, YDR073C-A, showed a weak beneficial phenotype from the ATG→AAG
566 mutation in some conditions, as did two other nORFs, YGR016C-A and YNL040C-A. The largest growth
567 reductions were observed from disabling translation in YDL204W-A: this strain reached only 64% of
568 wildtype growth in hydroxyurea and 63% in high salt concentration, with a smaller reduction to 94%
569 growth in rich media (YPDA). These growth defects were also observed in a liquid growth setting (**Figure**
570 **7C-E**). To confirm that these phenotypes were caused by loss of the YDL204W-A protein rather than *cis*
571 effects at the locus, we expressed the intact YDL204W-A ORF from a plasmid in the ATG→AAG mutant
572 strain. Plasmid expression of the ORF fully restored the wildtype phenotype in the mutant strains (**Figure**
573 **7F-H**), providing further evidence that blocking YDL204W-A translation causes a loss of function
574 phenotype mediated by loss of the encoded protein.

575 In our translation dataset, YDL204W-A has a translation rate at the top percentile among transient ORFs
576 (**Figure 7I**), higher than 10% of cORFs. Comparing its sequence to the homologous region of other

`

577 *Saccharomyces* genus species, only *S. paradoxus* and *S. mikatae* have a homologous start codon, but a
578 2bp insertion in *S. cerevisiae* results in a frameshift such that little of the ORF is shared in any other
579 species (**Figure 7I**); thus, this ORF has a reading frame conservation score of only 0.2 (**Table 2**). The other
580 transient ORFs with phenotypes induced by an ATG→AAG mutation also showed no signs of selection
581 **(Table 2)**. Thus, our results exemplify the potential for unannotated coding sequences with no evident
582 evolutionary conservation to affect cellular phenotypes and fitness.

`



583

584 **Figure 7: Translation inhibition of transient ORFs causes strong phenotypes. A)** A two-step strategy for
585 inhibiting nORF translation. An ORF may overlap a DNA regulatory element or an RNA with a noncoding
586 function (Wildtype ORF), both of which are disrupted in a gene replacement strategy in addition to the
587 loss of translation (Full gene replacement). This creates ambiguity in interpreting comparisons between
588 deletion mutants and wildtype strains. Following a deletion screen using gene replacement, we used a
589 second round of homologous recombination to restore either the full ORF (Restored ORF) or an ORF

`

590 with its start codon mutated from ATG to AAG (Translation inhibition). As these mutants differ only by
591 this single base, the specific effects of translation inhibition can be inferred. **B)** Inhibiting translation of
592 transient ORFs triggers colony growth phenotypes. The fitness of AAG mutants (translation inhibition) is
593 shown for seven transient ORFs under stress conditions (colors). Fitness is assessed by comparing colony
594 size between AAG mutants and ATG controls (restored ORFs). A cross symbol after the ORF names
595 indicates unannotated nORFs assigned systematic names in this study. Relative fitness values
596 significantly different from one are indicated as follows: *p<.05 **p<.01 ***p<.001. **C-E)** Deleterious
597 impact of inhibiting translation of transient nORF YDL204W-A in a liquid growth assay. Liquid growth
598 curve of a strain in which YDL204W-A translation is inhibited by mutating its start codon (AAG) and a
599 strain with the initial codon as ATG in: 1M NaCl (C), 100mM hydroxyurea (D), and YPDA (E), with three
600 technical replicates for each strain. **F-H**) Expression from plasmid restores wildtype growth to YDL204W-
601 A start codon mutants. Liquid growth curves of an attempted rescue of the YDL204W-A AAG mutant by
602 expressing intact YDL204W-A from a plasmid. The AAG start codon mutants were transformed with
603 either an empty plasmid or a plasmid expressing the intact ORF; the ATG controls were transformed
604 with an empty plasmid. All strains were then assayed for growth in liquid media in either 1M NaCl (F),
605 100 mM hydroxyurea (G) or YPDA (H) with three technical replicates each. The shaded area covers 1 SD
606 from the mean OD value among replicates. **I)** YDL204W-A is translated and not conserved. Top:
607 ribosome profiling reads mapped by iRibo to YDL204W-A show triplet periodicity. Bottom: alignment of
608 the YDL204W-A ORF against homologous DNA in the *Saccharomyces* genus.

609

610 **Discussion**

611 Since the advent of ribosome profiling, it has been evident that large parts of eukaryotic genomes are
612 translated outside of canonical protein-coding genes[1], but the nature and full significance of this
613 translation has remained elusive. To facilitate study of this noncanonical translatome, we developed
614 iRibo, a framework for integrating ribosome profiling data to sensitively detect ORF translation across a
615 variety of environmental conditions. The iRibo framework can be applied to any species and set of
616 candidate ORFs of interest. Here, we deployed iRibo to map a high confidence yeast reference
617 translatome almost five times larger than the canonical translatome. This resource can serve as the basis
618 for further investigations into the yeast noncanonical translatome, including the prioritization of nORFs
619 for experimental study.

620 We designed iRibo to be highly sensitive at detecting patterns of triplet periodicity through the genome,
621 but there are some limitations to our strategy. We focused exclusively on ORFs with AUG start codons
622 and therefore missed the non-AUG codons that are sometimes used as starts.[70] Similarly, we did not
623 consider ORFs overlapping canonical genes in a different frame on the same strand, though some such
624 nORFs are known to be translated.[71,72] Finally, candidate ORFs were selected as the longest ORF in any
625 reading frame, which means the true boundaries of identified ORFs could be shorter than described. We

`

626 expect these limitations to cause underestimation of the number of translated nORFs, suggesting that
627 the true count is even larger than identified here.

628 We used the iRibo yeast reference translatome to address a fundamental question: to what extent does
629 the noncanonical translatome consist of conserved coding sequences that were missed in prior
630 annotation attempts? In a thorough evolutionary investigation, we identified 14 translated nORFs that
631 show evidence of being conserved under purifying selection. Only one of these ORFs, YJR107C-A,
632 appears to have been previously described[34], though it was not annotated on Saccharomyces Genome
633 Database at the time of our analysis. Thus, even a genome as well-studied as *S. cerevisiae*'s contains
634 undiscovered conserved genes, likely missed in prior analyses due to difficulties in analyzing ORFs of
635 short length. These 14 nORFs are, however, the exception: the great majority of translated nORF show
636 no signatures of selection, comprising a large pool of evolutionarily transient translated sequences.

637 The yeast genome thus encodes two translatomes, one conserved, one transient. The conserved
638 translatome consists of coding sequences that are preserved by strong purifying selection and usually
639 have a long evolutionary history. They tend to be relatively long, well expressed, and with sequence
640 properties highly distinct from noncoding sequences. The transient translatome, by contrast, is
641 evolutionarily young, of recent *de novo* origin from previously noncoding sequence and still similar to
642 noncoding sequences in nucleotide composition. Evolving in the absence of strong purifying selection,
643 transient translated ORFs appear to be frequently lost to disrupting mutations, only to be replaced by
644 other transient translated ORFs upon translation-enabling mutations. Despite these profound
645 differences, transient translated ORFs, like conserved ones, can affect the phenotype and fitness of the
646 organism. Several well-characterized coding sequences unique to *S. cerevisiae*, such as *HUR1*[66] and
647 *MDF1*[62], play key roles in biological processes through encoding lineage-specific proteins that physically
648 interact with conserved proteins. Additionally, around 100 transient ORFs are annotated as coding genes
649 and have therefore been extensively screened; a majority express stable proteins and many have known
650 loss-of-function phenotypes. Their genetic interaction patterns suggest involvement in a wide array of
651 specialized cellular processes. Our experiments revealed that disabling the start codons of unannotated
652 transient translated ORFs can cause large fitness reductions in stress conditions. The strength of the
653 fitness reduction observed was highly dependent on the stressor applied in the environment, suggesting
654 again specialized cellular roles. In some cases, disabling the start codon resulted in growth increases,
655 perhaps indicating that disabling translation saved the cell energy.

`

656 Our work adds to the growing research on the roles noncanonical coding play across many species,
657 including humans.[7,73] We note that "noncanonical" is not a coherent biological category, as it simply
658 indicates the class of sequences that have not been annotated in genome databases. We demonstrate
659 that the division between "canonical" and noncanonical" translation in *S. cerevisiae* corresponds largely,
660 but not perfectly, to a biological division between transient and conserved. It is this biological division
661 that is fundamental: the 101 yeast canonical ORFs classified as transient have sequence and
662 evolutionary properties nearly identical to noncanonical transient ORFs, except for sequence length, and
663 should be placed in the same category. We can thus reclassify the translatome according to biology
664 rather than annotation history.

665 It is perhaps surprising that a coding sequence can affect organism phenotype despite showing no
666 evidence of selection. However, this result is consistent with evidence from the field of *de novo* gene
667 birth. Species-specific coding sequences have been characterized in numerous species.[32] For example,
668 Xie et al. 2019[74] identified a mouse protein contributing to reproductive success that experienced no
669 evident period of adaptive evolution. Sequences that contribute to phenotype without conservation
670 have also been described outside of coding sequences. Regulatory sequences, such as transcription
671 factor binding sites, are a mix of relatively well-conserved elements and elements that are not preserved
672 even between close species[75]; it is plausible that translated sequences also show such a division. There
673 are several explanations for why translated ORFs may lack detectable signatures of selection. Most
674 transient ORFs are expressed at much lower levels than canonical genes, and therefore may have
675 minimal effects on phenotype. For those that do have large and beneficial effects in some
676 environmental conditions, these may be balanced by deleterious effects in other conditions. Moreover,
677 selection may occur, and be biologically important, below the limits of detectability for the genomic
678 approaches we used. Our findings do not imply an absence of selective forces in shaping the patterns of
679 noncanonical translation. Rather, the particular selective environment favoring expression of these
680 sequences may be too short-lived to detect selection using traditional comparative genomics
681 approaches. Previous research, such as the proto-gene model of *de novo* gene birth[3], have proposed
682 that recently emerged translated ORFs serve as an intermediary between noncoding sequences and
683 mature genes. Our results add to the evidence that these ORFs provide many potential phenotypes from
684 which selection could preserve beneficial ones for the long term.[65] Still, the observation that even ORFs
685 with phenotypes lack evidence of conservation at the population level suggests that there are important
686 filters that prevent the vast majority of recently emerged translated ORFs, even those with beneficial
687 phenotypes, from evolving into mature genes that are preserved over long evolutionary time. The

`

688 primary significance of the great majority of transient translated ORFs is in their biological activity over
689 their short lifespans.

690 The yeast reference translatome resource we constructed with iRibo is meant to facilitate community
691 efforts to decipher the specific physiological implications of transient translated ORFs. Our proof-of
692 concept analyses of subcellular localization, genetic interactions and ATG->AAG mutants suggest
693 involvement in diverse cellular processes and pathways. It is important to note that some transient
694 translatome phenotypes may be mediated by a protein product, by the process of translation itself, or
695 both. Translation of both uORFs[76] and dORFs[77] can affect expression of nearby genes. Translation also
696 plays a major role in the regulation of RNA metabolism through the nonsense-mediated decay pathway.
697 [78,79] Dissection of the molecular mechanisms mediating transient translatome phenotypes is an exciting
698 area for future research.

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

707 **Acknowledgments**

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

716 **Author contributions**

`

717 Conceptualization, A.W. and A.-R.C. Methodology, A.W., A.-R.C., S.B.P., N.C.C., O.A. Investigation, A.W.,
718 N.C.C., S.B.P., O.A., C.H., L.C. Writing – Original Draft, A.W., S.B.P., O.A., N.C.C. Writing – Review &
719 Editing, A.W., A.-R.C., S.B.P., N.C.C., O.A., C.H., L.C. Supervision, A.-R.C.

720 **Declaration of interests**

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

722 **Tables**

723 **Table 1: Properties of well-conserved nORFs.** Systematic name refers to either the systematic name
724 annotated on SGD, or the name assigned here according to SGD conventions. BLASTP and TBLASTN e-
725 values are the minimum BLASTP or TBLASTN e-value observed in a search of the ORF against the yeast
726 genomes assembled by Shen et al.[41], excluding those in the *Saccharomyces* genus. BLAST coverage is the
727 length of the segment that aligns to the best identified homolog (lowest e-value) in the BLAST search.
728 RFC is reading frame conservation of the ORF among species in the Saccharomyces genus. Length is the
729 length of the ORF in nucleotides. The pN/pS ratio is obtained from nucleotide variation in the ORF
730 among the 1011 S. cerevisiae strains assembled by Peter et al.[40]; significant values below 1 indicate
731 purifying selection. The dN/dS ratio was obtained from nucleotide variation in the ORF among
732 Saccharomyces genus species; significant values below 1 indicate purifying selection. Translation
733 percentile indicates the percentage of nORFs with a lower ribo-seq read count per codon than the
734 indicated ORF.

Systematic Name	Coordinates	BLASTP e-value	TBLASTN e-value	RFC	BLAST coverage (nt)	Length (nt)	pN/pS (p-value)	dN/dS (p-value)	Translation percentile
YBL029W-B[a]	chrII:164192-164368	6.5×10^{-4}	8.0×10^{-3}	0.82	107	177	1.65 (.33)	0.88 (.68)	67
YBL014W-A[a]	chrII:196737-196889	4.1×10^{-5}	1.0×10^{-4}	1	116	153	0.47 (.11)	0.14 (3.46×10^{-12})	86
YBR085W-B[a]	chrII:417494-417556	1	1	0.86	0	63	0.72 (.48)	1.26 (.62)	58
YBR268W-A[a]	chrII:741844-742005	1	1	0.99	0	162	0.61 (.15)	0.35 (3.18×10^{-7})	97
YBR292W-A[a]	chrII:786745-786903	1.9×10^{-7}	5.0×10^{-3}	0.96	146	159	0.72 (.43)	0.57 (.0026)	83
YER186W-A[a]	chrV:565603-565800	6.0×10^{-6}	1	0.92	143	198	0.55 (.02)	1.0 (1)	97
YGL262W-A[a]	chrVII:4663-4872	1	1.0×10^{-3}	0.88	113	210	0.96 (.86)	1.0 (1)	86
YGR238W-A[a]	chrVII:969015-969089	1	1	0.87	0	75	0.20 (.01)	1.18 (.74)	94
YBL049C-A[a]	chrII:126330-126461	8.3×10^{-5}	6.0×10^{-4}	0.84	92	132	1.36 (.79)	1.5 (.22)	75
YBL026C-A[a]	chrII:169634-169870	6.8×10^{-12}	9.0×10^{-10}	0.88	116	237	1.30 (.6)	0.87 (.42)	99.96
YJR107C-A[a]	chrX:628457-628693	3.8×10^{-8}	3.0×10^{-18}	0.99	161	237	0.39 (.005)	1.42 (.13)	99.91

`

YLR349C-A[a]	chrXII:828276-828338	1	1	0.81	0	63	0.30 (.02)	0.73 (.24)	73
YNR062C-A[a]	chrXIV:745640-745792	5.0×10^{-14}	5.0×10^{-13}	0.89	110	153	0.65 (.44)	1.49 (.15)	44
YBR012C	chrII:259147-259566	6.51×10^{-59}	1×10^{-16}	0.70	120	420	.62 (.1)	.50 (.039)	92

[a]We assigned this unannotated ORF a systematic name based on SGD conventions.

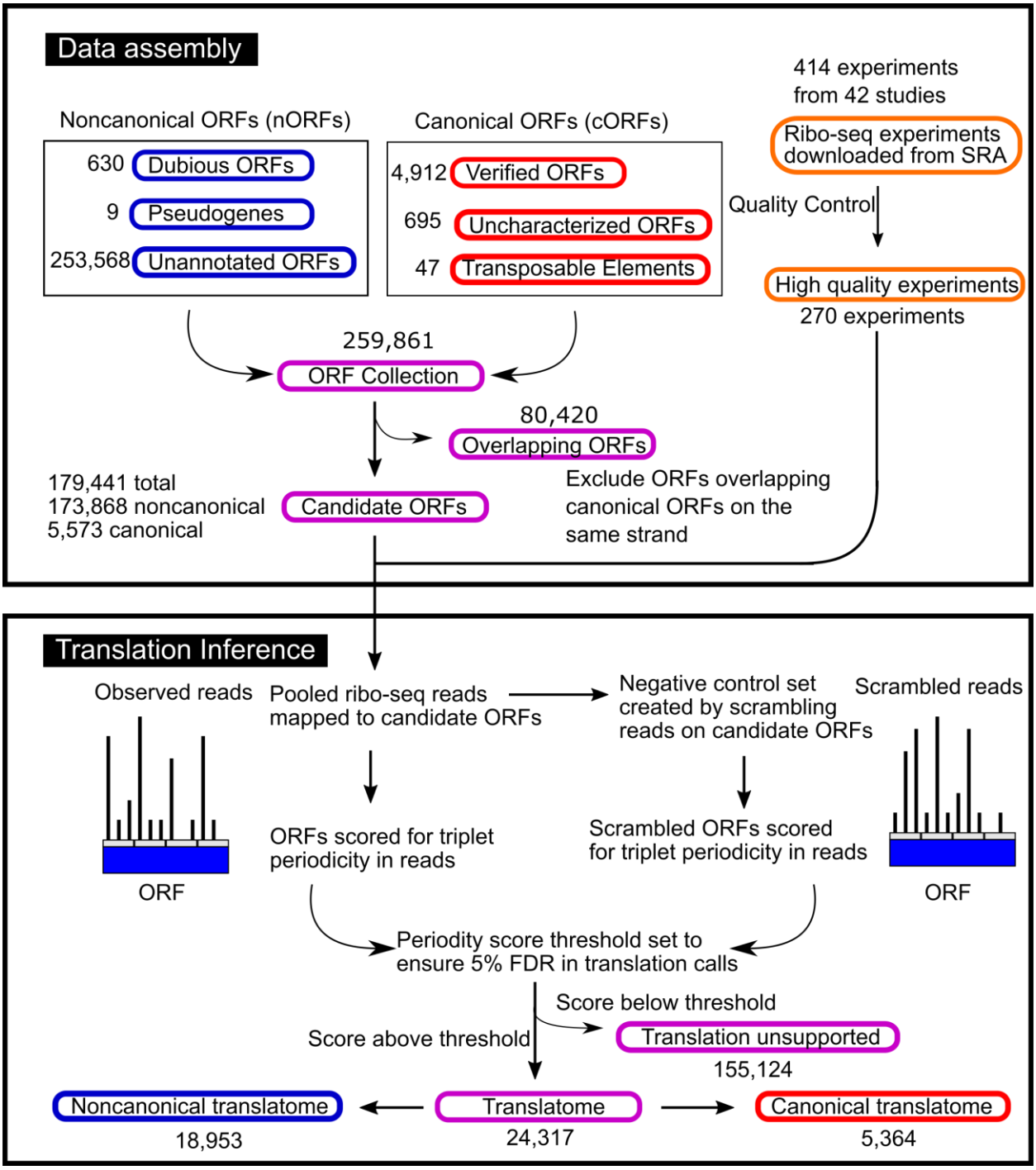
Table 2: Evolutionary properties of transient ORFs with phenotypes induced by inhibiting translation. The pN/pS ratio is obtained from nucleotide variation in the ORF among the 1011 S. cerevisiae strains assembled by Peter et al.[40] TBLASTN was run for each ORF against genomes in the subphylum *Saccharomycotina*, excluding the genus *Saccharomyces*, with an e-value threshold of 10^{-4}.

ORF Name	Reading frame conservation	pN/pS (p-value)	TBLASTN matches
YBR196C-A	.29	1.34 (0.65)	0
YDL204W-A[a]	.20	1.25 (0.83)	0
YGR016C-A[a]	.29	0.66 (0.36)	0
YJL077C	.21	0.74 (0.19)	0
YNL040C-A[a]	.38	0.97 (1.00)	0
YPR096C	.20	1.39 (0.47)	0

[a]We assigned this unannotated ORF a systematic name based on SGD conventions.

Supplementary figure legends

`

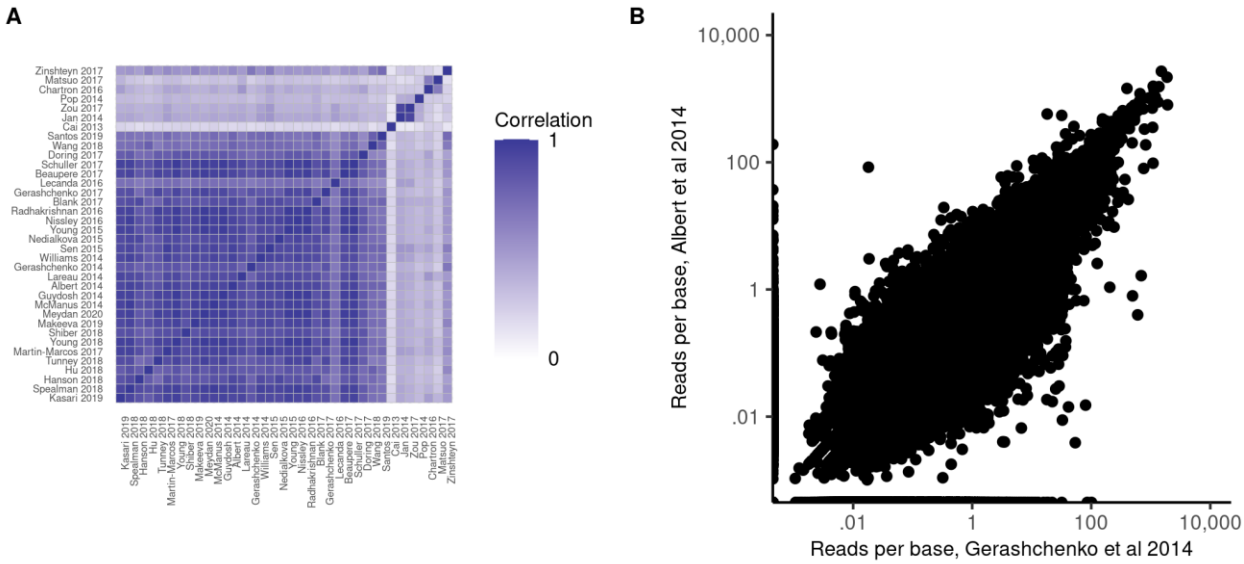


742

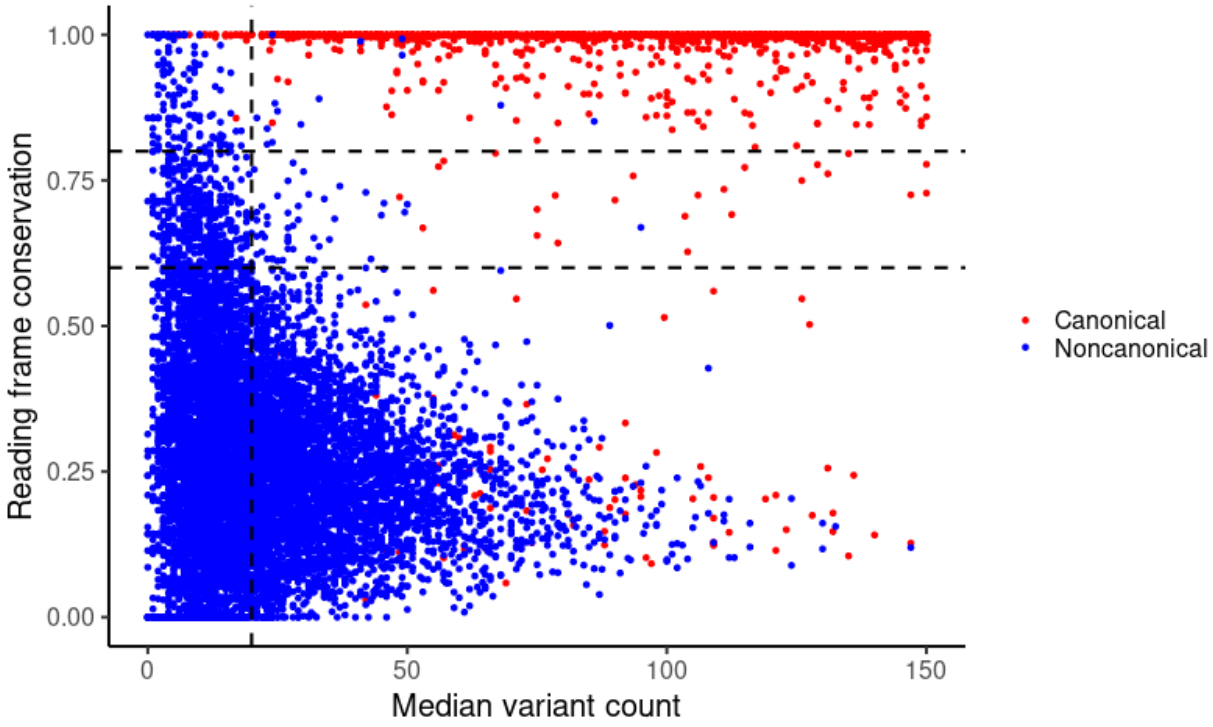
743 **Supplementary Figure 1:** Workflow to identify translated ORFs in the *S. cerevisiae* genome using

744 published datasets.

`



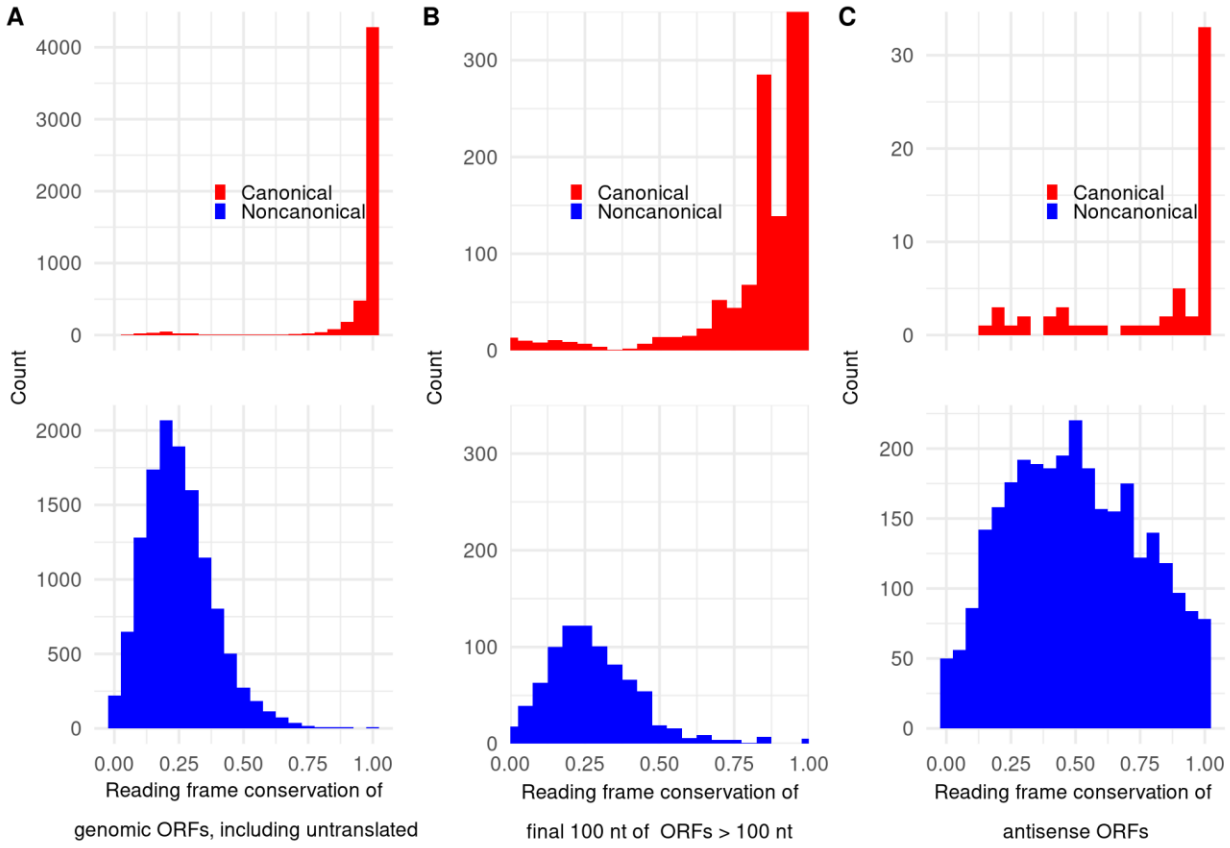
Supplementary Figure 2: Translation patterns in candidate ORFs show high replicability between studies. A) Pairwise correlation between ribo-seq coverage of all candidate ORFs between studies included in the dataset. B) For each candidate ORF, the reads per base (considering only in-frame reads) are plotted for the two largest studies in the dataset.



Supplementary Figure 3: Nucleotide variation determines ability to distinguish conserved ORFs. Reading frame conservation for each nonoverlapping ORF is plotted against the median count of

`

753 differences between the *S. cerevisiae* ORF and the aligned homologous sequence in each *Saccharomyces*
754 relative. Colors indicate SGD annotation categories. To the right of the vertical line, there are two
755 distinct populations separable by reading frame conservation; the intermediate region contains few
756 ORFs. For ORFs to the left of the vertical line, with few differences in the ORF between species, there is
757 no clear separation between high-RFC and low-RFC ORFs.

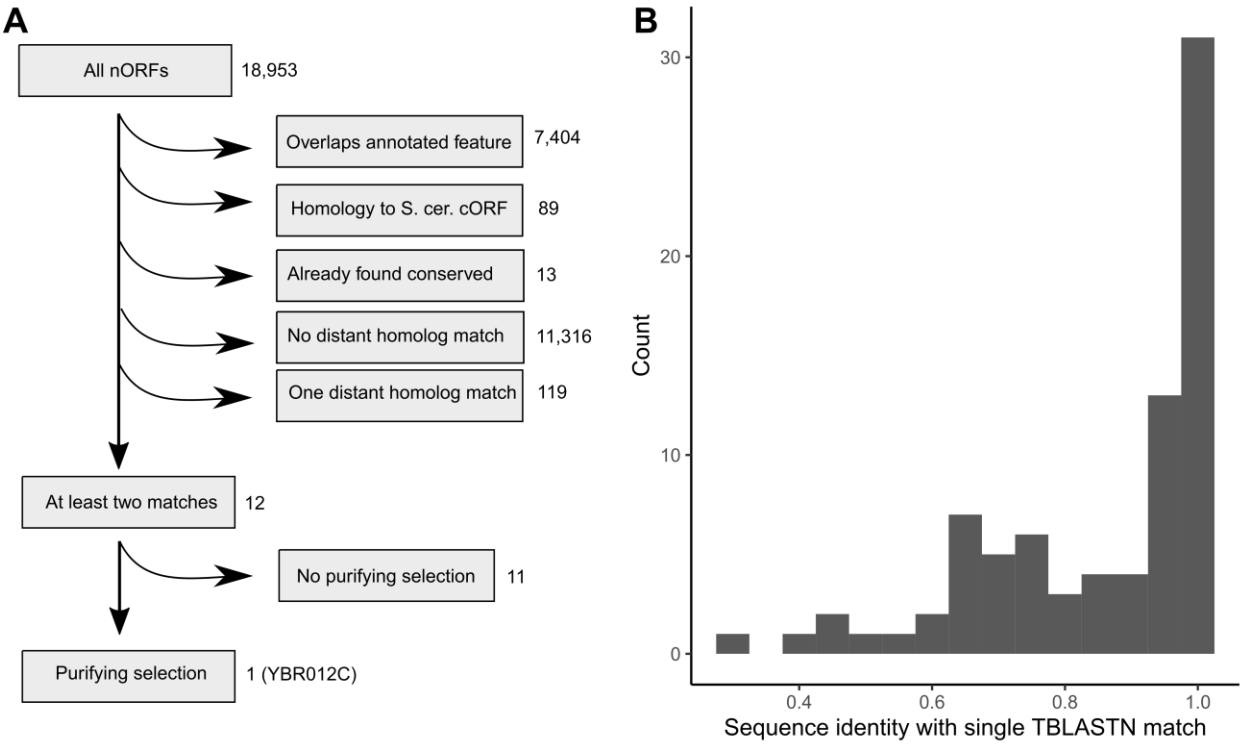


758

759 **Supplementary Figure 4: Distribution of frame conservation among classes of ORFs**. A) The distribution
760 of frame conservation among candidate ORFs in the genome, including both translated and untranslated
761 ORFs. B) For all ORFs in the high information set at least 100 nt in length, RFC was calculated considering
762 only the final 100 nt of the ORF. RFC was then plotted for both cORFs and nORFs. This was done to test
763 whether low RFC in nORFs could be caused by inferring start codons upstream of the actual start codons
764 for conserved nORF, which would lead to false inference of a low RFC value. However, the pattern
765 considering only the final 100 nt is similar to the pattern observed for the full ORFs in Figure 4B, with a
766 clear bimodal distribution, indicating that false start codon inference is likely not driving the pattern. C)
767 The distribution of frame conservation is plotted for translated cORFs and nORFs that are antisense to

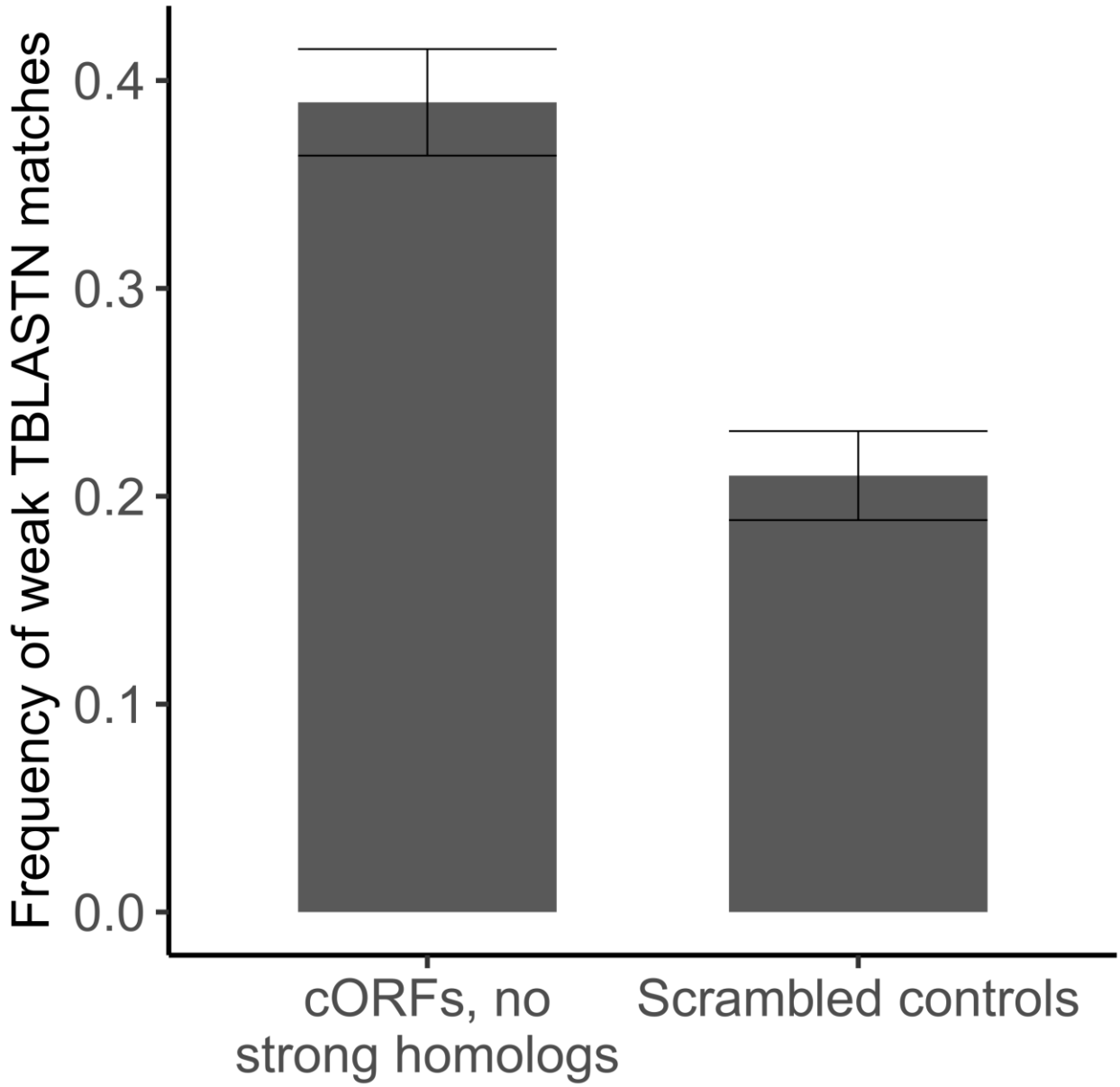
`

canonical genes. In contrast to frame conservation among nonoverlapping ORFs, the distribution does not appear bimodal.



Supplementary Figure 5: Identification of conserved genes in the noncanonical translatome using TBLASTN. A) Process for identification of conserved nORFs evolving under purifying selection. To be identified as conserved, an nORF could not overlap any annotated feature on the *S. cerevisiae* genome or have any homology to an *S. cerevisiae* cORF at a 10^{-4} BLASTP e-value threshold (as this makes BLAST results ambiguous) and have at least two identified homologs in a TBLASTN search at a 10^{-4} e-value threshold. Then, an additional indicator of selection was required (RFC > .8, or p-value < .05 in a test of neutrality using dN/dS or pN/pS). B) Among translated *S. cerevisiae* ORFs with a single TBLASTN hit (at a 10^{-4} e-value threshold) among budding yeasts outside the *Saccharomyces* genus, the distribution of sequence identities with that match is plotted. The existence of only a single match together with the prevalence of high sequence identities (>80%) suggests that the matches may be the result of genomic contamination rather than genuine homology, so at least two matches are required to accept homology as valid.

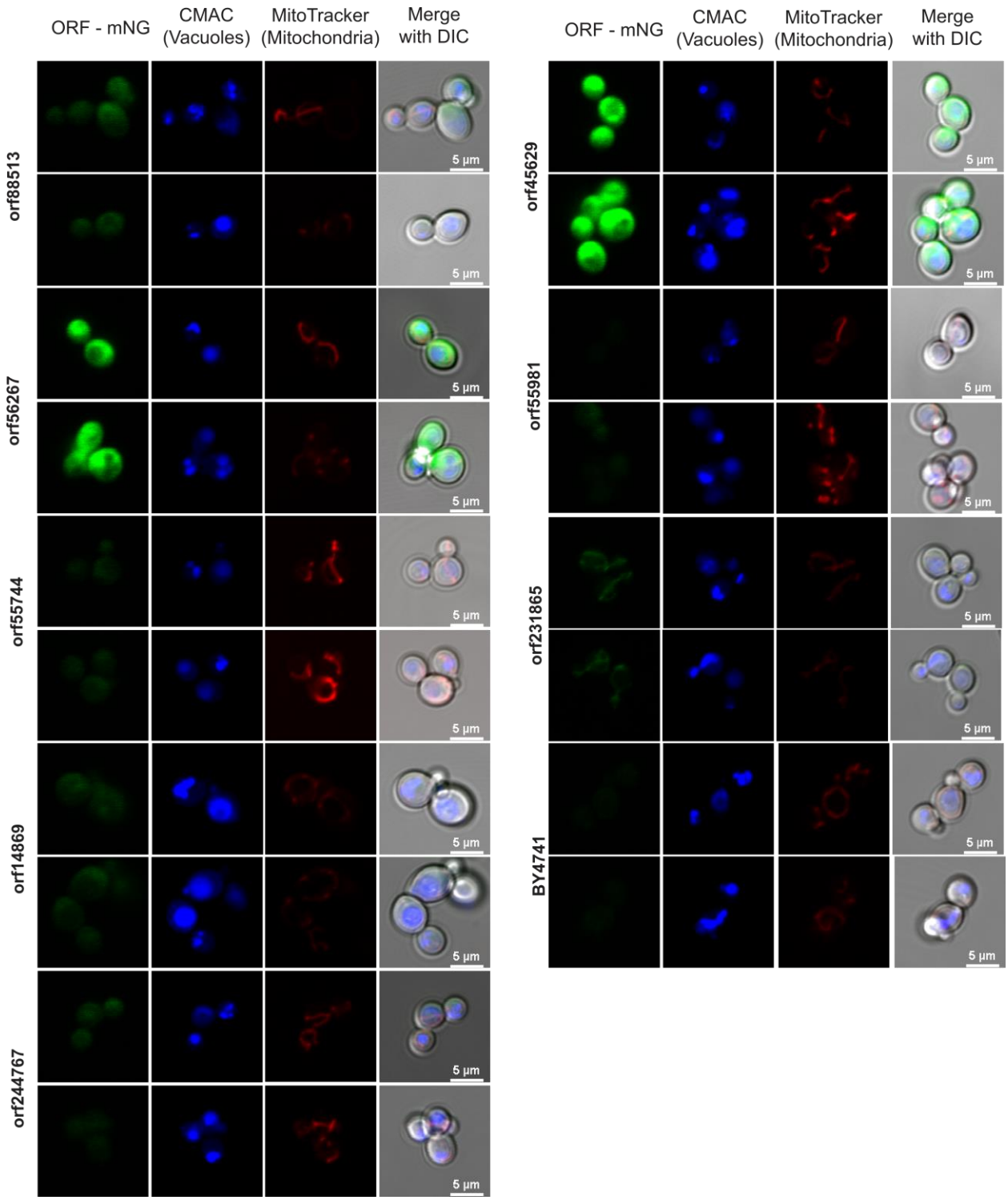
`



785

786 **Supplementary Figure 6: cORFs lacking high-confidence homologs are enriched in weak TBLASTN**
787 **matches.** The frequency of weak TBLASTN matches (10^{-4} < e-value < .05) among budding yeast genomes
788 for cORFs that lack any strong matches, and controls consisting of the same sequences randomly
789 scrambled. Error bars indicate standard errors estimated from bootstrapping.

`



790

791 **Supplementary Figure 7: Microscopy of detected transient nORFs.** Microscopy images of unannotated
792 transient nORFs taken at 40X. Left panel show the expression of the nORFs tagged with mNeonGreen,
793 middle panels the dyes CMAC Blue and MitoTracker Red for mitochondria and vacuoles identification,
794 respectively, and the right panel the merge all the above channels with DIC. Two representative images

`

795 are shown per strain; expression of orf55981 was not uniformly detected, with some cells showing
796 expression and some not.

797 **Supplementary tables**

798 **Supplementary Table 1:** Ribosome profiling experiments used for translation inference.

799 **Supplementary Table 2:** Ribosome profiling studies used for translation inference.

800 **Supplementary Table 3:** The yeast translatome.

801 **Supplementary Table 4:** Selection analysis of ORF groups in S. cerevisiae strains.

802 **Supplementary Table 5:** Phenotypes of canonical evolutionarily transient ORFs reported in literature.

803 **Supplementary Table 6:** Results of deletion mutant screen on transient nORFs using a gene replacement
804 strategy.

805 **Supplementary Table 7:** Gene ontology analysis of genetic interactors of annotated transient ORFs.

806 **Supplementary Table 8:** Information on ORFs tested in translation inhibition experiment.

807 **Supplementary Table 9:** Strains used in this study.

808

`

809 **STAR Methods**

810 **Key resources table**

Reagent or Resource	Source	Identifier
Deposited Data		
Deletion screen colony growth images	This paper	https://figshare.com/articles/dataset/A_vast_evolutionarily_transient_translatome_contributes_to_phenotype_and_fitness_-_Deletion_screen_data/21741434
C-SWAT collection	Meurer et al.[60]	Supplementary Table
YeastRGB collection	Dubreuil et al.[61]	Yeastrgb.org
CYCLoPs collection	Ko et al.[59]	https://thecellvision.org/cyclops/
Saccharomyces cerevisiae S288C reference genome	Saccharomyces genome database	S288C reference sequence R64.2.1
S. paradoxus genome	Liti et al. 2009[80]	http://www.saccharomycessensustricto.org/
S. arboricolus genome	Liti et al. 2013[81]	GCF_000292725.1
S. jurei genome	Naseeb et al. 2018[82]	GCA_900290405.1
S. mikatae, *S. bayanus var. uvarum*, *S. bayanus var. bayanus,* and *S. kudriavzevii* genome	Scannell et al. 2011[51]	http://www.saccharomycessensustricto.org/
TIF-seq data	Pelechano et al. 2014[47]	GSE39128
S. cerevisiae strain genomes	Peter et al. 2018[40]	http://1002genomes.u-strasbg.fr/files/
Budding yeast genomes	Shen et al. 2018[41]	https://y1000plus.wei.wisc.edu/data
Reagents		
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
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 study	
Saccharomyces cerevisiae: BY4741, ORF-mNG:HYG (mini collection with the selected ORFs tagged with mNeonGreen)	This study	
BY4741, YDL204W-A(wt):HYG	This study	
BY4741, YDL204W-A(ATG->AAG):HYG	This study	
BY4741, YBR196C-A(wt):HYG	This study	
BY4741, YBR196C-A(ATG->AAG):HYG	This study	
BY4741, YDR073C-A(wt):HYG	This study	
BY4741, YDR073C-A(ATG->AAG):HYG	This study	

`

BY4741, YGR016C-A(wt):HYG	This study	
BY4741, YGR016C-A(ATG->AAG):HYG	This study	
BY4741, YJL077C(wt):HYG	This study	
BY4741, YJL077C(ATG->AAG):HYG	This study	
BY4741, YNL040C-A(wt):HYG	This study	
BY4741, YNL040C-A(ATG->AAG):HYG	This study	
BY4741, YPR096C(wt):HYG	This study	
BY4741, YPR096C(ATG->AAG):HYG	This study	
BY4741, YDL204W-A(wt):HYG, pAG-GPD-ccdB1-KanMx	This study	
BY4741, YDL204W-A(ATG->AAG):HYG, pAG-GPD-ccdB1-KanMx	This study	
BY4741, YDL204W-A(wt):HYG, pAG-GPD-YDL204W-A-KanMx	This study	
BY4741, YDL204W-A(ATG->AAG):HYG, pAG-GPD-YDL204W-A-KanMx	This study	
Plasmids		
pAG-GPD-ccdB1-KanMx	This study	
pAG-GPD-YDL204W-A-KanMx	This study	
Software and algorithms		
Code for analyses conducted	This paper	https://zenodo.org/badge/latestdoi/446910374
R	R	R version 4.1.2

`

BLAST	National Library of Medicine	BLAST 2.9.0+
Ontologizer 2.0	Bauer et al. 2008[83]	http://ontologizer.de/
water	EMBOSS	https://www.ebi.ac.uk/Tools/psa/emboss_water/
MUSCLE 3.8.31	Edgar 2004[84]	https://www.drive5.com/muscle/

Resources 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.

Data and code availability

- All original code has been deposited on GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Plate images for colony growth assays are available at Figshare and are publicly available as of the data 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 Subject Details

Yeast strains

All strains used in this study are derived from BY4741 (Dharmacon, YSC1048). The parental strain and all derivatives produced in this study are listed in Supplementary Table 9. The lithium acetate method[85] 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-

`

836 exon reading frames starting with an ATG, ending with a canonical stop codon, and having at least one
837 additional codon between the start and stop. Among all ORFs that shared a stop codon, all but the
838 longest were discarded. An ORF was considered canonical if it shared a stop codon with an ORF
839 annotated as "verified", "uncharacterized", or "transposable element gene" on SGD. All other ORFs that
840 overlapped a canonical ORF on the same strand were removed (including pairs of overlapping canonical
841 genes) while ORFs that overlapped cORFs on the opposite strand were classified as antisense ORFs.

842 **Yeast ribo-seq dataset collection and read mapping**

843 A list of *S. cerevisiae* ribosome profiling (ribo-seq) studies was identified by conducting a broad literature
844 search. For each study, all ribo-seq experiments were added to the dataset except those conducted on
845 mutants designed to alter wildtype translation patterns. The full list of experiments and studies included
846 is given in **Supplementary Tables 1 and 2**, respectively. The fastq files associated with each experiment
847 were downloaded from Sequence Read Archive[86] or the European Nucleotide Archive.[87] If adaptors were
848 present in the fastq file, they were trimmed. Reads were filtered to exclude reads in which any base had
849 a Phred score below 20. For each remaining read, the number of perfect matches in the *S. cerevisiae*
850 genome were identified, and only unique perfect matches were kept.

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

`

867 For quality control, presence of triplet periodicity was then tested for each read length in each
868 experiment. The number of reads mapping (after remapping) to the first, second, and third position of
869 each codon was counted among annotated genes, requiring at least twice as many reads in the first
870 position than each of the second and third. If a read length failed this test for a given experiment it was
871 excluded from further analysis, and if all read lengths for an experiment failed the experiment itself was
872 excluded. All read lengths from 25 to 35 nucleotides were tested.

873 **Translation calling**

874 The iRibo program can be applied to any set of ribo-seq experiments to identify a set of ORFs with
875 evidence of translation among those experiments. To construct a reference translatome, translation was
876 inferred using ribo-seq data from the full set of experiments we collected that passed quality control
877 (**Supplementary Table 3**). Separately, iRibo was also run on specific subsets of the full collection,
878 including: experiments with or without the drug cycloheximide, experiments only on cells grown in YPD;
879 only on cells grown on SD; and only on cells grown in YPD without cycloheximide (**Supplementary Table**
880 **3**). iRibo was also run separately for each individual study, generating lists of translated ORFs within
881 each study.

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

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

`

896 to give a 5% FDR among noncanonical ORFs; all ORFs with p-values below this threshold were then
897 included in the translated set, whether canonical or noncanonical.

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

899 All nORFs were partitioned into genomic contexts, with nonoverlapping nORFs classified by the relation
900 between the nORF and any cORF located on the same transcript and antisense nORFs classified by
901 partial or complete overlap of the opposite strand gene. The transcripts reported in Pelechano et al.
902 2014[47] based on TIF-seq data were used for this analysis. An nORF was considered antisense if it
903 overlapped an ORF annotated as "verified", "uncharacterized", "transposable element" or "blocked" on
904 SGD on the opposite strand and nonoverlapping otherwise (ORFs overlapping annotated genes on the
905 same strand were excluded from analysis, as described above). A nonoverlapping nORF was considered
906 to share a transcript with a cORF or annotated non-coding RNA if any transcript fully contained both the
907 nORF and the cORF or annotated RNA sequence; the ORF was then further classified as being either a
908 uORF or dORF based on whether it was upstream or downstream of the cORF or RNA. If an nORF shared
909 a transcript with both its upstream and downstream neighboring cORFs, it was classified according to
910 the cORF that was closer.

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

912 Genomes were obtained from seven relatives of *S. cerevisiae* within the *Saccharomyces* genus: *S.*
913 *paradoxus* from Liti et al. 2009[80], *S. arboricolus* from Liti et al. 2013[81], *S. jurei* from Naseeb et al. 2018[82],
914 and *S. mikatae*, *S. bayanus var. uvarum*, *S. bayanus var. bayanus,* and *S. kudriavzevii* from Scannell et al.
915 2011.[51] Alignments were constructed between each *S. cerevisiae* ORF and its homologs in each
916 *Saccharomyces* relative using synteny information. To identify anchor genes for syntenic blocks, BLASTP
917 was run for each annotated ORF in *S. cerevisiae* against each ORF in the comparison species. Identified
918 homolog pairs with e-value < 10^{-7} were selected as potential anchors. For each ORF in the *S. cerevisiae*
919 genome, the upstream anchor G_0 and downstream anchor G_1 were selected that minimized the sum of
920 the distance between the anchors in *S. cerevisiae* and the distance between the anchors in the
921 comparison species; this sum was required to be less than 60 kb. The sequence between and including
922 G_0 and G_1 were then extracted from both the *S. cerevisiae* genome and the comparison species and a
923 pairwise alignment of the syntenic region was generated using MUSCLE 3.8.31.[84]

924 To confirm that the ORF was matched to genuinely homologous DNA, the alignment of the *S. cerevisiae*
925 ORF along with its 50 bp flanking regions was extracted from the full syntenic alignment. The extracted

`

926 region was then realigned using the Smith-Waterman algorithm[88] with a match bonus of 5, a mismatch
927 penalty of 4, and a gap penalty of 4. To test homology, 1000 alignments were constructed using the
928 same score system in which the sequence of the comparison species was shuffled at random, reflecting
929 a null hypothesis that the region was not homologous. The proportion of times the alignment of the real
930 sequence scored better than the shuffled ones is a p-value indicating the strength of the null hypothesis
931 against the alternative that the region is homologous. Homology was accepted as confirmed if the p-
932 value was less than 1%, and alignments were excluded from analysis if homology was not confirmed.

933 If a syntenic alignment could not be constructed for a particular *S. cerevisiae* ORF and comparison
934 species (because homology failed or there were no appropriate anchors), BLAST was attempted as an
935 alternative method of finding the homologous DNA sequence. For these ORF sequences, BLASTn was
936 run against the genome of the comparison species. For each reciprocal best matching pair with e-value <
937 10^{-4}, the matched sequences in both species were extracted, together with a 1000 bp flanking region in
938 both ends, and aligned using MUSCLE.[84] DNA homology was then tested using Smith-Waterman
939 alignment as described above.

940 **Division of ORFs into high information and low information sets**

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

950 **Reading frame conservation**

951 Reading frame conservation is a measure of conservation of codon structure developed by Kellis et al.
952 2003[20] and used here with some modifications. Calculation of reading frame conservation was done on a
953 pairwise alignment of a genomic region containing the *S. cerevisiae* ORF (either a syntenic block
954 between conserved genes or the 1000 bp flanking region around a BLAST hit). All single-exon ORFs (ATG
955 to stop codon) in the comparison species were identified across this region. For each ORF in the

`

956 comparison species, the reading frame conservation was calculated by summing up all points in the
957 alignment where the pair of aligned bases are in the same position within the codon (i.e., both are in
958 either the first, second, or third position) and dividing by the length of the *S. cerevisiae* ORF in
959 nucleotides (including start and stop codons). Positions that align to gaps or are outside the range of the
960 *S. cerevisiae* ORF are always considered to be not in the same codon position and do not add to the
961 numerator. The ORF in the comparison species with the highest reading frame conservation is
962 considered the best match, and the reading frame conservation of the *S. cerevisiae* ORF in relation to
963 each other *Saccharomyces* species is defined as its reading frame conservation with its best match. In
964 addition to the pairwise reading frame conservation of each *S. cerevisiae* ORF in relation to its homologs
965 in all other species, an index of reading frame conservation (RFC) was defined equal to the average
966 reading frame conservation of the *S. cerevisiae* ORF against all species in the *Saccharomyces* genus for
967 which homologous DNA could be identified.

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

969 The genomes of 332 budding yeasts were taken from Shen et al. 2018.[41] We applied TBLASTN and
970 BLASTP for each *S. cerevisiae* translated ORF against each genome in this dataset (excluding the
971 *Saccharomyces* genus). Default settings were used except for setting an e-value threshold of 0.1; results
972 were then filtered by a stricter e-value threshold as described in each analysis. The BLASTP analysis was
973 run against the list of protein coding genes used in Shen et al. 2018[41] while the TBLASTN analysis was
974 run against each entire genome. In the TBLASTN analysis, scrambled sequences of each *S. cerevisiae* ORF
975 were also included as queries to serve as a negative control.

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

977 Variant call file data for 1011 *S. cerevisiae* isolates was taken from Peter et al. 2018.[40] For each ORF,
978 nucleotide diversity was estimated from the full set of isolates. Nucleotide diversity was estimated as
979 the mean number of differences per site in the ORF between any pair of isolates. To calculate dN/dS, the
980 consensus sequence among all isolates was determined. At each position in the consensus, the three
981 possible nucleotide variations were recorded as possible polymorphisms and distinguished by
982 polymorphism type (12 possible combinations of consensus and variant nucleotide) and whether they
983 would result in a synonymous or nonsynonymous difference from the consensus. If at least one isolate
984 had the polymorphism, the polymorphism was also recorded as observed. All possible and observed
985 polymorphisms were counted among all considered ORFs.

`

986 The pN/pS ratio was calculated in a similar manner to Ruiz-Orera et al. 2018[28] and could be applied to
987 either a single ORF or a group of ORFs. For each ORF under consideration, the consensus sequence
988 among all isolates was determined. At each position in the consensus, the three possible nucleotide
989 variations were recorded as possible polymorphisms and distinguished by polymorphism type (12
990 possible combinations of consensus and variant nucleotide) and whether they would result in a
991 synonymous or nonsynonymous difference from the consensus. If at least one isolate had the
992 polymorphism, the polymorphism was also recorded as observed. All possible and observed
993 polymorphisms were counted among all considered ORFs.

994 Consider a variant $X \rightarrow Y$ where X is the consensus at a site and Y is a possible variant. The probability of
995 observing variant Y at a position with consensus X, $p_{x \rightarrow Y}$ was estimated as the observed count of X→Y
996 variant sites divided by the possible count of X→Y variant sites. Under neutrality, the expected count of
997 either synonymous or nonsynonymous X→Y variant sites is then the product of $p_{x \rightarrow Y}$ and the number of
998 possible synonymous or nonsynonymous X→Y variant sites. In this manner the expected and observed
999 counts of synonymous and nonsynonymous variants were calculated. The pN/pS ratio is then estimated
1000 as:

1001
$$\omega = \frac{nonsyn_{obs}/nonsyn_{exp}}{syn_{obs}/syn_{exp}}$$

1002 Under neutrality, then, the expected count of X→Y nonsynonymous variant sites is the number of
1003 possible such variant sites times the expected probability of this variant. In this manner the expected
1004 and observed counts of all synonymous variant types were calculated. To test for deviation from
1005 neutrality, we used a chi-squared test with one degree of freedom to compare observed vs. expected
1006 counts of synonymous and nonsynonymous variants. Standard errors for the pN/pS ratio in group
1007 analyses were estimated by bootstrapping: the ORFs in the group were resampled with replacement
1008 1000 times and the pN/pS ratio was calculated each time. The standard error was then estimated as the
1009 sample standard deviation among the 1000 pN/pS ratios.

1010 The dN/dS ratio was calculated based on differences in the pairwise ORF alignments *S. cerevisiae* and its
1011 closest relative *S. paradoxus*. Each S. cerevisiae ORF was associated with an *S. paradoxus* ORF for which
1012 the pair had the highest reading frame conservation (or none if homology with *S. paradoxus* was not
1013 confirmed or the highest reading frame conservation was 0). Counts of differences were made only for
1014 codons that shared the same frame between these ORFs and with at most one nucleotide difference
1015 between the codons. For every eligible position in the *S. cerevisiae* ORF, each possible *S. paradoxus*

`

1016 difference was counted and distinguished by whether the difference was synonymous or
1017 nonsynonymous and by type (four *S. cerevisiae* nucleotides, each with three possible *S. paradoxus*
1018 differences). These observed and possible differences were then used to estimate the dN/dS ratio in the
1019 same way as described above for the pN/pS ratio.

1020 Among nORFs with high RFC, the strong conservation in *Saccharomyces* permitted calculation of dN/dS
1021 over the entire *Saccharomyces* tree, and so this was done as an additional test of selection (as reported
1022 in Table 1). For this analysis, ancestral reconstruction of the *Saccharomyces* phylogeny was conducted
1023 using PRANK[89] with parameters -showanc -showevents -once -prunetree -keep. Ancestral reconstruction
1024 included all species in which DNA homology was confirmed. Codons were only used for counting
1025 substitutions if they shared frame conservation among all species. Observed and possible substitutions
1026 were counted across each branch and distinguished by substitution type and whether the substitutions
1027 were synonymous or nonsynonymous. Then, dN/dS was estimated in the same way as described for
1028 pN/pS above.

1029 **Classification of ORFs into transient and conserved sets**

1030 All high-information nonoverlapping translated ORFs with RFC > 0.8 were classified as conserved (**Figure**
1031 **4A**). An nORF was also classified as conserved if it overlapped no annotated feature on SGD, had
1032 TBLASTN matches with e-value < 10^{-4} with at least two species outside the *Saccharomyces* genus and
1033 showed at least one additional signature of purifying selection (RFC > 0.8 or a p-value < 0.05 in a test of
1034 neutrality using dN/dS or pN/pS) (**Supplementary Figure 5A**).

1035 Nonoverlapping ORFs were excluded from classification in the transient set if they showed homology to
1036 an ORF classified as conserved in *S. cerevisiae* (e-value < 10^{-4} using BLASTP) or to any sequence among
1037 budding yeasts outside *Saccharomyces*[41] (e-value < 10^{-4} using TBLASTN). Among remaining translated
1038 ORFs, all high-information ORFs with RFC < 0.6 were classified as transient. Low information ORFs were
1039 divided into groups and classified as transient if no group they belonged to showed evidence of
1040 selection in dN/dS analysis, pN/pS analysis, or weak homology matching analysis. Two low-information
1041 groups were cORFs and antisense nORFs. Low information nonoverlapping nORFs were each assigned to
1042 three groups corresponding to deciles of translation rate, coding score and ORF length. Analyses of
1043 dN/dS and pN/pS are described above. For weak homology detection, the number of ORFs with at least
1044 two weak TBLASTN matches (e-value < 0.05) to budding yeast genomes collected by Shen et al. 2018[41]
1045 (excluding *Saccharomyces* species) was counted for both actual and scrambled ORF sequences. Selection

`

1046 was inferred if actual matches significantly ($p < 0.05$) exceeded scrambled matches using Fisher's exact
1047 test. Only ORFs that did not overlap any annotated feature on SGD were included in weak homology
1048 detection analysis.

1049 **Coding score calculation**

1050 The coding score, described by Ruiz-Orera et al. 2014[90], is a measure of how close the hexamer (i.e., the
1051 nucleotide sequence of a pair of adjacent codons) frequency of an ORF is to the hexamer of coding vs.
1052 noncoding sequences. Higher scores indicate a more gene-like hexamer distribution. Coding hexamer
1053 frequencies were calculated among all ORFs annotated as "verified" or "uncharacterized" by
1054 Saccharomyces Genome Database.[43] Noncoding hexamer frequencies were calculated for all intergenic
1055 sequences (sequences in between verified or uncharacterized ORFs) in the *S. cerevisiae* genome. As
1056 intergenic sequence has no codon structure, hexamer frequencies for intergenic sequence were counted
1057 as if read in each possible coding frame. The score was then calculated as described in Ruiz-Orera et al.
1058 2014.[90]

1059 **Identification to transient ORFs with detectable translation products in published microscopy studies**

1060 Published results were examined from fluorescent tagging experiments where the expression of ORFs
1061 was driven by native promoters and terminators. A list of ORFs detected in 15 GFP-tagged screens on
1062 wildtype strains in either normal conditions or with chemical treatment (hydroxyurea or rapamycin)
1063 were retrieved from the CYCLoPs database.[58,59] Lists of ORFs detected in the C-SWAT tagging library
1064 were taken from Meurer et al. 2018[60] and from YeastRGB[61]. ORFs with fluorescent intensity below the
1065 reported detection threshold in each screen were filtered out. Transient ORFs that showed detectable
1066 translation products in at least one screen were considered as detected.

1067 **Literature analysis of transient translatome cORFs**

1068 For each transient translatome cORF, we examined all publications listed on SGD as "primary" or
1069 "additional" literature for the ORF. If the ORF had a phenotype in any listed publication, we noted the
1070 evidence for the phenotype (**Supplementary Table 5**).

1071 **Genetic interaction analysis**

1072 Single mutant fitness and genetic interaction data were downloaded from TheCellMap.org.[91] In this
1073 dataset, mutants of nonessential genes are full deletions and mutants of essential genes are
1074 temperature-sensitive alleles. Transient ORFs were all nonessential. Different temperature-sensitive

`

1075 alleles for the same essential gene were treated separately. We removed all genes or transient ORFs
1076 with a genomic overlap to another genetic element from our analyses as it is not possible to assign the
1077 observed phenotypes to either of the overlapping pairs.

1078 We counted the number of transient ORF and nonessential genes that showed at least one genetic
1079 interaction with Ɛ<-.2 and p-value < 0.05 (a negative genetic interaction) or Ɛ<-.35 with a p-value<0.05 (a
1080 synthetic lethal interaction). We then divided this number by the total number of transient ORFs or
1081 nonessential genes in the Costanzo et al. 2016[69] genetic interaction dataset to calculate the percentage
1082 showing at least one genetic interaction. We used Fisher's exact test to assess the significance of
1083 differences between percentages of nonessential genes and transient ORFs.

1084 Gene ontology analysis of the interactors of each ORF was conducted with Ontologizer[83], using
1085 Benjamini-Hochberg multiple testing correction and the term-for-term calculation method. The gene
1086 association file was downloaded from SGD. Gene ontology evidence codes relating to genetic
1087 interactions (IGI and HGI) were not used.

1088 **Creation of yeast strains**

1089 Deletion mutant strains for 49 transient nORFs and 3 transient cORFs were created by using homologous
1090 recombination to replace the ORFs with a KanMX cassette. Transformations were done using the
1091 LiAc/PEG protocol[85] in the background BY4741 strain, and selected in media containing G-418. After an
1092 initial screen of these strains, a subset of the deletion strains that showed strong deleterious effects
1093 were transformed a second time, also using the LiAc/PEG protocol[85], to replace the KanMx cassette with
1094 either an intact copy of the original ORF, or a mutant copy of the ORF with the start codon ATG and (in
1095 some cases) additional in-frame ATG codons mutated to AAG to prevent translation. This was
1096 accomplished by using homologous recombination to replace the KanMx cassette with a construct
1097 containing the intact or mutant ORF followed by a hygromycin resistance cassette. These constructs
1098 were synthesized by IDT (Integrated DNA Technologies). The resulting transformants were selected in
1099 agar plates containing hygromycin. All positive clones were sequenced to confirm presence of either the
1100 restored wildtype ORF or the ORF with a mutated start codon.

1101 Strains containing an mNeonGreen tag for microscopy purposes were also made by homologous
1102 recombination using the LiAc/PEG protocol[85] in the BY4741 background. The mNeonGreen and
1103 hygromycin cassette sequences were amplified from a plasmid using primers containing homology to
1104 the 3' of each ORF. The primers were designed to remove the STOP codon of each ORF and place the

`

1105 mNeonGreen in frame with the ORF, to be expressed under its native promoter. Positive clones were
1106 selected on agar plates containing hygromycin.

1107 All strains were kept in glycerol stocks at −80 °C in 96 and 384-well format until used for screening.
1108 Strain genotypes are listed in **Supplementary Table 9**.

1109 **Screening strategy for fitness estimation**

1110 Both rounds of deletion screening were conducted at 1536 colony density, with 1 in 4 colonies on the
1111 plate being reference strains used to correct for spatial biases as described in Parikh et al. 2021.[92] In the
1112 initial deletion screen, each mutant strain was tested using 12 replicates; 72 replicates were tested per
1113 strain in the start codon mutant screen. Conditions tested were YPDA and YPDA+DMSO as unstressed
1114 conditions and five stress conditions: YPDA supplemented with 1M NaCl, 100mM Hydroxyurea, 0.6µM
1115 Tunicamycin, 25µg/ml Fluconazole, or 30mM Hydrogen peroxide (H_2O_2). Agar plates were incubated and
1116 imaged periodically until the colonies reached saturation. The plate handler Singer ROTOR (Singer
1117 Instrument Co. Ltd) was used to prepare all plates starting from glycerol stocks. Serial imaging of the
1118 plates was conducted using the spImager Automated Imaging System (S & P Robotics Inc., Ontario,
1119 Canada). The images were analyzed in bulk using a custom script made using functions from the
1120 MATLAB Colony Analyzer Toolkit[92] to provide colony size estimations
1121 (https://github.com/sauriiiin/lid_personal/blob/master/justanalyze.m). The output files containing
1122 colony size information along with the images is available at https://bit.ly/3xtzHJO. The LI Detector
1123 analytical pipeline[92] was used to correct for spatial biases in colony size and obtain colony fitness
1124 estimates. Strain fitness was estimated as the median of bias-corrected colony size among replicates of
1125 the strain at 40 hours in the initial screen and 90 hours in the start codon mutant screen. In the LI
1126 Detector pipeline[92], sets of reference colonies are treated as if they were replicates of a mutant strain,
1127 with their median fitness calculated in order to construct an empirical null distribution of median fitness
1128 values to compare with estimated strain fitness. Strains were called as beneficial or deleterious using a
1129 5% false discovery rate threshold based on this empirical null distribution. For any selected fitness
1130 threshold used to infer deleterious strains, the false discovery rate can be calculated as the proportion
1131 of null distribution fitness values below that threshold divided by the proportion of mutant strain fitness
1132 values below the threshold. Thus, fitness thresholds were selected such that a 5% FDR was obtained and
1133 strains with fitness below that threshold were inferred to be deleterious. In the same manner, a list of
1134 beneficial strains at 5% FDR was also selected.

`

1135 **Liquid growth assay**

1136 For liquid growth assays, cells were first grown in liquid YPDA media overnight at 30°C in a 96-density
1137 microplate. These were then used to inoculate a new 96-density microplate with 150µl YPDA+ stress
1138 conditions (1M NaCl, 100mM Hydroxyurea)) using the Singer ROTOR (Singer Instrument Co. Ltd). This
1139 microplate was incubated at 30°C with constant double orbital shaking for a period of 72h on microplate
1140 reader Biotek Synergy H1 (Aligent Technology Inc.). Optical density readings at 600nm (OD_{600}) were
1141 taken every 15 minutes.

1142 **Microscopy**

1143 The strains containing the ORFs tagged with mNeonGreen were imaged on a Nikon TiE2 inverted A1R
1144 confocal microscope. A first screening was done at high density in 96-well plates with a 40x water
1145 objective, to assess the success of the transformations. Plates were incubated with CellTracker Blue
1146 CMAC Dye (Invitrogen) and MitoTracker Red CMXRos Dye (Invitrogen) at least 10 min prior to imaging.
1147 Plates were then imaged in 4 channels (405, 488, 561, and DIC), and 3 fields of view were taken for each
1148 strain that contained many cells. Strains that demonstrated visibly higher signal in the green channel
1149 (488nm) compared to a non-transformed background strain were selected to examine in single dishes
1150 under a 100X oil objective to more accurately evaluate sub-cellular localization. All strains were imaged
1151 in triplicate at high density and triplicate in dishes (once without CMAC and MitoTracker and two times
1152 with the dyes).

1153 Quantification and statistical analysis

1154 Statistical analyses were performed in R version 4.1.2. Details for each statistical test and analysis can be
1155 found in the results section and figure legends.

1156

1157

`

1158

1159 **References**

1160 1. Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R.,
1161 and Weissman, J.S. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated
1162 Protein-Coding Genes. Cell Rep. *8*, 1365–1379. 10.1016/j.celrep.2014.07.045.

1163 2. Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome
1164 scale. Nat. Rev. Genet. *15*, 205–213. 10.1038/nrg3645.

1165 3. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux,
1166 B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and *de novo* gene birth.
1167 Nature *487*, 370–374. 10.1038/nature11184.

1168 4. Wilson, B.A., and Masel, J. (2011). Putatively Noncoding Transcripts Show Extensive Association with
1169 Ribosomes. Genome Biol. Evol. *3*, 1245–1252. 10.1093/gbe/evr099.

1170 5. Pruitt, K.D., and Maglott, D.R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic
1171 Acids Res. *29*, 137–140. 10.1093/nar/29.1.137.

1172 6. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic,
1173 S., Zimmer, R., and Dölken, L. (2018). Improved Ribo-seq enables identification of cryptic translation
1174 events. Nat. Methods *15*, 363–366. 10.1038/nmeth.4631.

1175 7. Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann,
1176 M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open
1177 reading frames. Science *367*, 1140–1146. 10.1126/science.aay0262.

1178 8. Prensner, J.R., Enache, O.M., Luria, V., Krug, K., Clauser, K.R., Dempster, J.M., Karger, A., Wang, L.,
1179 Stumbraite, K., Wang, V.M., et al. (2021). Noncanonical open reading frames encode functional
1180 proteins essential for cancer cell survival. Nat. Biotechnol., 1–8. 10.1038/s41587-020-00806-2.

1181 9. van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M.,
1182 Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The Translational Landscape of the Human
1183 Heart. Cell *178*, 242-260.e29. 10.1016/j.cell.2019.05.010.

1184 10. Laumont, C.M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados,
1185 D.P., Durette, C., Lemieux, S., Thibault, P., et al. (2016). Global proteogenomic analysis of human
1186 MHC class I-associated peptides derived from non-canonical reading frames. Nat. Commun. *7*,
1187 10238. 10.1038/ncomms10238.

1188 11. Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A.G., Khan, O.M., Brewer, J.R.,
1189 Skadow, M.H., Duizer, C., et al. (2018). The translation of non-canonical open reading frames
1190 controls mucosal immunity. Nature *564*, 434. 10.1038/s41586-018-0794-7.

1191 12. Makarewich, C.A., and Olson, E.N. (2017). Mining for Micropeptides. Trends Cell Biol. *27*, 685–696.
1192 10.1016/j.tcb.2017.04.006.

`

1193 13. Anderson, D.M., Anderson, K.M., Chang, C.-L., Makarewich, C.A., Nelson, B.R., McAnally, J.R.,
1194 Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al. (2015). A Micropeptide Encoded by a
1195 Putative Long Noncoding RNA Regulates Muscle Performance. Cell *160*, 595–606.
1196 10.1016/j.cell.2015.01.009.

1197 14. Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A.,
1198 Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are
1199 regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228–232.
1200 10.1038/nature21034.

1201 15. Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S.E., Hildenbrand, C., Rom, J.,
1202 Aulmann, S., Sinn, H.-P., Vandesompele, J., et al. (2018). The cancer-associated microprotein
1203 CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet
1204 formation. Oncogene *37*, 4750–4768. 10.1038/s41388-018-0281-5.

1205 16. Housman, G., and Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long
1206 noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochim.
1207 Biophys. Acta BBA - Gene Regul. Mech. *1859*, 31–40. 10.1016/j.bbagrm.2015.07.017.

1208 17. Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct.
1209 Mol. Biol. *14*, 103–105. 10.1038/nsmb0207-103.

1210 18. Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K.,
1211 Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of
1212 large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. *19*,
1213 208. 10.1186/s13059-018-1590-2.

1214 19. Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for
1215 selection within long noncoding RNAs. Genome Res. *17*, 556–565. 10.1101/gr.6036807.

1216 20. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison
1217 of yeast species to identify genes and regulatory elements. Nature *423*, 241. 10.1038/nature01644.

1218 21. Ward, L.D., and Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently
1219 Acquired Regulatory Functions. Science *337*, 1675–1678. 10.1126/science.1225057.

1220 22. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E.,
1221 Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome.
1222 Proc. Natl. Acad. Sci. *111*, 6131–6138. 10.1073/pnas.1318948111.

1223 23. Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates, J.R., Lockhart, D.J., and Winzeler, E.A. (2002).
1224 Parallel Identification of New Genes in Saccharomyces cerevisiae. Genome Res. *12*, 1210–1220.
1225 10.1101/gr.226802.

1226 24. Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casarégola, S.,
1227 Montigny, J. de, Gaillardin, C., Lépingle, A., et al. (2000). Genomic Exploration of the
1228 Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited. FEBS Lett. *487*,
1229 31–36. 10.1016/S0014-5793(00)02275-4.

`

1230 25. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar,
1231 C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014). Identification of small ORFs in vertebrates
1232 using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993.
1233 10.1002/embj.201488411.

1234 26. Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., and
1235 Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide
1236 search for novel putatively coding sORFs. BMC Genomics *14*, 648. 10.1186/1471-2164-14-648.

1237 27. Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A.K., Nielly-Thibault, L., Namy, O., and
1238 Landry, C.R. (2019). Turnover of ribosome-associated transcripts from de novo ORFs produces gene-
1239 like characteristics available for de novo gene emergence in wild yeast populations. Genome Res.
1240 *29*, 932–943. 10.1101/gr.239822.118.

1241 28. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J.L., Messeguer, X., and Albà, M.M. (2018).
1242 Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat. Ecol.
1243 Evol. *2*, 890. 10.1038/s41559-018-0506-6.

1244 29. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2016).
1245 sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *44*, D324–
1246 D329. 10.1093/nar/gkv1175.

1247 30. Lee, J., Wacholder, A., and Carvunis, A.-R. (2021). Evolutionary Characterization of the Short Protein
1248 SPAAR. Genes *12*, 1864. 10.3390/genes12121864.

1249 31. Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Calvet, F., Jungreis, I., Gonzalez, J.M.,
1250 Magrane, M., Martinez, T.F., Schulz, J.F., et al. (2022). Standardized annotation of translated open
1251 reading frames. Nat. Biotechnol. *40*, 994–999. 10.1038/s41587-022-01369-0.

1252 32. Van Oss, S.B., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet. *15*.
1253 10.1371/journal.pgen.1008160.

1254 33. Blevins, W.R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J.L., Espinar, L.,
1255 Díez, J., Carey, L.B., and Albà, M.M. (2021). Uncovering de novo gene birth in yeast using deep
1256 transcriptomics. Nat. Commun. *12*, 604. 10.1038/s41467-021-20911-3.

1257 34. Yagoub, D., Tay, A.P., Chen, Z., Hamey, J.J., Cai, C., Chia, S.Z., Hart-Smith, G., and Wilkins, M.R.
1258 (2015). Proteogenomic Discovery of a Small, Novel Protein in Yeast Reveals a Strategy for the
1259 Detection of Unannotated Short Open Reading Frames. J. Proteome Res. *14*, 5038–5047.
1260 10.1021/acs.jproteome.5b00734.

1261 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
1262 hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res. *47*, 8111–8125.
1263 10.1093/nar/gkz646.

1264 36. Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E.,
1265 Knisbacher, B.A., Le, P.M., et al. (2020). Thousands of novel unannotated proteins expand the MHC I
1266 immunopeptidome in cancer. bioRxiv, 2020.02.12.945840. 10.1101/2020.02.12.945840.

`

1267 37. Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., and Dieterich, C. (2017). Bayesian
1268 prediction of RNA translation from ribosome profiling. Nucleic Acids Res. *45*, 2960–2972.
1269 10.1093/nar/gkw1350.

1270 38. Ji, Z. (2018). RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome
1271 Profiling. Curr. Protoc. Mol. Biol. *124*, e67. 10.1002/cpmb.67.

1272 39. Calviello, L., and Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the
1273 Functions of the Transcriptome. Trends Genet. *33*, 728–744. 10.1016/j.tig.2017.08.003.

1274 40. Peter, J., Chiara, M.D., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel,
1275 K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates.
1276 Nature *556*, 339. 10.1038/s41586-018-0030-5.

1277 41. Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B.,
1278 Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and Mode of Genome Evolution in the
1279 Budding Yeast Subphylum. Cell *175*, 1533-1545.e20. 10.1016/j.cell.2018.10.023.

1280 42. Choudhary, S., Li, W., and D. Smith, A. (2020). Accurate detection of short and long active ORFs
1281 using Ribo-seq data. Bioinformatics *36*, 2053–2059. 10.1093/bioinformatics/btz878.

1282 43. Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T.,
1283 Schroeder, M., et al. (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res. *26*, 73–79.
1284 10.1093/nar/26.1.73.

1285 44. Gerashchenko, M.V., and Gladyshev, V.N. (2014). Translation inhibitors cause abnormalities in
1286 ribosome profiling experiments. Nucleic Acids Res. *42*, e134–e134. 10.1093/nar/gku671.

1287 45. Santos, D.A., Shi, L., Tu, B.P., and Weissman, J.S. (2019). Cycloheximide can distort measurements of
1288 mRNA levels and translation efficiency. Nucleic Acids Res. *47*, 4974–4985. 10.1093/nar/gkz205.

1289 46. Duncan, C.D.S., and Mata, J. (2017). Effects of cycloheximide on the interpretation of ribosome
1290 profiling experiments in Schizosaccharomyces pombe. Sci. Rep. *7*, 10331. 10.1038/s41598-017-
1291 10650-1.

1292 47. Pelechano, V., Wei, W., Jakob, P., and Steinmetz, L.M. (2014). Genome-wide identification of
1293 transcript start and end sites by transcript isoform sequencing. Nat. Protoc. *9*, 1740–1759.
1294 10.1038/nprot.2014.121.

1295 48. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
1296 Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.
1297 10.1111/j.2517-6161.1995.tb02031.x.

1298 49. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are
1299 translated and some are likely to express functional proteins. eLife *4*, e08890. 10.7554/eLife.08890.

1300 50. Moro, S.G., Hermans, C., Ruiz-Orera, J., and Albà, M.M. (2021). Impact of uORFs in mediating
1301 regulation of translation in stress conditions. BMC Mol. Cell Biol. *22*, 29. 10.1186/s12860-021-
1302 00363-9.

1303 51. Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and
1304 Hittinger, C.T. (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences
1305 and Strain Resources for the Saccharomyces sensu stricto Genus. G3 Genes Genomes Genet. *1*, 11–
1306 25. 10.1534/g3.111.000273.

1307 52. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-
1308 Bonet, T., and Albà, M.M. (2015). Origins of De Novo Genes in Human and Chimpanzee. PLOS Genet.
1309 *11*, e1005721. 10.1371/journal.pgen.1005721.

1310 53. Firth, A.E. (2014). Mapping overlapping functional elements embedded within the protein-coding
1311 regions of RNA viruses. Nucleic Acids Res. *42*, 12425–12439. 10.1093/nar/gku981.

1312 54. Sealfon, R.S., Lin, M.F., Jungreis, I., Wolf, M.Y., Kellis, M., and Sabeti, P.C. (2015). FRESCo: finding
1313 regions of excess synonymous constraint in diverse viruses. Genome Biol. *16*, 38. 10.1186/s13059-
1314 015-0603-7.

1315 55. Hardison, R.C. (2003). Comparative Genomics. PLOS Biol. *1*, e58. 10.1371/journal.pbio.0000058.

1316 56. Dujon, B. (1996). The yeast genome project: what did we learn? Trends Genet. *12*, 263–270.
1317 10.1016/0168-9525(96)10027-5.

1318 57. Dujon, B., Alexandraki, D., André, B., Ansorge, W., Baladron, V., Ballesta, J.P.G., Banrevi, A., Bolle,
1319 P.A., Bolotin-Fukuhara, M., Bossier, P., et al. (1994). Complete DNA sequence of yeast chromosome
1320 XI. Nature *369*, 371–378. 10.1038/369371a0.

1321 58. Chong, Y.T., Koh, J.L.Y., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., and
1322 Andrews, B.J. (2015). Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis.
1323 Cell *161*, 1413–1424. 10.1016/j.cell.2015.04.051.

1324 59. Koh, J.L.Y., Chong, Y.T., Friesen, H., Moses, A., Boone, C., Andrews, B.J., and Moffat, J. (2015).
1325 CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance
1326 and Subcellular Localization Patterns in Saccharomyces cerevisiae. G3 GenesGenomesGenetics *5*,
1327 1223–1232. 10.1534/g3.115.017830.

1328 60. Meurer, M., Duan, Y., Sass, E., Kats, I., Herbst, K., Buchmuller, B.C., Dederer, V., Huber, F., Kirrmaier,
1329 D., Štefl, M., et al. (2018). Genome-wide C-SWAT library for high-throughput yeast genome tagging.
1330 Nat. Methods *15*, 598–600. 10.1038/s41592-018-0045-8.

1331 61. Dubreuil, B., Sass, E., Nadav, Y., Heidenreich, M., Georgeson, J.M., Weill, U., Duan, Y., Meurer, M.,
1332 Schuldiner, M., Knop, M., et al. (2019). YeastRGB: comparing the abundance and localization of
1333 yeast proteins across cells and libraries. Nucleic Acids Res. *47*, D1245–D1249. 10.1093/nar/gky941.

1334 62. Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., and Wang, W. (2010). A *de novo* originated gene depresses
1335 budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell
1336 Res. *20*, 408–420. 10.1038/cr.2010.31.

1337 63. Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and
1338 Lafontaine, I. (2018). A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. *35*, 631–645.
1339 10.1093/molbev/msx315.

`

1340 64. Li, D., Yan, Z., Lu, L., Jiang, H., and Wang, W. (2014). Pleiotropy of the de novo-originated gene
1341 MDF1. Sci. Rep. *4*, 7280. 10.1038/srep07280.

1342 65. Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S.B., Wacholder, A., Medetgul-Ernar, K.,
1343 Bowman, R.W., Hines, C.P., Iannotta, J., et al. (2020). De novo emergence of adaptive membrane
1344 proteins from thymine-rich genomic sequences. Nat. Commun. *11*, 1–18. 10.1038/s41467-020-
1345 14500-z.

1346 66. Omidi, K., Jessulat, M., Hooshyar, M., Burnside, D., Schoenrock, A., Kazmirchuk, T., Hajikarimlou, M.,
1347 Daniel, M., Moteshareie, H., Bhojoo, U., et al. (2018). Uncharacterized ORF HUR1 influences the
1348 efficiency of non-homologous end-joining repair in Saccharomyces cerevisiae. Gene *639*, 128–136.
1349 10.1016/j.gene.2017.10.003.

1350 67. Hajikarimlou, M., Moteshareie, H., Omidi, K., Hooshyar, M., Shaikho, S., Kazmirchuk, T., Burnside, D.,
1351 Takallou, S., Zare, N., Jagadeesan, S.K., et al. (2020). Sensitivity of yeast to lithium chloride connects
1352 the activity of YTA6 and YPR096C to translation of structured mRNAs. PLOS ONE *15*, e0235033.
1353 10.1371/journal.pone.0235033.

1354 68. Alesso, C.A., Discola, K.F., and Monteiro, G. (2015). The gene ICS3 from the yeast Saccharomyces
1355 cerevisiae is involved in copper homeostasis dependent on extracellular pH. Fungal Genet. Biol. *82*,
1356 43–50. 10.1016/j.fgb.2015.06.007.

1357 69. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M.,
1358 Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of
1359 cellular function. Science *353*, aaf1420. 10.1126/science.aaf1420.

1360 70. Kearse, M.G., and Wilusz, J.E. (2017). Non-AUG translation: a new start for protein synthesis in
1361 eukaryotes. Genes Dev. *31*, 1717–1731. 10.1101/gad.305250.117.

1362 71. Loughran, G., Zhdanov, A.V., Mikhaylova, M.S., Rozov, F.N., Datskevich, P.N., Kovalchuk, S.I.,
1363 Serebryakova, M.V., Kiniry, S.J., Michel, A.M., O'Connor, P.B.F., et al. (2020). Unusually efficient CUG
1364 initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. Proc. Natl.
1365 Acad. Sci. *117*, 24936–24946. 10.1073/pnas.2001433117.

1366 72. McVeigh, A., Fasano, A., Scott, D.A., Jelacic, S., Moseley, S.L., Robertson, D.C., and Savarino, S.J.
1367 (2000). IS1414, an Escherichia coliInsertion Sequence with a Heat-Stable Enterotoxin Gene
1368 Embedded in a Transposase-Like Gene. Infect. Immun. *68*, 5710–5715. 10.1128/IAI.68.10.5710-
1369 5715.2000.

1370 73. Wright, B.W., Yi, Z., Weissman, J.S., and Chen, J. (2022). The dark proteome: translation from
1371 noncanonical open reading frames. Trends Cell Biol. *32*, 243–258. 10.1016/j.tcb.2021.10.010.

1372 74. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and Tautz,
1373 D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife *8*.
1374 10.7554/eLife.44392.

1375 75. Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y.,
1376 Gerstein, M., and Snyder, M. (2007). Divergence of Transcription Factor Binding Sites Across Related
1377 Yeast Species. Science *317*, 815–819. 10.1126/science.1140748.

`

1378 76. Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene
1379 expression. WIREs RNA *5*, 765–768. 10.1002/wrna.1245.

1380 77. Wu, Q., Wright, M., Gogol, M.M., Bradford, W.D., Zhang, N., and Bazzini, A.A. (2020). Translation of
1381 small downstream ORFs enhances translation of canonical main open reading frames. EMBO J. *39*,
1382 e104763. 10.15252/embj.2020104763.

1383 78. Andjus, S., Szachnowski, U., Vogt, N., Hatin, I., Papadopoulos, C., Lopes, A., Namy, O., Wery, M., and
1384 Morillon, A. (2022). Translation is a key determinant controlling the fate of cytoplasmic long non-
1385 coding RNAs. 2022.05.25.493276. 10.1101/2022.05.25.493276.

1386 79. Wery, M., Descrimes, M., Vogt, N., Dallongeville, A.-S., Gautheret, D., and Morillon, A. (2016).
1387 Nonsense-Mediated Decay Restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded
1388 RNA Structure. Mol. Cell *61*, 379–392. 10.1016/j.molcel.2015.12.020.

1389 80. Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N.,
1390 Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature
1391 *458*, 337–341. 10.1038/nature07743.

1392 81. Liti, G., Ba, A.N.N., Blythe, M., Müller, C.A., Bergström, A., Cubillos, F.A., Dafhnis-Calas, F.,
1393 Khoshraftar, S., Malla, S., Mehta, N., et al. (2013). High quality de novo sequencing and assembly of
1394 the Saccharomyces arboricolus genome. BMC Genomics *14*, 69. 10.1186/1471-2164-14-69.

1395 82. Naseeb, S., Alsammar, H., Burgis, T., Donaldson, I., Knyazev, N., Knight, C., and Delneri, D. (2018).
1396 Whole Genome Sequencing, de Novo Assembly and Phenotypic Profiling for the New Budding Yeast
1397 Species Saccharomyces jurei. G3 Genes Genomes Genet. *8*, 2967–2977. 10.1534/g3.118.200476.

1398 83. Bauer, S., Grossmann, S., Vingron, M., and Robinson, P.N. (2008). Ontologizer 2.0--a multifunctional
1399 tool for GO term enrichment analysis and data exploration. Bioinforma. Oxf. Engl. *24*, 1650–1651.
1400 10.1093/bioinformatics/btn250.

1401 84. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput.
1402 Nucleic Acids Res. *32*, 1792–1797. 10.1093/nar/gkh340.

1403 85. Dunham, M.J., Dunham, M.J., Gartenberg, M.R., and Brown, G.W. (2015). Methods in yeast genetics
1404 and genomics: a cold spring harbor laboratory course manual (Cold Spring Harbor Laboratory Press).

1405 86. Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, on behalf of the I.N.S.D. (2011). The
1406 Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21. 10.1093/nar/gkq1019.

1407 87. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque,
1408 N., Goodgame, N., Gibson, R., et al. (2011). The European Nucleotide Archive. Nucleic Acids Res. *39*,
1409 D28–D31. 10.1093/nar/gkq967.

1410 88. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol.
1411 Biol. *147*, 195–197. 10.1016/0022-2836(81)90087-5.

`

1412 89. Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In Multiple Sequence Alignment
1413 Methods Methods in Molecular Biology., D. J. Russell, ed. (Humana Press), pp. 155–170.
1414 10.1007/978-1-62703-646-7_10.

1415 90. Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a
1416 source of new peptides. eLife *3*. 10.7554/eLife.03523.

1417 91. Usaj, M., Tan, Y., Wang, W., VanderSluis, B., Zou, A., Myers, C.L., Costanzo, M., Andrews, B., and
1418 Boone, C. (2017). TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global
1419 Yeast Genetic Interaction Network. G3 GenesGenomesGenetics *7*, 1539–1549.
1420 10.1534/g3.117.040220.

1421 92. Parikh, S.B., Castilho Coelho, N., and Carvunis, A.-R. (2021). LI Detector: a framework for sensitive
1422 colony-based screens regardless of the distribution of fitness effects. G3 GenesGenomesGenetics
1423 *11*, jkaa068. 10.1093/g3journal/jkaa068.

1424