# Exploring the noncanonical translatome using massively integrated coexpression analysis

**April Rich*[1,2,3], Omer Acar*[1,2,3], Anne-Ruxandra Carvunis[2,3]**

*[1]Joint Carnegie Mellon University-University of Pittsburgh Computational Biology PhD Program, University of Pittsburgh, Pittsburgh, PA, USA; [2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [3]Pittsburgh Center for Evolutionary Biology and Medicine (CEBaM), University of Pittsburgh, Pittsburgh, PA, USA*

*co-first authors

# Abstract

Cells transcribe and translate thousands of noncanonical open reading frames (nORFs) whose impacts on cellular phenotypes are unknown. Here, we investigated nORF transcription, evolution, and potential cellular roles using a coexpression approach. We measured coexpression between ~6,000 nORFs and ~6000 canonical ORFs (cORFs) in the *Saccharomyces cerevisiae* genome by massively integrating thousands of RNA sequencing samples and developing a dedicated computational framework that accounts for low expression levels. Our findings reveal that almost all cORFs are strongly coexpressed with at least one nORF. However, almost half of nORFs are not strongly coexpressed with any cORFs and form entirely new transcription modules. Many nORFs recently evolved *de novo* in genomic regions that were non-coding in the *Saccharomyces* ancestor. Coexpression profiles suggest that half of *de novo* nORFs are functionally associated with conserved genes involved in cellular transport or homeostasis. Furthermore, we discovered that *de novo* ORFs located downstream of conserved genes leverage their neighbors' transcripts resulting in high expression levels. Where a *de novo* nORF emerges could be just as important as its sequence for shaping how it can influence cellular phenotype. Our coexpression dataset serves as an unprecedented resource for unraveling how nORFs integrate into cellular networks, contribute to cellular phenotypes and evolve.

# Introduction

Eukaryotic genomes contain thousands of open reading frames (ORFs), including noncanonical ORFs (nORFs) that have long been considered unlikely to have any cellular roles[1]. Being very short and lacking evolutionary conservation, nORFs were historically excluded from genome annotations[2,3]. Yet the development of RNA sequencing (RNA-seq)[4] and ribosome profiling[5,6] has shown genome-wide transcription and translation of nORFs in yeast[5,7], zebrafish[8], flies[9], mammalian cell lines[10,11] and in humans[12–14]. As a result, unraveling the cellular, physiological and evolutionary implications of nORFs' has become an active area of research[7,15].

There are two major obstacles to studying nORFs. The first is detection, as nORF expression levels are typically low and dependent on specific conditions[16,17]. However, it has been recently shown that massive integration of RNA-seq[18,19] or ribosome profiling[7] experiments is an effective way to overcome these detection issues. For example, Wacholder et al.[7] recently discovered around 19,000 translated nORFs in *Saccharomyces cerevisiae* by massive integration of ribosome profiling data. This number is four times larger than the number of canonical ORFs (cORFs) annotated in the yeast genome, demonstrating the power of aggregating publicly available data for identifying translated nORFs. The second obstacle is biological interpretation. Transcription or translation of nORFs could be attributed to expression noise[20,21] whereby non-specific binding of RNA polymerases and ribosomes to DNA and RNA might cause promiscuous transcription or translation, respectively. However, many studies have shown that nORFs can form stable peptides and affect phenotypes including cell growth[22], cell cycle regulation[23], muscle physiology[24], and immunity[25]. These studies showed that there is a pool of nORFs whose expression is not mere noise and whose translation products are important for cellular life.

Many translated nORFs have evolved *de novo* from previously non-coding loci[16,26]. However, how *de novo* ORFs gain the ability to be transcribed in the first place is unclear[27]. One possibility is that novel regulatory regions emerge alongside with the emergence of the ORF, enabling new ORFs to be regulated by their own regulatory landscape (ORF-first) as it was shown for *de novo* ORFs in *Drosophila melanogaster*[28], codfish[29], human[17,30] and chimpanzee[17]. Alternatively, ORFs may emerge on actively transcribed loci such as long noncoding RNAs as it was shown for *de novo* ORFs in primates[31] or upstream or downstream of transcripts containing genes[30] (transcription-first)[32–34]. Transcriptional activation has a ripple effect causing coordinated activation of nearby genes[35,36]. Thus, *de novo* ORFs that emerge near established genes or regulatory regions may acquire the ability to be transcribed by 'piggybacking'[36] on the regulatory context of their established gene neighbors[32,37].

Research investigating the impacts of existing regulatory context on *de novo* ORF transcription has yielded conflicting results. For instance, Vakirlis et al.[32] proposed that *de novo* ORFs located upstream of genes on the opposite strand tend to be transcribed from bi-directional promoters of genes and likely to be on nucleosome free regions (NFRs). However, Blevins et al.[38] proposed the opposite, that *de novo* ORFs are more likely to overlap with genes on the

opposite strand. Thus, there is a need to systematically assess the impacts of existing regulatory context, e.g., the orientation and distance relative to a gene that a *de novo* ORF emerges on, on a *de novo* ORF's transcriptional profile. Transcription is indeed the first step of expression that constrains where and when a *de novo* ORF might be translated and impact cellular functions.

Coexpression networks, where nodes represent ORFs and edges represent high correlation between transcriptional profiles, have been used successfully to identify new gene function[39,40], new disease-related genes[41–43] and for studying the conservation of the regulatory machinery[40,44] or gene modules[45] between species. Based on the assumption that genes involved in similar pathways have correlated expression patterns, coexpression networks can reveal relationships between genes and other transcribed genetic elements[46,47]. Most coexpression studies have focused on cORFs but a few recent studies suggest that coexpression networks are a useful tool for investigating nORFs as well. For instance, Stiens et al.[48] constructed a coexpression network for *Mycobacterium tuberculosis* to study unannotated transcripts and other studies have employed coexpression networks to study small ORFs in plants[49,50]. Work by Li et al.[18] in *S. cerevisiae* showed that many transcribed nORFs form coexpression clusters. These studies underscore the utility of coexpression as a valuable approach for studying the biology of functional elements at the genome scale.

The recent identification of nearly 19,000 translated nORFs in *S. cerevisiae*[7] – which have the potential to generate peptides that affect cellular phenotypes but are almost entirely uncharacterized – provides an unprecedented opportunity to leverage coexpression for understanding nORF biology and evolution at scale. Here, we addressed the statistical challenges arising from the low expression of nORFs to build the first high-quality coexpression network spanning the canonical and noncanonical translatome of any species.

# Results

## Massive integration of RNA-seq data shows noncanonical ORFs are poorly expressed yet are transcriptionally regulated

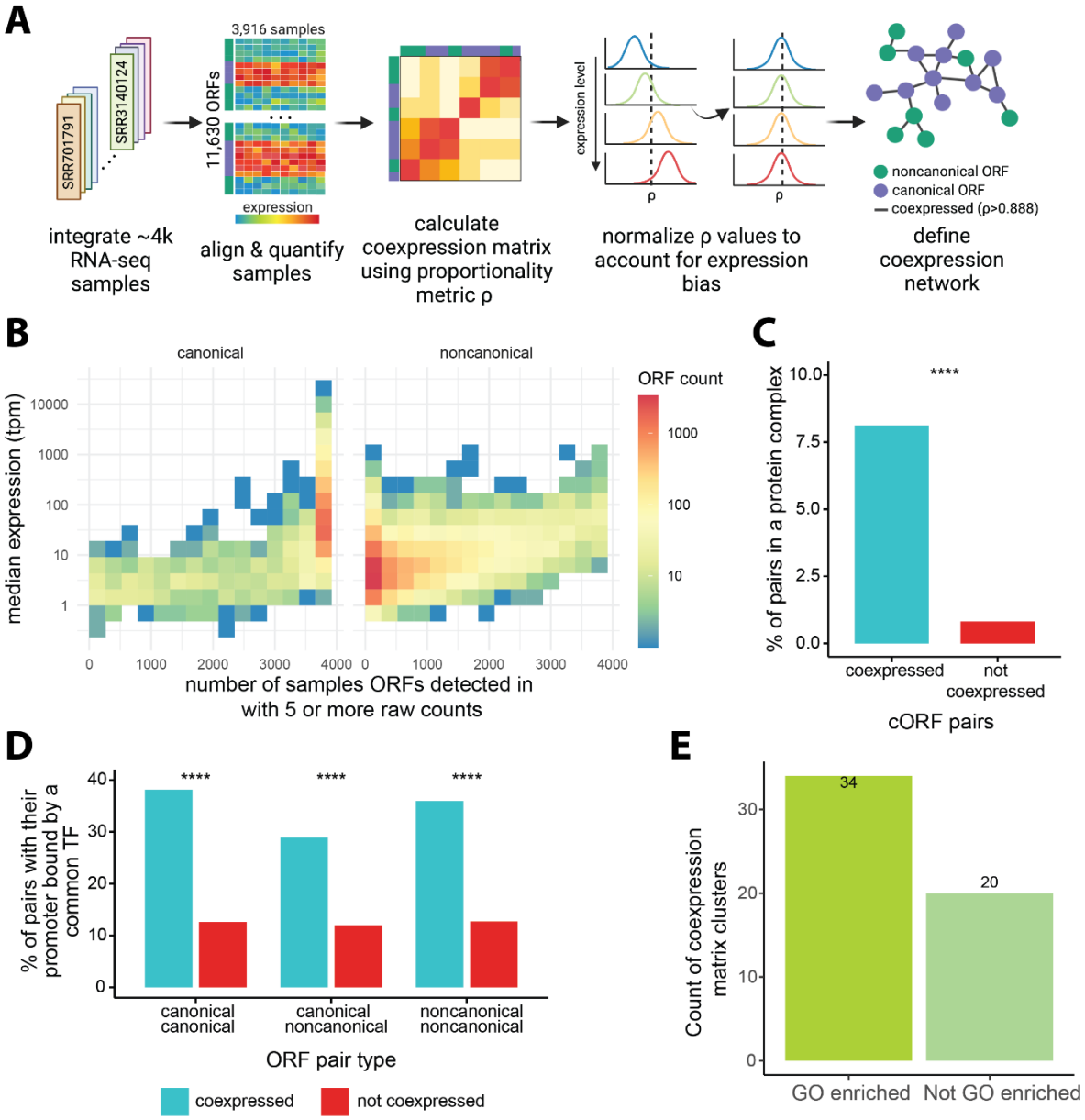



**Figure 1: Coexpression network including nORFs is biologically meaningful**
A) Workflow for creating expression dataset and coexpression analysis; 3,916 samples were analyzed to create an expression matrix for 11,630 ORFs, including 5,803 cORFs and 5,827 nORFs; expression values were used to calculate the coexpression matrix using proportionality

metric, ρ and were normalized to correct for expression bias. The coexpression matrix was thresholded using ρ > 0.888 to create a coexpression network (top 0.2% of all pairs). B) Distribution of the number of ORFs binned based on their median expression values (y-axis) and the number of samples the ORFs were detected in with at least 5 raw counts (x-axis). C) Using annotated protein complexes[51], coexpressed cORF pairs (ρ > 0.888) are more likely to form protein complexes than non-coexpressed cORF pairs (Fisher's exact test p < 2.2e-16). D) Using genome wide transcription factor (TF) binding profiles[52], coexpressed ORF pairs (ρ > 0.888) are more likely to have their promoters bound by a common TF than non-coexpressed ORF pairs (Fisher's exact test p < 2.2e-16). E) Hierarchical clustering of the coexpression matrix reveals functional enrichments for most clusters that contain at least 5 canonical ORFs; functional enrichments estimated by gene ontology enrichment analysis at FDR < 0.05 using Fisher's exact test.

To study coexpression at the translatome scale in *S. cerevisiae*, we considered all cORFs annotated as "verified", "uncharacterized", or "transposable element" in the *Saccharomyces* Genome Database (SGD)[53], as well as all nORFs, annotated as "dubious" and "pseudogene" or those that were unannotated, with evidence of translation according to Wacholder et al.[7] To maximize detection of transcripts containing nORFs, we integrated 3,916 publicly available RNA-seq samples from 174 studies (Figure 1A, Supplementary Data 1). Many nORFs were not detected in most of the samples we collected, creating a very sparse dataset (Figure 1B). The issue of sparsity has been widely studied in the context of single cell RNA-seq (scRNA-seq). A recent study looking at 17 measures of association for constructing coexpression networks from scRNA-seq showed that proportionality methods coupled with center log ratio (clr) transformation consistently outperformed other measures of coexpression in a variety of tasks including identification of disease-related genes and protein-protein network overlap analysis[54]. Thus, we used clr to transform the raw counts and quantified coexpression relationships using the proportionality metric, ρ[55].
We further addressed the issue of sparsity with two sample thresholding approaches. First, any observation with a raw count below five was discarded, such that when calculating ρ only the samples expressing both ORFs with at least five counts were considered. Second, we empirically determined that a minimum of 400 samples were required to obtain reliable coexpression values by assessing the effect of sample counts on the stability of ρ values (Supplementary Figure 1).

The combined use of clr, ρ, and sample thresholding accounts for statistical issues in estimating coexpression deriving from sparsity, but the large difference in expression levels between cORFs and nORFs poses yet another challenge. Indeed, Wang et al. showed that the distribution of coexpression values is biased by the expression level of a given ORF pair[56], where highly expressed genes also tended to be highly coexpressed due to statistical artifacts. Since nORFs are lowly expressed compared to cORFs, we observed this artifactual bias in our dataset (Supplementary Figure 2A). Therefore, we corrected this bias using spatial quantile normalization (SpQN) as recommended by Wang et al.[56] (Supplementary Figure 2B). Because of both the coexpression metric used and the quantile normalization of the values, the distribution of ρ in our expanded coexpression matrix is not centered at zero. These steps

resulted in a 11,630 by 11,630 coexpression matrix (Supplementary Data 2), with 5,803 cORFs and 5,827 nORFs (Supplementary Data 3).

We created a network representation of our coexpression matrix by considering only the top 0.2% of ρ values between all ORF pairs (ρ > 0.888). This threshold was chosen to include 90% of cORFs (Supplementary Figure 3). Altogether, our analysis resulted in an expanded coexpression network of 9,303 nodes (4,112 nORFs and 5,191 cORFs) and 124,382 edges (Figure 1A).

To assess whether our expanded coexpression network captures meaningful biological and regulatory relationships, we examined its overlap with orthogonal datasets. Using a previously published[51] curated protein complex dataset for cORFs, we found that coexpressed cORF pairs are significantly more likely to be in a protein complex together compared to non-coexpressed pairs (Odds ratio = 10.8 Fisher's exact test $p < 2.2e-16$; Figure 1C). Using a previously published[52] genome wide ChIP-exo dataset containing DNA-binding information for 73 sequence-specific transcription factors (TFs), we observed that coexpressed ORF pairs were more likely to have their promoters bound by a common TF than non-coexpressed ORF pairs, whether the pairs consist of nORFs or cORFs (canonical-canonical pairs: Odds ratio = 4.28, canonical-noncanonical pairs: Odds ratio = 3.0, noncanonical-noncanonical pairs: Odds ratio = 3.86, Fisher's exact test $p < 2.2e-16$ for all three comparisons; Figure 1D). Using the WGCNA[57] method to cluster the weighted coexpression matrix, we found that more than half of the clusters identified contained functionally related ORFs (GO biological process enrichments at FDR < 0.05; Figure 1E; Supplementary Figure 4). Altogether these analyses demonstrate the high quality of our expanded coexpression network and confirm that it captures meaningful biological relationships for both canonical and noncanonical ORFs.

# nORFs tend to be located at the periphery of the coexpression network and form new noncanonical transcription modules

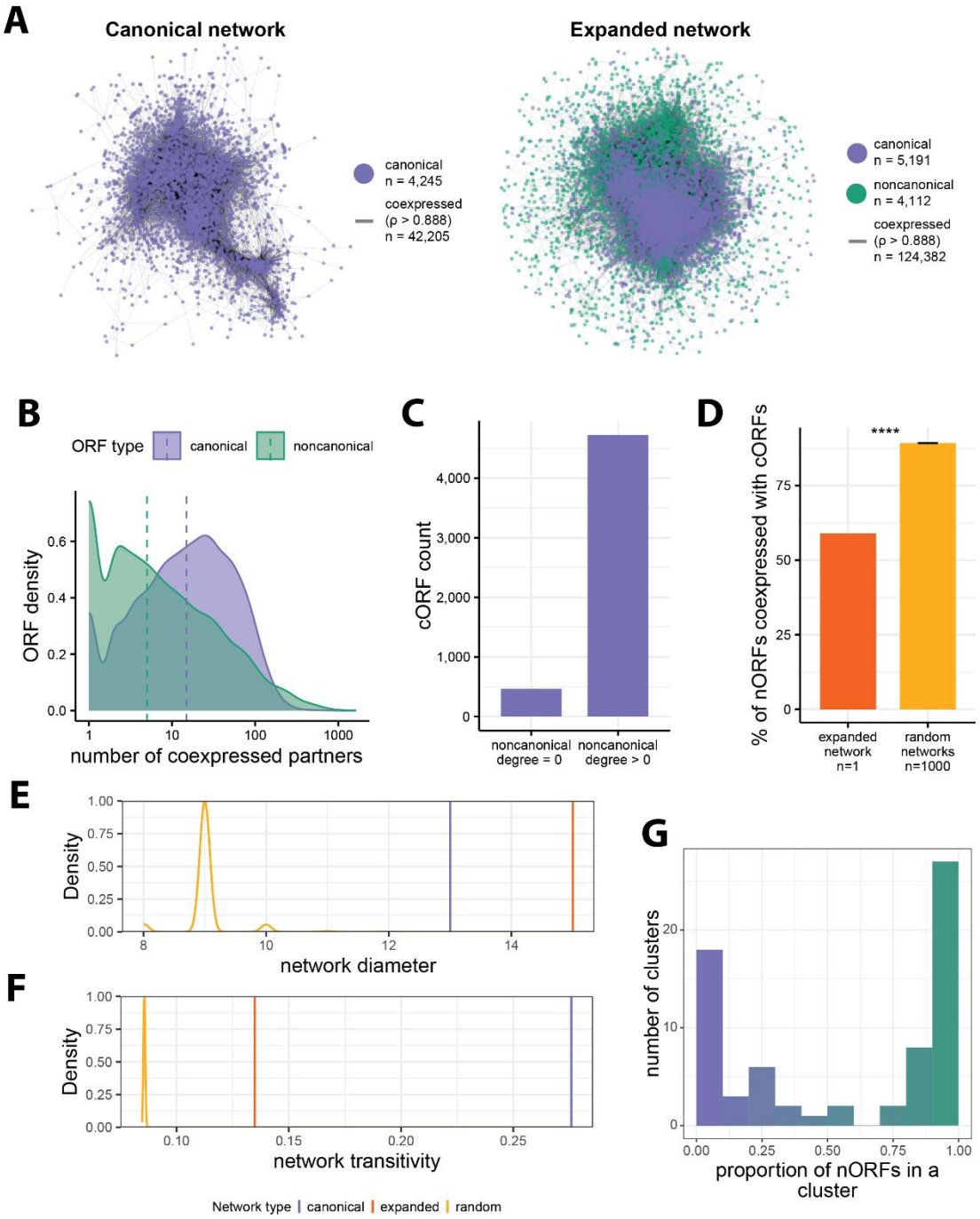



**Figure 2 nORFs have fewer connections yet coexpress with most cORFs and form new noncanonical transcription modules at the periphery of the coexpression network**

A) Visualization for canonical and expanded coexpression networks using spring embedded graph layout. Expanded network contains more cORFs than the canonical only network since addition of nORFs also results in addition of many cORFs that are only connected to an nORF.

B) nORFs have fewer coexpression partners (degree in expanded network) than cORFs (Mann-Whitney U-test $p < 2.2e-16$). C) Most cORFs are coexpressed with at least one nORF. D) 59% of nORFs are coexpressed with at least one cORFs and this is less than expected by chance (Fisher's exact test $p < 2.2e-16$). E) Addition of nORFs to the canonical network results in the expanded network being less dense, whereas the opposite is expected by chance, shown by the decrease in diameters for the 1,000 randomized networks. F) Addition of nORFs to the canonical network decreases local clustering in the expanded network, however this is to a lesser extent than expected by chance as shown by the distribution for the 1,000 randomized networks. G) Distribution of coexpression matrix cluster composition by WGCNA shows that most clusters are formed either by primarily nORFs or primarily cORFs (n= 69 clusters, *green* represents nORF majority clusters, *purple* represents cORF majority clusters).

We sought to investigate how the expanded coexpression network differs from the canonical coexpression network, which only contains cORFs and has 42,205 edges (Figure 2A). On average, nORFs have fewer coexpressed partners (degree) than cORFs, suggesting that nORFs have distinct transcriptional profiles (Cliff's Delta d = -0.29, Mann-Whitney U-test $p < 2.2e-16$, Figure 2B). We found that 91% of cORFs are coexpressed with at least one nORF (n = 4,726, Figure 2C), whereas only 59% of nORFs are coexpressed with at least one cORF. In contrast, we would have expected an average of 89% of nORFs to be coexpressed with a cORF according to degree preserving simulations of 1,000 randomized networks where edges from nORFs were shuffled (Odds ratio = 0.174, Fisher's exact test $p < 2.2e-16$, Figure 2D; Supplementary Figure 5). The randomized networks further helped us to confirm that the changes we observed when comparing the canonical and expanded networks were biologically meaningful and not a mere consequence of having a larger network. We analyzed two network properties for this purpose: diameter, which is the longest shortest path between any two ORFs, and transitivity, which is the tendency for ORFs that are coexpressed with a common neighbor to also be coexpressed with each other. The diameter of the randomized networks decreased upon the addition of nORFs, indicating a denser network when connections are random. This is in sharp contrast to the expanded network, where the addition of nORFs led to a larger diameter (Figure 2E). The network is thus much less dense compared to the canonical network and what would be expected by chance suggesting that nORFs tend to be located at the periphery of the network. Moreover, the transitivity of the expanded network decreased with the addition of nORFs compared to the canonical network but was higher than in the randomized networks, indicating that nORFs' connections create a more clustered network structure than expected by chance (Figure 2F). To further investigate this result, we inspected the ratio of nORFs and cORFs among the cluster assignments from WGCNA hierarchical clustering (Supplementary Figure 4). Strikingly, we observed a bimodal distribution of network clusters, with approximately half of the clusters consisting mostly of nORFs and the other half containing mostly cORFs (Figure 2G). Overall, our findings suggest that nORFs exhibit distinct expression patterns compared to cORFs but share well-structured coexpression relationships among themselves, leading to both integration within existing transcriptional modules as well as the creation of new ones.

# Coexpression profiles reveal associations between *de novo* ORFs and specific cellular processes

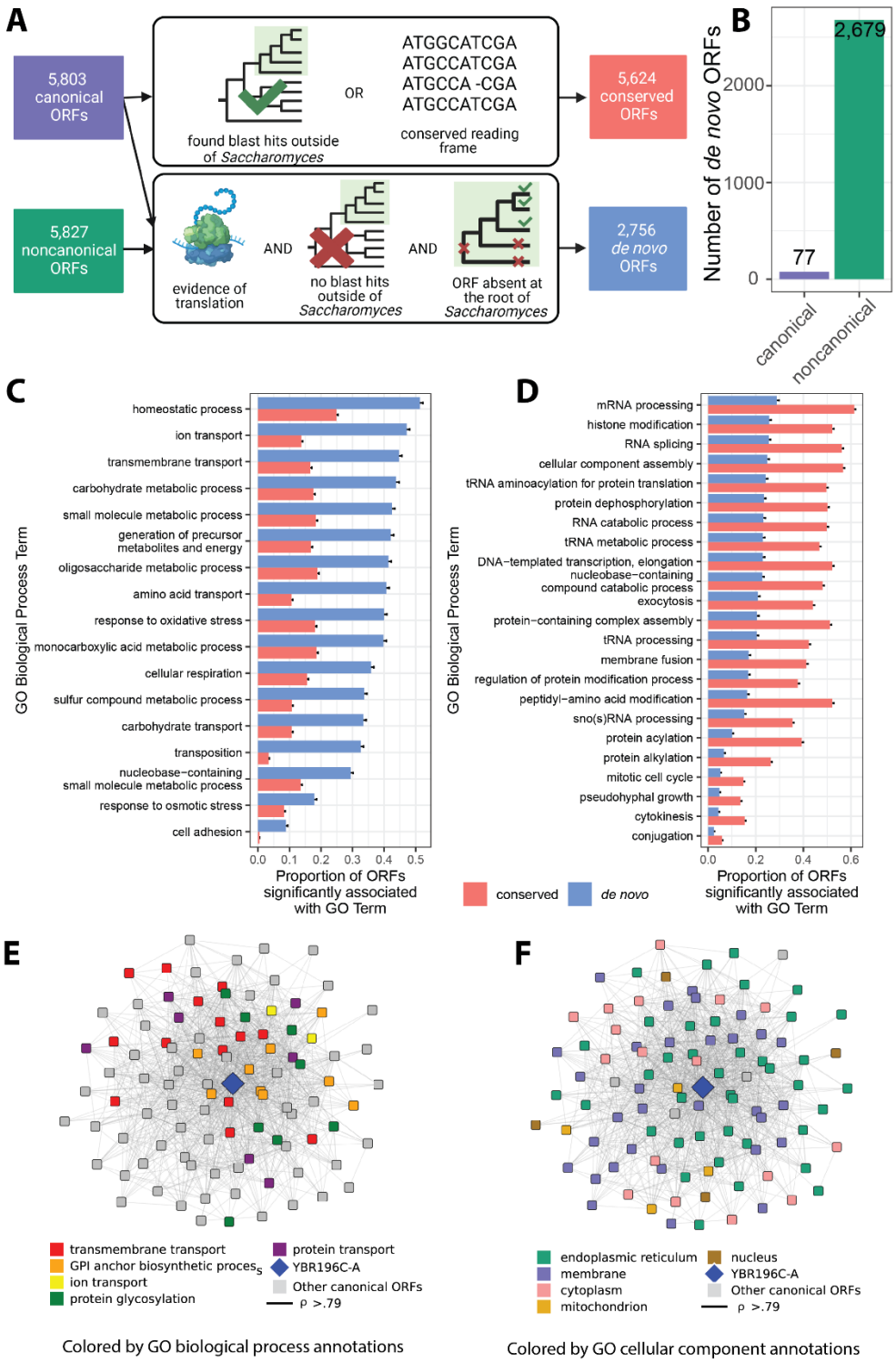


**Figure 3 More than half of nORFs evolved *de novo* and are associated with specific cellular processes as revealed by coexpression relationships**

A) Pipeline used to reclassify ORFs as conserved or *de novo*. cORFs were considered for both conserved and *de novo* classification while nORFs were only considered for *de novo* classification. Conserved ORFs were determined by either detection of homology outside of *Saccharomyces* or reading frame conservation (*top*). *De novo* ORFs were determined by evidence of translation, lack of homology outside of *Saccharomyces* as well as lack of a homologous ORF in the two most distant *Saccharomyces* branches. B) Counts of cORFs and nORFs that were found *de novo*. C-D) GO terms that proportionally have more (C) (Odds ratio > 2, n=17 terms) or less (D)(Odds ratio < 0.5, n=23 terms) GSEA enrichments with *de novo* ORFs compared to conserved ORFs (y-axis ordered by *de novo* ORF enrichment proportion from highest to lowest, FDR < 0.001 for all terms, Fisher's exact test). E-F) The top 100 coexpression partners of *de novo* cORF YBR196C-A colored by biological process (E) or cellular component annotations (F).

To define which ORFs in our dataset were evolutionarily conserved and which were of recent *de novo* evolutionary origins, we developed a multistep pipeline combining sequence similarity searches and syntenic alignments (Figure 3A). ORFs were considered conserved if they had homologues detectable by sequence similarity in budding yeasts outside of the *Saccharomyces* genus or if their reading frames were maintained within the *Saccharomyces* genus. ORFs were considered *de novo* if they lacked homologues detectable by sequence similarity outside of the *Saccharomyces* genus and if less than 60% of syntenic orthologous nucleotides in the two most distant *Saccharomyces* branches were in the same reading frame as in *S. cerevisiae*. These analyses identified 5,624 conserved cORFs and 2,756 *de novo* ORFs including 77 *de novo* cORFs and 2,679 *de novo* nORFs (Figure 3B).

To determine whether *de novo* ORFs might be associated with specific cellular processes, we performed gene set enrichment analyses[58] (GSEA) on the coexpression profiles with cORFs for all conserved and *de novo* ORFs in our coexpression matrix. GSEA takes an ordered list of genes, in this case sorted by coexpression level, and seeks to find if those genes placed primarily in front of the ordered list are annotated with specific GO terms. For each conserved and *de novo* ORF, GSEA allowed us to detect if an ORF's highly coexpressed partners are significantly associated with any GO terms (Supplementary Figure 6). We then calculated, for each GO term, the number of conserved and *de novo* ORFs that had GSEA enrichments at FDR < 0.01 (Supplementary Data 4). These analyses identified 17 specific GO terms that are more prevalent among the coexpression partners of *de novo* ORFs relative to those of conserved ORFs (Figure 3C, FDR < 0.001, Odds ratio > 2, Fisher's exact test, Supplementary Data 5). Homeostatic process (GO:0042592) had the highest proportion of *de novo* ORFs enrichments with 51% (n=1416) followed by ion transport (GO:0006811, 47%), transmembrane transport (GO:0055085, 44%) as well as other transport-related processes, metabolic processes, or response to stressors such as oxidative or osmotic stresses. We additionally found 23 terms that are less likely to be among coexpression partners of *de novo* ORFs (FDR < 0.001, Odds ratio < 0.5, Fisher's exact test). These terms were related to a variety of processes such as protein modifications, cellular growth, or RNA processing (Figure 3D, Supplementary

Data 5). These results suggest that *de novo* ORF expression is regulated by transcriptional programs.

We found the enrichment of transport-related GO terms particularly interesting. Ion transport, transmembrane transport, amino acid transport and carbohydrate transport, all had odds ratio > 2. Almost half (n=1,289, 47%) of *de novo* ORFs tend to have higher coexpression with the cORFs annotated for the parent 'transport' GO term (GO:0006810, GSEA FDR < 0.01). Among these 1,289 transport-enriched *de novo* ORFs, 31 were cORFs, which allowed us to search literature to find orthogonal evidence for their role in transport. Four *de novo* cORFs, *YIR020C*, *YLR406C-A*, *YEL068C* and *YBR196C-A*, were previously found to be localized[22,59] to the endoplasmic reticulum (ER) or the vacuole. Our GSEA enrichments and their localizations suggest that they might be involved in transport at these organelles.

*YBR196C-A* was previously described as an example of *de novo* ORF that is adaptive when overexpressed and integrates into the membrane of the endoplasmic reticulum.[22] Coexpression relationships in our matrix showed GSEA enrichments for *YBR196C-A* in transmembrane transport (GO:0055085, FDR = 1.95e-03), ion transport (GO:0006811, FDR = 1.46e-05), Golgi-vesicle transport (GO:0048193, FDR = 5.56e-04), vesicle-mediated transport (GO:0016192, FDR = 1.82e-07), and endosomal transport (GO:0016197, FDR = 5.01e-03, Supplementary Figure 7). Out of the top 100 ORFs coexpressed with YBR196C-A, 33 were involved in various biological processes related to membranes (Figure 3E) and 70 were annotated with membrane localization (FDR = 7.7e-13, Fisher's exact test) including 39 with ER membrane (FDR = 7.21e-12, Fisher's exact test, Figure 3F). Overall, the results from our coexpression matrix suggest that the expression of *de novo* ORFs is associated with a diverse set of cellular processes and around half of *de novo* ORFs are associated with the transcriptional programs regulating transport and homeostasis.

# *de novo* ORF expression and regulation are shaped by genomic location

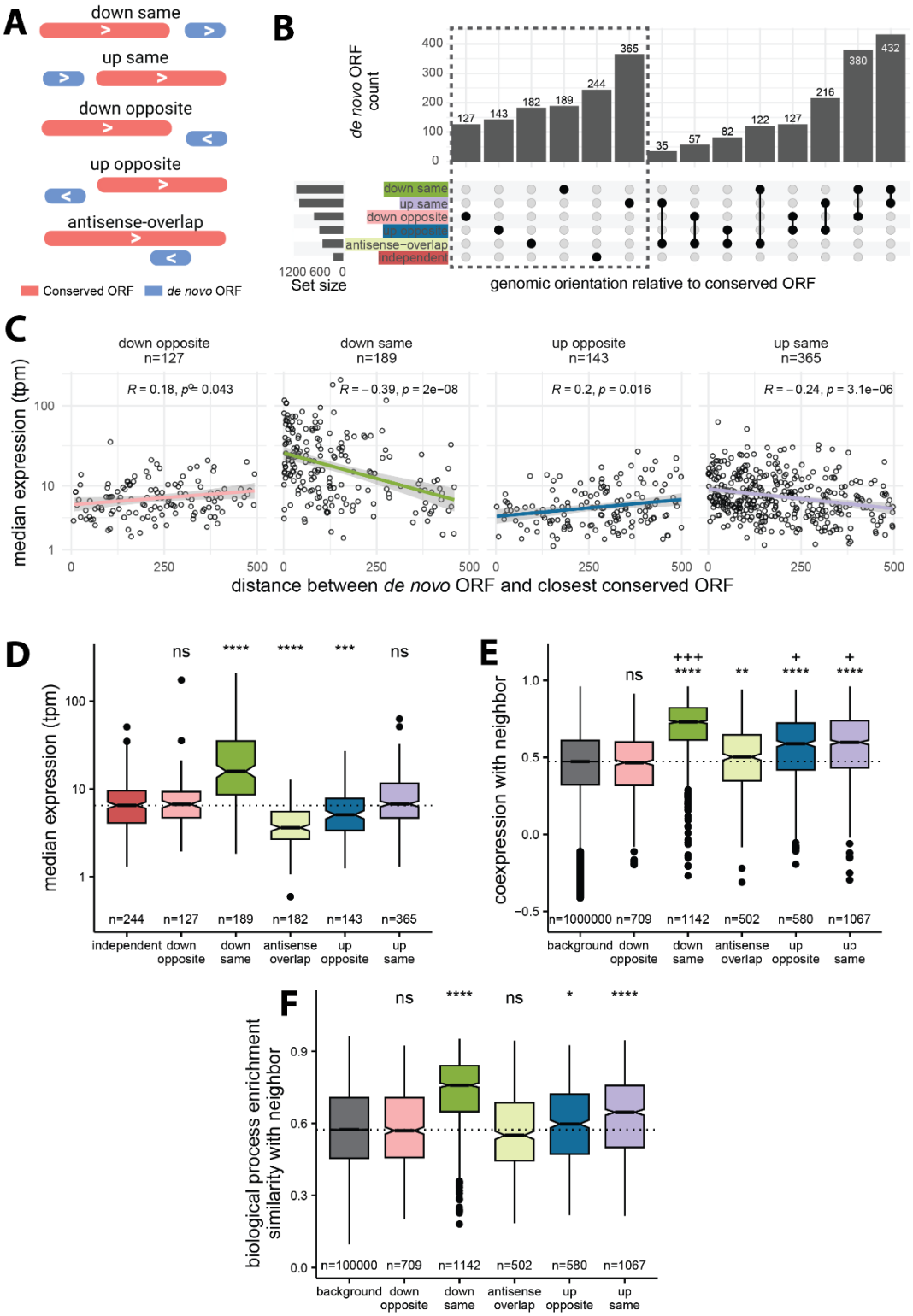


**Figure 4 Genomic orientation influences the expression and potential cellular roles of *de novo* ORFs**

A) Possible genomic orientations of *de novo* ORFs relative to neighboring conserved ORFs within 500bp. ORFs that are further away than 500bp are classified as independent. B) Counts of *de novo* ORFs that are within 500 bp of a conserved ORF in different genomic orientations (dashed box represents subgroups that were used for panels *C* and *D*). C) Expression level (median TPM across all RNA-seq samples) of *de novo* ORFs is influenced by distance and genomic orientation (considering only ORFs in a single orientation, dashed box in panel B, *R*: Spearman's correlation coefficient and *p*: p value for significance of correlation). D) *De novo* ORFs located downstream on the same strand as conserved ORFs have the highest expression among different orientations (considering only ORFs in only a single orientation, dashed box in panel B). E) Down same *de novo* ORFs tend to be highly coexpressed with their neighbors; background: *de novo*-conserved ORF pairs located on different chromosomes. F) Down same *de novo* ORFs tend to be coexpressed with cORFs associated with similar biological processes as the cORFs that are coexpressed with their neighbors; similarity measured by calculating semantic similarity between GSEA enrichments for neighboring *de novo*-conserved ORF pairs using relevance metric. (0 = no similarity, 1 = perfect overlap) (Mann-Whitney U-test, ****: $p \leq 0.0001$, ***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$, ns: not-significant, +: small effect size (Cliff's $d < .33$), ++: medium effect size (Cliff's $d < .474$), +++: large effect size (Cliff's $d \geq .474$))

Having shown that many of these nORFs are evolutionarily young with *de novo* origins and yet are transcribed, translated, and exhibit biologically coherent coexpression patterns with cORFs, we investigated how the genomic locus where a *de novo* ORF emerges influences its potential evolutionary path and integration into the cellular network.

We categorized *de novo* ORFs based on their positioning relative to neighboring conserved ORFs. The *de novo* ORFs that were not within 500 bp of a conserved ORF were classified as independent. The remaining *de novo* ORFs were categorized as either upstream or downstream on the same strand (up same or down same), upstream or downstream on the opposite strand (up opposite or down opposite), or as overlapping on the opposite strand (anti-sense overlap) based on their orientation to nearest conserved ORF (Figure 4A-B).

We investigated how orientation and distance from a conserved ORF influence the expression of *de novo* ORFs, by restricting our analyses to those *de novo* ORFs assigned to only a single orientation (dashed box in Figure 4B). We found that the ORFs located on the same strand as their neighboring conserved ORFs exhibit a negative correlation in expression with their distance to their conserved neighbors, while those located on the upstream on the opposite strand have a positive correlation on their expression. (up same: $R = -0.24$, $p = 3.1e-6$; down same: $R = -0.39$, $p = 2e-8$; up opposite: $R = 0.2$, $p = 0.016$; Spearman's Correlation Coefficient, Figure 4C). Transcription of the neighboring conserved ORFs seems to enhance the transcription of nearby *de novo* ORFs on the same strand possibly through piggybacking, while has an inhibiting effect for *de novo* ORFs on the opposite strand possibly through transcriptional interference.

Generally, down same *de novo* ORFs display significantly higher expression levels compared to independent *de novo* ORFs (Cliff's Delta d = 0.58, Mann-Whitney U-test, p < 2.2e-16) as well as to *de novo* ORFs in other orientations (Figure 4D). In contrast, antisense overlap and up opposite *de novo* ORFs have significantly lower expression levels than independent *de novo* ORFs (antisense overlap: Cliff's Delta d = -0.52, p < 2.2e-16; up opposite: Cliff's Delta d = -0.22, p = 2.6e-4; Mann-Whitney U-test). We did not observe any expression differences between down opposite or up same ORFs compared to independent *de novo* ORFs.

Distance negatively influenced coexpression in an orientation dependent manner, except for down opposite ORFs (Supplementary Figure 8). To further investigate this, we examined whether neighboring pairs of *de novo*-conserved ORFs exhibited higher coexpression compared to a background distribution based on randomly selected *de novo*-conserved ORF pairs located on separate chromosomes (Figure 4E). Down same *de novo* ORFs showed the largest increase in coexpression with neighboring conserved ORFs compared to background pairs (Cliff's Delta d = 0.64, Mann-Whitney U-test, p < 2.2e-16), while up opposite and up same *de novo* ORFs had small increases in coexpression with neighboring conserved ORFs (up opposite: Cliff's Delta d = 0.28; up same: Cliff's Delta d = 0.32; Mann-Whitney U-test, p < 2.2e-16 for both comparisons, Figure 4E). Despite showing an increase in coexpression with nearby conserved ORFs, the *de novo* ORFs did not exhibit the strongest coexpression with their immediate neighboring conserved ORFs, i.e., neighboring conserved ORFs were only rarely (6%, n=240) in the top 10 coexpressed ORFs (15% of down same, 4.5% of up same, 3% of up opposite and 1 % of anti-sense overlap). To explore the implications of these observations, we compared the biological processes associated by coexpression of each *de novo* ORF to those of their neighboring conserved ORFs (Figure 4F). To calculate biological process similarity between two ORFs, we used significant GO terms at FDR<0.01 determined by GSEA for each ORF and calculated similarity between these two sets of GO terms using the relevance method.[60] If two ORFs are enriched in same specialized terms, their relevance metric would be higher and if they are enriched in different terms or same but generic terms, their relevance would be lower. We found that *de novo* ORFs in the down same and up same orientations are significantly more likely to share similar biological process enrichments with neighboring conserved ORFs than background ORF pairs (Cliff's Delta d = 0.5 and d = 0.17, respectively, Mann-Whitney U-test, p-value < 2.2e-16 for both) and pairs in other orientations.
Overall, these results show that *de novo* ORFs located downstream on the same strand as conserved ORFs exhibit higher expression levels as well as higher coexpression and functional similarity with their neighboring conserved ORF than *de novo* ORFs that are located further away or in other orientations

## Sharing a transcript is a major piggybacking mechanism for *de novo* ORFs located near conserved ORFs

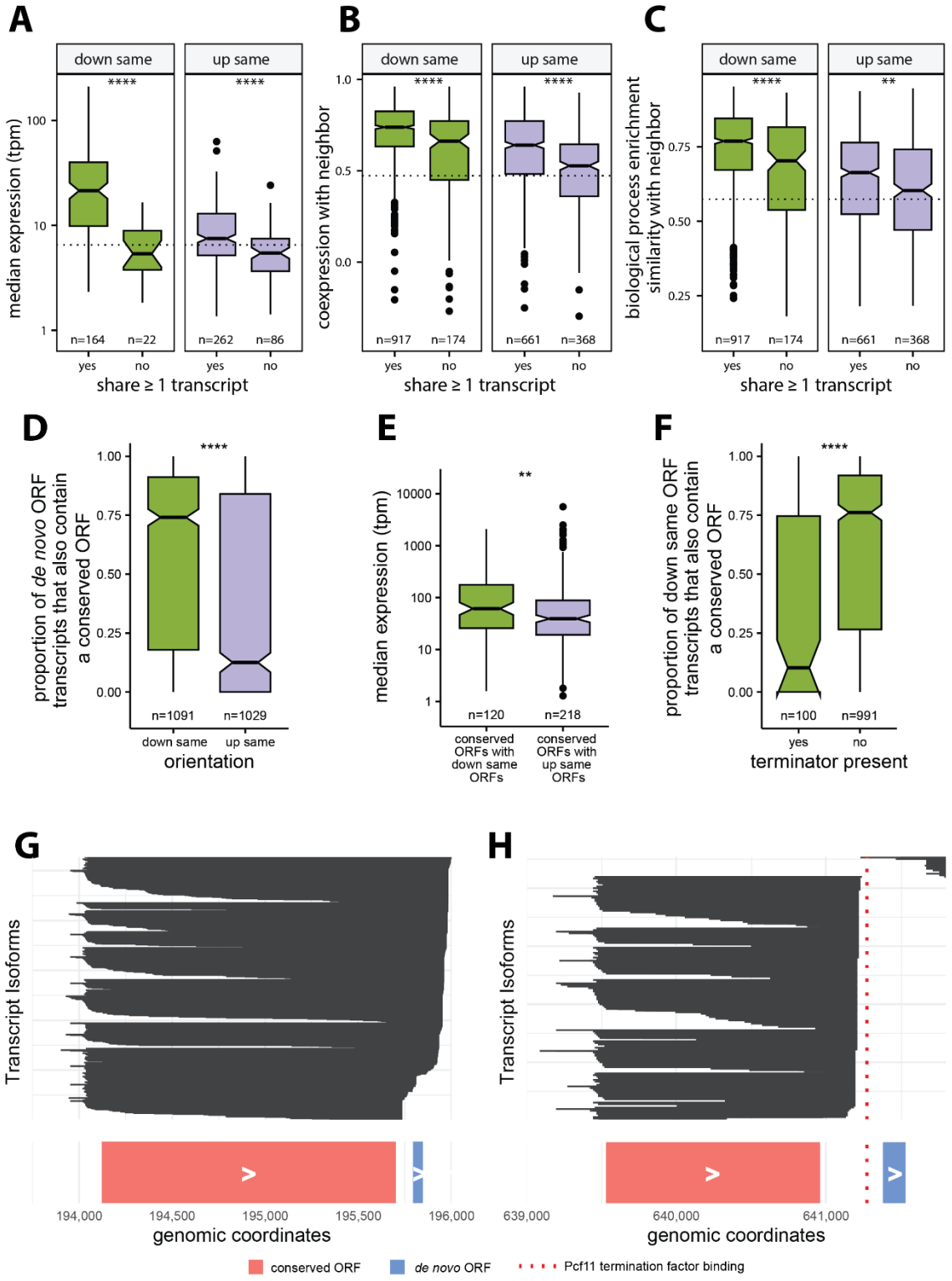


**Figure 5 Down same *de novo* ORFs are positioned advantageously to share transcripts with neighboring conserved ORFs**

A) *De novo* ORFs that share a transcript with neighboring conserved ORFs, as determined by TIF-seq transcript boundaries, have significantly higher expression levels than *de novo* ORFs that do not. Dashed line represents the median expression of independent *de novo* ORFs. Down same and up same *de novo* ORFs are compared. B) *De novo* ORFs that share a transcript with neighboring conserved ORFs have higher coexpression with their neighbors than *de novo* ORFs that do not share a transcript. Down same and up same *de novo* ORFs are compared. Dashed line represents median coexpression of *de novo*-conserved ORF pairs on separate chromosomes. C) *De novo* ORFs that share a transcript have more similar functional enrichments with neighboring conserved ORFs than *de novo* ORFs that do not share a transcript. Down same and up same *de novo* ORFs are compared. Dashed line represents median functional enrichment similarities of the background distribution of *de novo*-conserved ORF pairs on separate chromosomes. D) Down same *de novo* ORFs share a transcript with neighboring conserved ORFs more often than up same ORFs. E) Conserved ORFs with downstream *de novo* ORFs have a small but significant increase in expression compared to conserved ORFs with upstream *de novo* ORFs. F) Loci where transcription termination factors Pcf11 and Nrd1 are not present between a conserved ORF and neighboring down same *de novo* ORF are more likely to share a transcript than loci where termination factors are present. G) Transcript isoforms (black) at an example locus where there are no transcription termination factors present between conserved ORF YBL015W (pink) and downstream *de novo* ORF chr2_195794 (blue). H) Transcript isoforms (black) at an example locus where there is Pcf11 transcription terminator present (red line) between conserved ORF YIL090W (pink) and downstream *de novo* ORF chr9_195520 (blue).
(****: $p \leq 0.0001$, ***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$, ns: not-significant; Mann-Whitney U-test)

*De novo* ORFs that are located downstream of conserved ORFs may be in an advantageous location, as they exhibit increased expression, coexpression, and functional enrichment with neighboring conserved ORF. We hypothesized that the molecular mechanism resulting in these increases is transcriptional read-through which leads to *de novo* ORFs sharing a transcript with their neighboring conserved ORFs. To test this hypothesis, we analyzed publicly available transcript isoform sequencing (TIF-seq) data[61]. Of the ORF pairs that were detected using TIF-seq, we found that 84% of down same and 64% of up same *de novo* ORFs share at least one transcript with their neighboring conserved ORFs. For both orientations, *de novo* ORFs that share at least one transcript with neighboring conserved ORFs have a significantly higher median expression level compared to *de novo* ORFs that do not (down same Cliff's Delta d = 0.75, p = 1.06e-8, up same: Cliff's Delta d = 0.38, p = 1.23e-7; Mann-Whitney U-test, Figure 5A). We also observed a significant increase in coexpression and biological process enrichment similarity between *de novo* ORFs and their neighboring conserved ORFs when they are found on the same transcript at least once compared to when they are not (coexpression: down same: Cliff's Delta d = 0.28, Mann-Whitney U-test, p = 2.99e-9; up same: Cliff's Delta d = 0.31, Mann-Whitney U-test, p < 2.2e-16; BP enrichment similarity: down same: Cliff's Delta d = 0.21, Mann-Whitney U-test, p = 1.49e-5; up same: Cliff's Delta d = 0.108, Mann-Whitney U-test, p = 3.78e-3, Figures 5B and 5C, respectively). While sharing a transcript led to increases for both up same

and down same orientations, down same ORFs that share a transcript have much higher expression, as well as coexpression and biological process enrichment similarity with their conserved neighbor, compared to up same ORFs that share a transcript. This could be due to down same ORFs' tendency to share transcripts more often than up same ORFs (Cliff's Delta d = 0.26, Mann-Whitney U-test $p < 2.2e-16$; Figure 5D) and the slightly higher expression of conserved ORFs with down same ORFs on their transcripts than conserved ORFs with up same ORFs on their transcripts (Cliff's Delta d = 0.19, Mann-Whitney U-test, p = 4.29e-3; Figure 5E).

Additionally, we examined the impact of transcription terminators Pcf11 or Nrd1 on the frequency of transcript sharing between a conserved ORF and its downstream *de novo* ORF. Analyzing publicly available ChiP-exo data[52], we found that ORF pairs lacking transcriptional terminators had a notably higher percentage of shared transcripts than those with a transcriptional terminator (Cliff's Delta d = 0.39, Mann-Whitney U-test, p = 1.591e-10, Figure 5F). Therefore, we conclude that sharing a transcript via transcriptional readthrough is a major transcriptional piggybacking mechanism for down same *de novo* ORFs.

As an illustration, consider the genomic region on chromosome II from bases 194,000 to 196,000, containing the conserved ORF YBL015W and a downstream *de novo* ORF (positions 195,794 to 195,847). No terminator factor is bound to the intervening DNA between these two ORFs. This pair has high coexpression, with $\rho = 0.96$ and we observed that nearly all transcripts in this region containing the *de novo* ORF also include YBL015W (Figure 5G). In contrast, the genomic region on chromosome XVI from 639,000 to 641,800, containing the conserved ORF YPR034W and downstream *de novo* ORF (positions 641,385 to 641,534), does have a Pcf11 terminator factor between the pair, and as expected, none of the transcripts in this region contain both YPR034W and the *de novo* ORF, which have poor coexpression as a result ($\rho = 0.1$, Figure 5H)

# Discussion

In this study we built a high quality expanded coexpression network including both cORFs and nORFs in *S. cerevisiae* by integrating thousands of publicly available RNA-seq samples. Our goal was to analyze the transcription profiles of nORFs, whose low expression creates statistical issues. We utilized a dedicated statistical approach which enabled us to uncover expression and coexpression patterns for thousands of nORFs despite their low expression. While previous studies have used coexpression to investigate nORFs and their cellular roles in various species[48–50], our study represents a significant technical advancement in that it is the first to combine thousands of RNA-seq samples with computational methods that account for sparsity and low expression levels when calculating coexpression[55,56].

We explored the transcription of nORFs from multiple angles including analyzing network topology, conducting evolutionary analyses, investigating associations with cellular processes, and examining the influence of genomic orientation on expression. Delving into network topology, we find that nORFs have distinct expression profiles that are correlated with only a few other ORFs. Nearly all cORFs are coexpressed with at least one nORF, but the converse is not true. Numerous nORFs form new structured transcriptional modules, possibly involved in both known and unknown cellular processes.

Next, we investigated the evolutionary origins of all ORFs in the expanded network. Similar to previous reports, we found that many nORFs have evolved *de novo* from previously non-genic regions[16]. We leveraged the expanded coexpression network to generate hypotheses about the potential cellular roles of *de novo* ORFs. We discovered that half of *de novo* ORFs tend to show higher coexpression with cORFs that are involved in homeostasis and transport, which could mean they are also involved in such processes. While future studies will be needed to test these hypotheses since nORFs are entirely uncharacterized, there are several consistent observations in the literature[18,38,62]. For instance, Li et al.[18] showed that many *de novo* ORFs are upregulated in heat shock. Wilson and Masel[63] found higher translation of *de novo* ORFs under starvation conditions. Carvunis et al.[16] found *de novo* cORFs are enriched for the GO term ‘response to stress’. Other studies showed examples of how *de novo* ORFs could be involved in stress response[29,64] or homeostasis[64,65]. For instance the *de novo* antifreeze glycoprotein AFGP allows Arctic codfish to live in colder environments[29] or *MDF1* in yeast, provides resistance to certain toxins and mediates ion homeostasis[66]. When combined with these previous investigations, our results provide further evidence that *de novo* ORFs may provide adaptation to environmental stresses and help maintain homeostasis, perhaps through modulation of transport processes.

Recent research in yeast has revealed a significant enrichment of transmembrane domains[16,22] (TMDs) within putative peptides of *de novo* ORFs, suggesting an association with cellular membranes. Notably, many studies identified *de novo* ORFs in yeast[22,59] and small nORFs in humans[14,67] that localize to diverse cellular membranes, such as ER, vacuole, endosome, or mitochondria. These findings have raised the possibility that *de novo* ORFs could play a role in

a range of transport processes, such as ion, amino acid, and protein transport across cellular membranes. Despite these observations, the precise functional relationship between *de novo* ORFs and transport processes remains unclear. In this context, our study provides the first evidence of a large-scale association between the expression of *de novo* ORFs and cellular transport. However, given the complex nature of transport processes, further experimental validations are warranted to elucidate the underlying molecular mechanisms that may be involved. Nevertheless, our results underscore the potential importance of *de novo* ORFs in cellular transport processes, and could pave the way for future investigations into their functional roles in this context.

Lastly, we conducted a comparative analysis to examine how different genomic orientations of *de novo* ORFs could affect their expression and coexpression. We found that *de novo* ORFs located downstream on the same strand as conserved ORFs exhibit large increases in expression, coexpression, and biological process similarity with their neighboring conserved ORFs compared to ORFs in other orientations. The underlying mechanism that facilitates this increase is transcriptional readthrough leading to *de novo* ORFs sharing a transcript with their neighboring conserved ORF. These findings suggest that certain genomic regions may provide a more favorable environment for the transcription of *de novo* ORFs. A previous study in humans showed that readthrough transcription downstream of some genes is responsible for roughly 15%–30% of intergenic transcription and is induced by osmotic and heat stress creating extended transcripts with chromatin localization that play a role in maintaining nuclear stability during stress[68]. Another study in humans and zebrafish showed that the translation of small ORFs located in the 3' UTR of mRNAs (dORFs) increased the translation rate of the upstream gene[69]. These examples suggest that the transcription of regions downstream of conserved ORFs is functional and regulated.

Our study could change our understanding of how the down same *de novo* ORFs gain cellular roles. When an ORF emerges downstream of a conserved ORF, it is more likely that RNA polymerase will continue transcribing over the length of the ORF, generating transcripts that contain both the *de novo* ORF and the conserved ORF, and perhaps in turn facilitate translation of the *de novo* ORF. This is particularly true when transcription terminators are absent, allowing for uninterrupted transcription. This transcription, which is presumably under the regulation of the conserved ORF, creates a pool of transcripts that evolution can select for or against. The likelihood of a *de novo* ORF being expressed or repressed under the same conditions as the neighboring ORF is largely determined by the extent to which it piggybacks on the neighboring ORF's transcription. Therefore, in addition to the evolutionary pressure acting on the sequence of emerging ORFs, our results suggest that transcriptional regulation and genomic context also play crucial roles in determining their functional potential.

# Methods

## Creating ORF list

To create our initial ORF list, we utilized two sources. First, we took annotated ORFs in the *S. cerevisiae* genome R64-2-1 from SGD on April 23, 2020, which included 6,600 ORFs. Second, we utilized the translated ORF list from Wacholder et al.[7] reported in their *Supplementary Table 3.* We filtered to include canonical ORFs (Verified, Uncharacterized or Transposable element genes) as well as any noncanonical ORFs with evidence of translation at q value < 0.05 (Dubious, Pseudogenes and unannotated ORFs). We removed ORFs with lengths shorter than the alignment index kmer size of 25nt used for RNA-seq alignment. In situations where ORFs overlapped on the same strand with greater than 75% overlap of either ORF, we removed the shorter ORF. We removed ORFs that were exact sequence duplicates of another ORF. This left 5,878 canonical and 18,636 noncanonical ORFs, for a total of 24,514 ORFs used for RNA-seq alignment.

## RNA-seq data preprocessing

Strand specific RNA-seq samples were obtained from the Sequencing Read Archive (SRA) using the search query *(saccharomyces cerevisiae[Organism]) AND rna sequencing*. Each study was manually inspected and only studies that had an accompanying paper or detailed methods on Gene Expression Omnibus (GEO) were included. Samples were aligned to the ORF list explained above and quantified using Salmon[70] version 0.12.0 and an index kmer size of 25. Samples with less than 1 million reads mapped or unstranded samples were removed, resulting in an expression dataset of 3,916 samples from 174 studies. ORFs were removed to limit sparsity and increase the number of observations in the subsequent pairwise coexpression analysis. Only ORFs that had at least 400 samples with a raw count > 5 were included for downstream coexpression analysis, n = 11,630 ORFs (5,803 canonical and 5,827 noncanonical).

## Coexpression calculations

The raw counts were transformed using centered log ratio (clr). Pairwise proportionality was calculated using $\rho$[55] for each ORF pair and only the pairs with at least 400 observations were included, i.e., ORFs that had at least 400 samples with greater than 5 raw counts for both ORFs. Spatial quantile normalization (SpQN)[56] of the coexpression network was performed using the mean clr expressions values for each ORF as confounders to correct for mean expression bias, which resulted in similar distributions of coexpression values across varying expression levels.

Since zero values cannot be used with log ratio transformations, all zeros must be removed from the dataset. Proposed solutions in the literature on how to remove zeros, all of which have

their cons, include removing all genes that contain any zeros, imputing the zeros, or adding a pseudo count to all genes[71,72]. Removing all ORFs that contain any zeros is not possible for this analysis since the ORFs of interest are lowly and conditionally expressed. The addition of pseudocounts can be problematic when dealing with lowly expressed ORFs, for the addition of a small count is much more substantial for an ORF with a read count of 5 compared to an ORF with a read count of 100[73]. For these reasons, all raw counts below 5 were set to NA prior to clr transformation. These observations are then excluded when calculating the clr transformation and in the ρ calculations.

We used clr and ρ implementations in R package *Propr*[55] and implementation of SpQN from Wang et al. (2022)[56].

# Sample thresholding

To determine the minimum number of samples needed expressing both ORFs in a pair we determined the number of samples needs for coexpression values to converge within $\rho \pm 0.05$ or $\rho \pm 0.1$ for 2,167 nORF-cORF pairs which have a $\rho$ > 99th percentile (before SpQN). All samples expressing both ORFs in a pair were randomly binned into groups of 10, and ρ was calculated after each addition of another sample. Fluctuations were calculated as max(ρ)-min(ρ) within a sample bin. Convergence was determined as the first sample bin with fluctuations ≤ fluctuation threshold, either 0.05 or 0.01.

# Transcription factor binding enrichments

A ChIP-exo dataset from Rossi et al.[52] containing DNA-binding information for 73 sequence-specific transcription factors (TFs) across the whole genome was used. For each ORF we identified which TFs had binding within 200 bp upstream of the ORF's transcription start site (TSS).

The transcription start site (TSS) for all ORFs in the coexpression matrix was determined by the median 5' TIF start using TIF-seq[61] dataset. If no transcript containing the ORF was found in the TIF-seq data, then the median distance from ATG to TSS for all other ORFs is used to infer the TSS for the given ORF.

To calculate enrichments, the coexpression matrix was first filtered to only include ORFs that have at least 1 TF binding within 200 bp upstream of its TSS (n = 1,909). Fisher's exact test was used to calculate association between coexpression and sharing a TF. Coexpressed was defined as $\rho > 0.888$.

# Protein Complex enrichments

A manually curated list of 408 protein complexes in *S. cerevisiae* was retrieved from the CYC2008 database by Pu et al[51]. The coexpression matrix was subsetted to contain only the 1,617 cORFs found in the CYC2008 database prior to creating the contingency table. Coexpressed was defined as $\rho > 0.888$. Fisher's exact test was used to calculate the significance of association between coexpression and protein complex formation.

## Coexpression matrix clustering

We used a package called weighted gene coexpression network analysis (*WGCNA*)[57] in R to cluster our coexpression matrix. To do this, we first transformed our coexpression matrix into a weighted adjacency matrix by applying a process called soft thresholding. Soft thresholding involved raising the coexpression matrix to the power of 12, which removed weak coexpression relationships from the matrix. We then used the topological overlap matrix (TOM) similarity to calculate the distances between each column and row of the matrix. Using the *hclust* function in R with the *ward* clustering method, we created a hierarchical clustering dendrogram. We then used the dynamic tree cutting method within the *WGCNA* package to assign ORFs to coexpression clusters, resulting in 73 clusters of which 69 were mapped to the expanded coexpression network.

## Gene ontology analysis of clusters

GO trees (file: go-basic.obo) and annotations (files: sgd.gaf) were downloaded from http://geneontology.org/ on March 10, 2022. We used the Python package, GOATools[74], to calculate the number of genes associated with each GO term in a cluster and the overall population of (all) genes in the coexpression matrix. We excluded annotations based on the evidence codes ND (no biological data available). We identified GO term enrichments by calculating the likelihood of the ratio of the cORFs associated with a GO term within a cluster given the total number of cORFs associated with the same GO term in the background set of all cORFs in the coexpression matrix. We applied Fisher's exact test and Benjamini-Hochberg false discovery rate (FDR)[75] multiple testing correction to calculate corrected p-values for the enrichment of GO term in the clusters. FDR < 0.05 was taken as a requirement for significance. We applied GO enrichment calculations only when there were at least 5 canonical ORFs in the cluster.

## Network analyses

To create random networks while preserving the same degree distribution, we used an edge swapping method (Supplementary Figure 4). This involved randomly selecting two edges in the network, which were either cORF-nORF or nORF-nORF edges, and swapping them. The swap was accepted only if it did not disconnect any nodes from the network and the newly generated edges were not already present in the network. We repeated this process for at least ten times the number of edges in the network. Network diameter and transitivity were calculated using R package *igraph*[76] and networks were plotted using spring embedded layout in Python package *networkx*[77].

## Gene set enrichment analysis

Gene set enrichment analysis (GSEA) calculates gene ontology (GO) enrichments of an ordered list of genes. For each ORF in our dataset, we used $\rho$ values to order annotated ORFs and provided this sorted set to GSEA software. We used the GO slim file downloaded from

*Saccharomyces* Genome Database (SGD) on 20 January 2021 for GO annotations and *fgsea*[78] R package to calculate enrichments. To calculate GO terms that are enriched or depleted for *de novo* ORFs compared to conserved ORFs, we calculated the number of conserved and *de novo* ORFs that had GSEA enrichments at FDR < 0.01. Using these counts we calculated the proportion of *de novo* and conserved ORFs associated with a GO term and used Fisher's exact test to assess the significance of association. P values returned by Fisher's exact test were corrected for multiple hypothesis testing using Benjamini-Hochberg FDR correction. Odds ratios were calculated by dividing proportion of *de novo* ORFs to proportion of conserved ORFs. Proportions for the GO terms with FDR < 0.001 and Odds ratio greater than 2 or less than 0.5 are plotted Figure 3C-D and are reported in Supplementary Data 4.

# Calculation of GO term similarities

GO term similarities were calculated using the Relevance method developed in Schlicker et al.[60] This method considers both the information content (IC) of the GO terms that are being compared and the IC of their most informative ancestor. IC represents the frequency of a GO term; thus, an ancestral GO term has lower IC than a descendant. We used the *GOSemSim*[79] package in R that implements these similarity measures.

# Detection of homologs using BLAST

We obtained the genomes of 332 budding yeasts from Shen et al. 2018[80]. To investigate the homology of each non overlapping ORF in our dataset, we used TBLASTN and BLASTP[81] against each genome in the dataset, excluding the *Saccharomyces* genus. Default settings were used, with an e-value threshold of 0.0001. The BLASTP analysis was run against the list of protein coding genes used in Shen et al. 2018, while the TBLASTN analysis was run against each entire genome. We also applied BLASTP to annotated ORFs within the *S. cerevisiae* genome to identify homology that could be caused by whole genome duplication or transposons.

# Identification of *de novo* and conserved ORFs

To identify *de novo* ORFs, we applied several strict criteria. Firstly, we obtained translation q-values and reading frame conservation (RFC) data from Wacholder et al.[7] All cORFs and only nORFs with a translation q-value less than 0.05 were considered as potential *de novo* candidates. We excluded ORFs that overlapped with another cORF on the same strand or had TBLASTN or BLASTP hits outside of the *Saccharomyces* genus at e-value < 0.0001. Moreover, we eliminated ORFs that had BLASTP hits to another canonical ORF in *S. cerevisiae*. From the remaining list of candidate *de novo* ORFs, we investigated whether their ancestral sequence could be noncoding. To do this, we utilized RFC values for each species within *Saccharomyces* genus. We classified ORFs as *de novo* if the RFC values for the most distant two species were

less than 0.6, suggesting the absence of a homologous ORF in those two species (Supplementary Figure 9).

We identified conserved ORFs if a nonoverlapping cORF has an average RFC > .8 or has either TBLASTN or BLASTP hit at e-value < 0.0001 threshold.

To identify conserved cORFs with overlaps we first considered if the cORFs had a BLASTP outside of *Saccharomyces* genus. Then for two overlapping ORFs, if one has RFC > 0.8 and the other has RFC < 0.8, we considered the one with higher RFC as conserved. For the ORF pairs that were not assigned as conserved using these two criteria, we applied TBLASTN for the non-overlapping parts of the overlapping pairs. Those with a TBLASTN hit with e-value < 0.0001 were considered conserved. We found a total of 5,624 conserved ORFs and 2,756 *de novo* ORFs.

## Termination factor binding analysis

ChIP-exo data for Pcf11 and Nrd1 termination factor binding sites are taken from Rossi et al.[52] This study reports binding sites at base pair resolution for *S. cerevisiae* for around 400 proteins. We used supplementary bed formatted files for Pcf11 and Nrd1, which are known transcriptional terminators, and used in house R scripts to find binding sites within the regions between the stop codon of conserved ORFs and the start codon of down same *de novo* ORFs. ORF pairs were classified as having terminators present between them if there was either Pcf11 or Nrd1 binding.

## Shared transcript isoforms

To determine whether two ORFs shared transcripts, we reused the TIF-Seq dataset from Pelechano et al.[61] TIF-Seq is a sequencing method that detects the boundaries of transcript isoforms (TIFs). We extracted all reported TIFs from the supplementary data file S1 and identified all TIFs that fully cover each ORF in both YPD and galactose. We then used this information to find ORF pairs that mapped to the same TIFs. ORF pairs where the conserved ORF was not found in the TIF-seq dataset were not included and pairs where the *de novo* ORF was not found were considered to be not sharing a transcript.

# Acknowledgments

Figures 1A, 3A, 4A, and supplementary Figure 6 were created with BioRender.com. The authors are grateful to Dr. Aaron Wacholder, Carly Houghton, Nelson Coelho, Dr. Saurin Bipin Parikh, Jiwon Lee, Lin Chou, and Alistair Turcan for reviewing the manuscript prior to submission.

# Author Contributions

Conceptualization: A.R., O.A., and A.-R.C.; Methodology: A.R, O.A.; Investigation: A.R, O.A.; Writing-original draft: A.R, O.A.; Writing-review and editing: A.R., O.A., and A.-R.C.; Supervision: A.-R.C.

# Funding

This work was supported by: the National Science Foundation Graduate Research Fellowship under Grant No. 2144349 awarded to A.-R.C and the National Science Foundation Graduate Research Fellowship under Grant No. 2139321 awarded to A.R. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

# Competing interests

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

# Source code

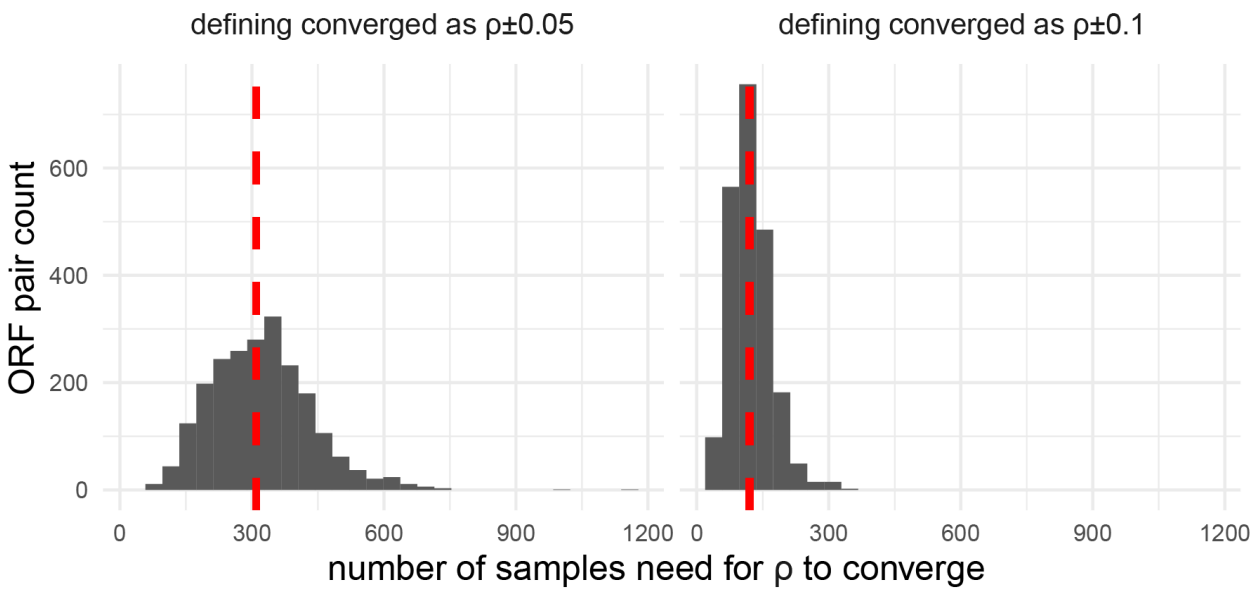
All source codes for the analyses conducted are accessible online at https://www.github.com/oacar/noncanonical_coexpression_network

# Supplementary Data

Supplementary data files are available on Figshare https://doi.org/10.6084/m9.figshare.22289614

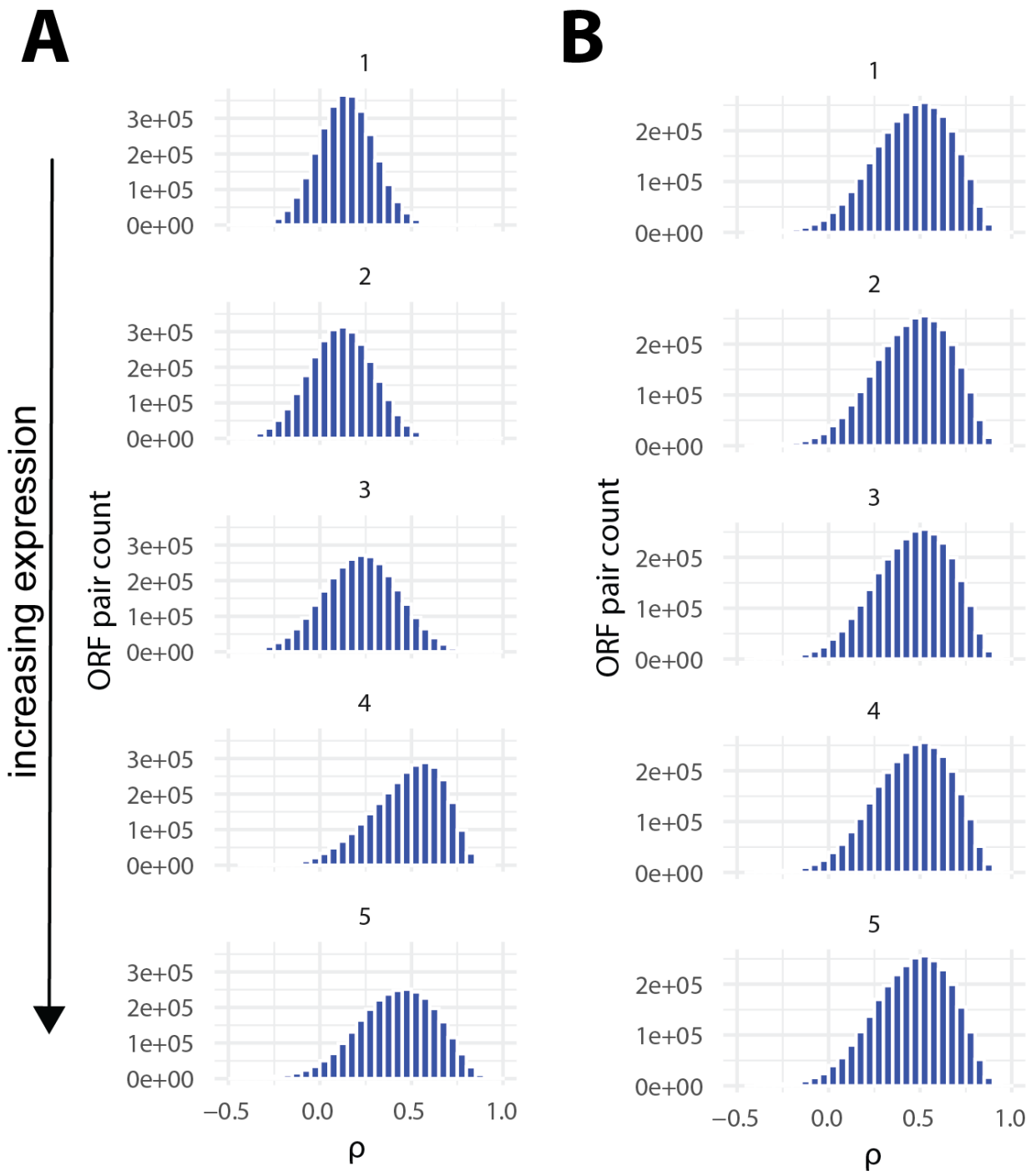
# Supplementary Figures

## Supplementary Figure 1



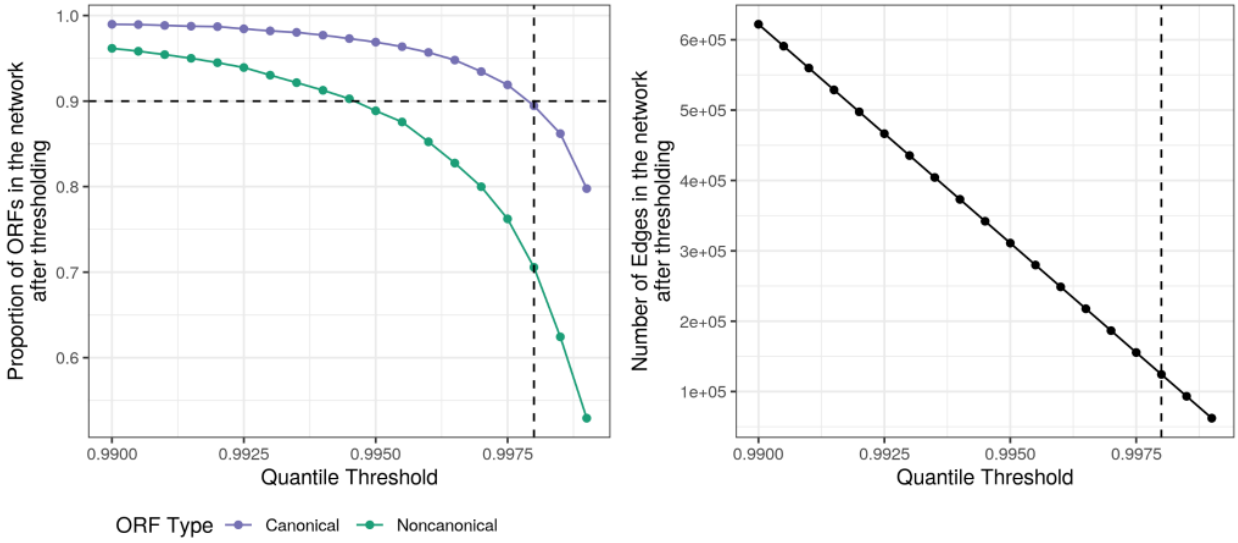

Supplementary Figure 1 To understand the effect of sample size on coexpression values and to determine how many samples is sufficient for ρ to converge, we recalculated coexpression for a given ORF pair using n = 2 samples through n = all samples. Fluctuations were calculated as max(ρ)-min(ρ) within bins of 10 samples. The number of samples needed for ρ to converge was calculated as the first sample bin where ρ fluctuations ≤ fluctuation threshold, either 0.1 or 0.05. Histogram showing the minimum number of samples needed for ρ values to converge within ρ ± 0.05 (left) and ρ ± 0.1 (right) for 2,167 cORF-nORF pairs with ρ > 99th percentile. Red dashed lines show the median number of samples needed.

# Supplementary Figure 2



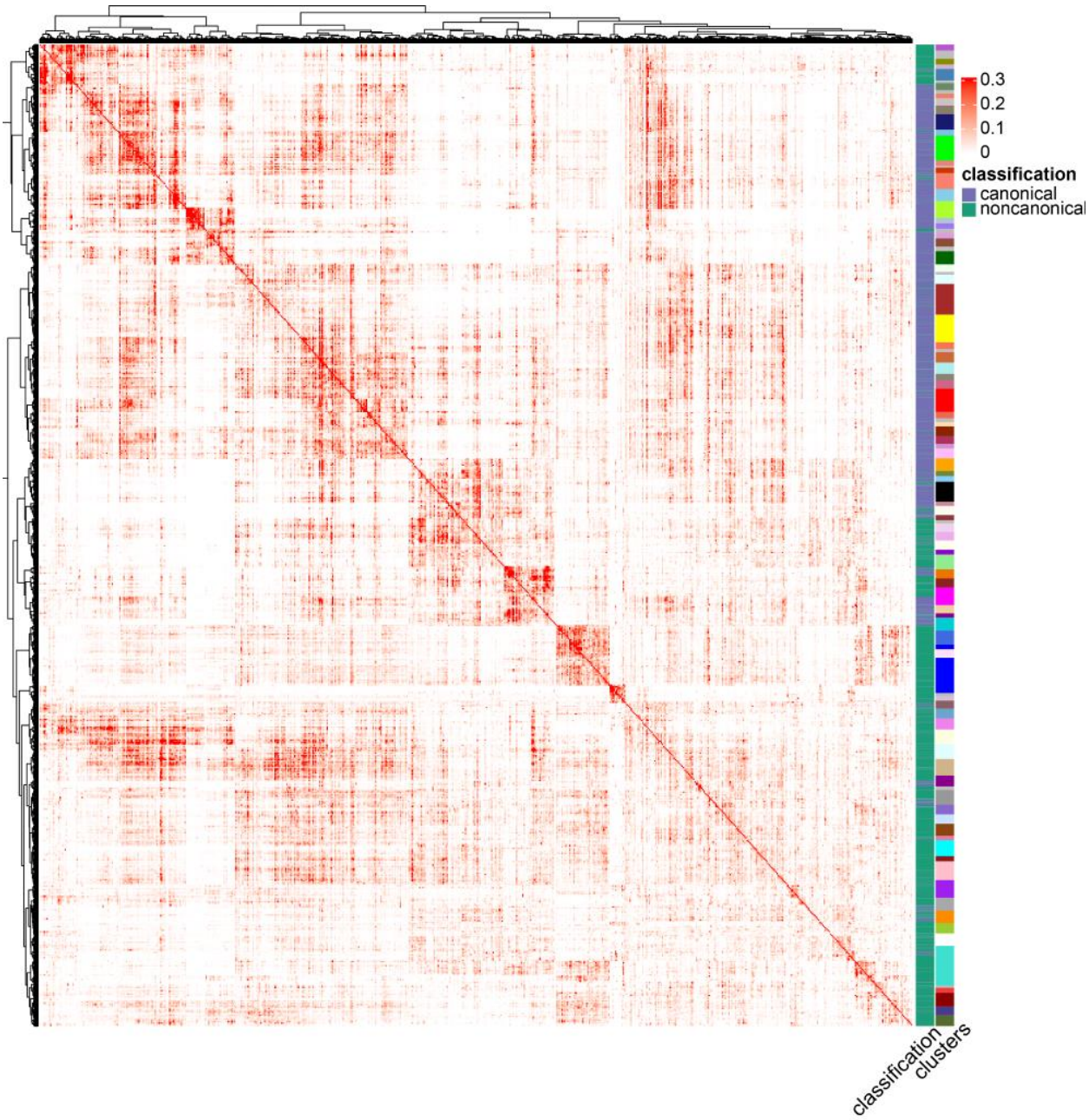

Supplementary Figure 2 Distribution of coexpression values (ρ) for ORF pairs binned by expression level, from lowly expressed pairs *top* to highly expressed pairs *bottom*, A) before spatial quantile normalization (SpQN) and B) after SpQN, which normalizes the coexpression values so that the distribution within each expression bin is similar.

# Supplementary Figure 3



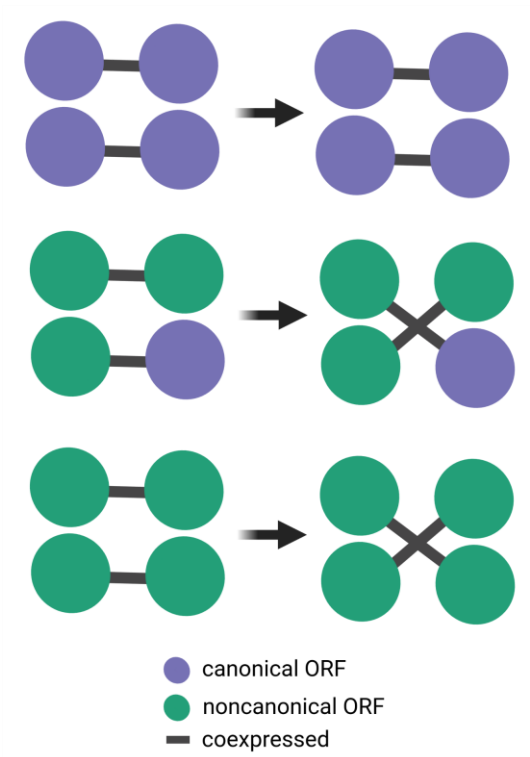

Supplementary Figure 3 Network threshold affects cORFs and nORFs differently. *Left* shows the proportion of cORFs or nORFs in the network at each quantile threshold and the right shows the number of connections in the network. Dashed line represents 0.9998 quantile which was chosen for creating the network.

# Supplementary Figure 4



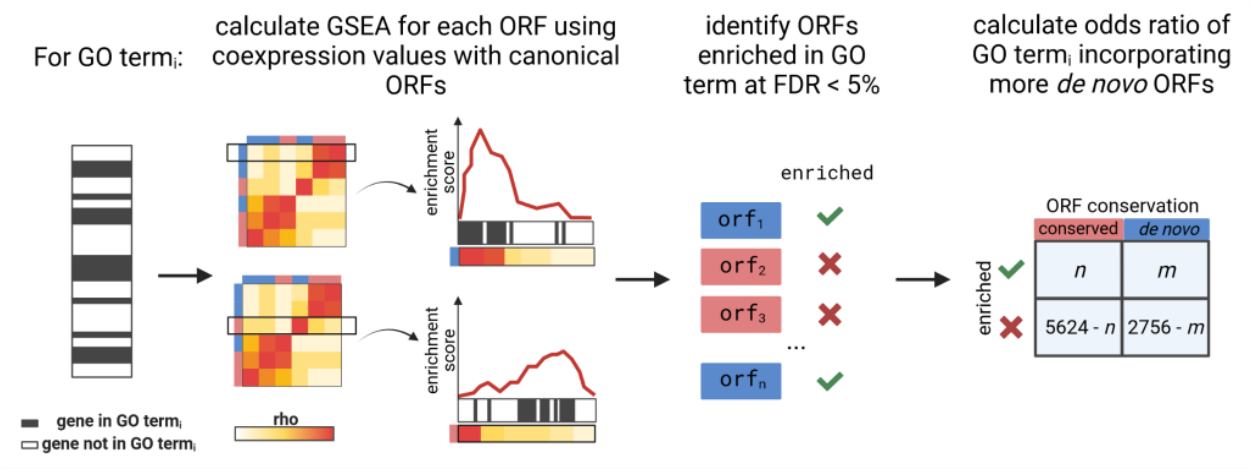

Supplementary Figure 4 Clustered matrix heatmap. Coexpression values are first transformed by taking power of 12 and then WGCNA pipeline is applied. Clusters are determined by cutting dendrograms. Colors on 'clusters' section represent the different clusters. Values of 0.3 and above are represented by red to show the structure of the heatmap.

# Supplementary Figure 5



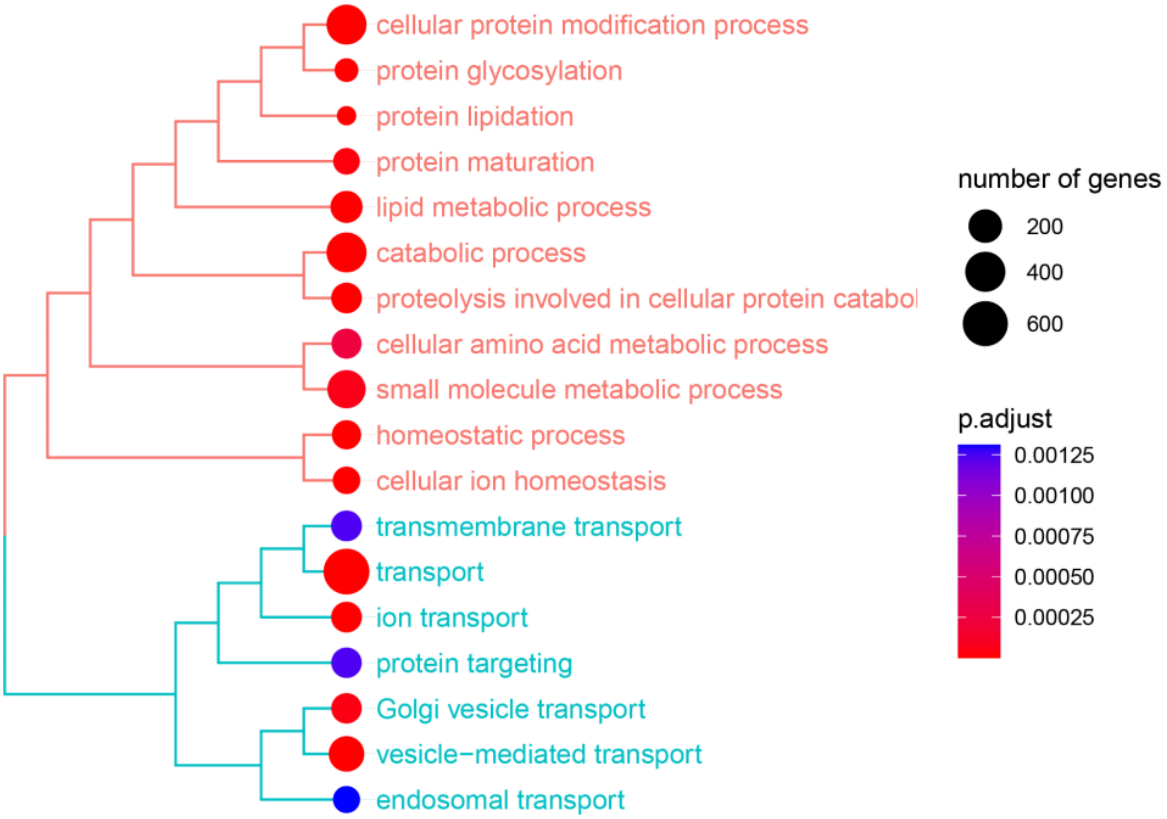

Supplementary Figure 5 Schema for generating randomized networks. Edges between cORF-nORF and nORF-nORF pairs were swapped in a pairwise manner such that the degree of each node stayed the same. Edges between cORF-cORF pairs were not randomized.

# Supplementary Figure 6



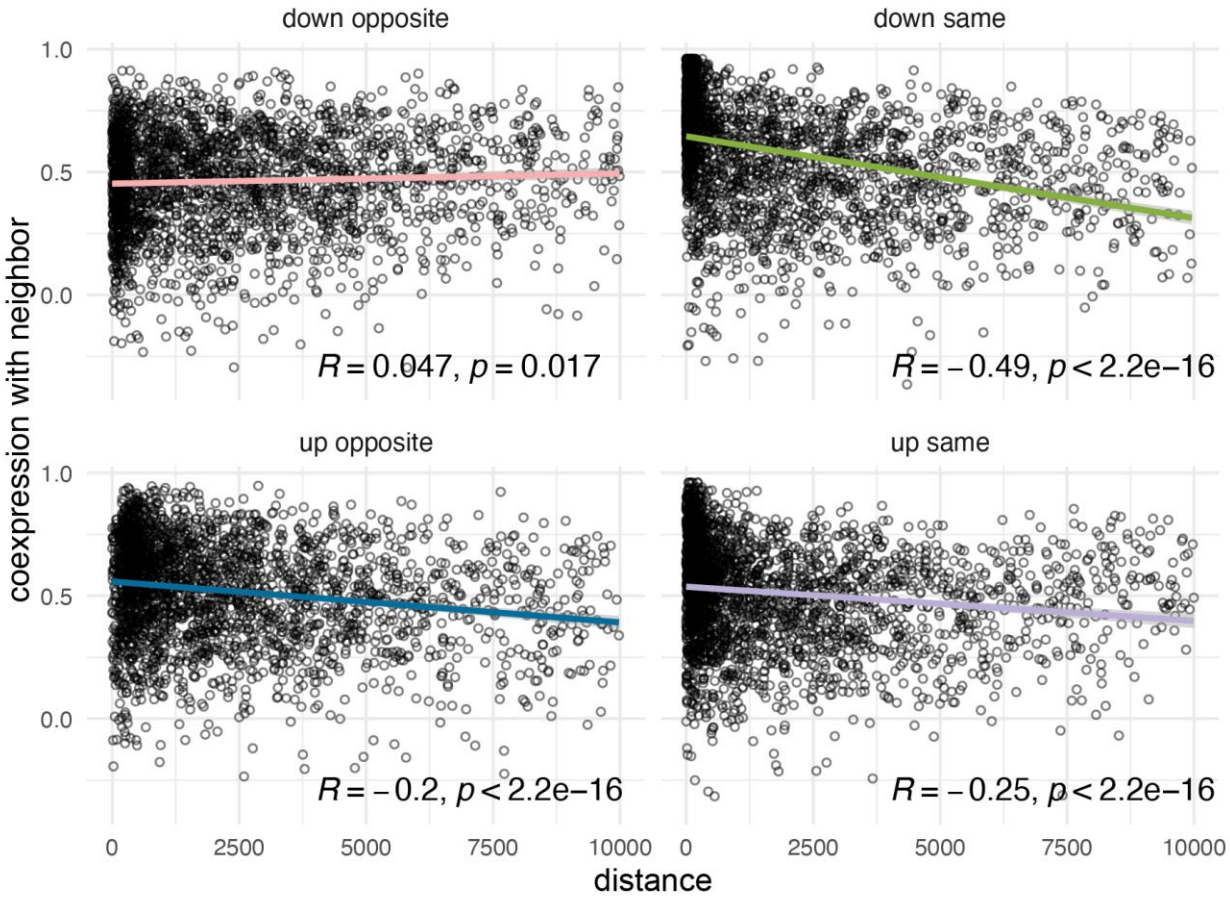

Supplementary Figure 6 Gene set enrichment analysis (GSEA) pipeline using coexpression profiles to find gene ontology terms that are more likely to incorporate *de novo* ORFs.

# Supplementary Figure 7



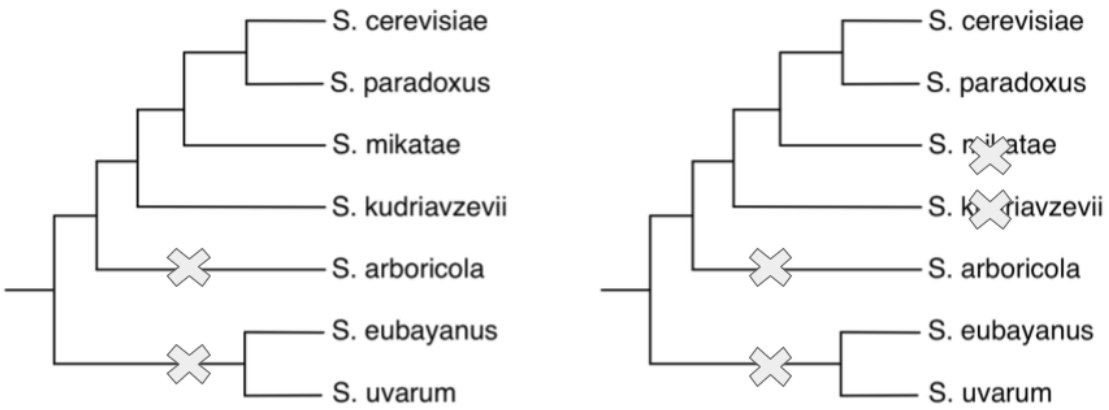

Supplementary Figure 7 The 20 most significant GSEA enrichments of YBR196C-A. Dendrogram was calculated using semantic similarities between GO terms. Node sizes correspond to the number of enriched genes. Node colors correspond to Benjamini-Hochberg corrected p-values (FDR).

# Supplementary Figure 8




Supplementary Figure 8 Coexpression level of *de novo* ORFs with neighboring conserved ORF is influenced by distance and orientation. There is a negative correlation between distance and coexpression for *de novo* ORFs located in the down same, up same and up opposite orientations (down opposite: R = 0.047, p = 0.017; down same: R = -0.49, p < 2.2e-16; up opposite: R = -0.2, p < 2.2e-16; up same: R = -0.25, p < 2.2e-16; Spearman's correlation coefficient).

# Supplementary Figure 9




Supplementary Figure 9 Determination of *de novo* ORFs using parsimony. Using *Saccharomyces* species, if an ORF is lost (represented by a gray 'X' in the plot) in the two of the most distant species and lacks BLAST hits, it is classified as *de novo*.

# References


1. Basrai, M. A., Hieter, P. & Boeke, J. D. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Res.* **7**, 768–771 (1997).
2. Dujon, B. The yeast genome project: what did we learn? *Trends Genet. TIG* **12**, 263–270 (1996).
3. Fisk, D. G. *et al.* Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. *Yeast Chichester Engl.* **23**, 857–865 (2006).
4. Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344–1349 (2008).
5. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
6. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).
7. Wacholder, A. *et al.* A vast evolutionarily transient translatome contributes to phenotype and fitness. 2021.07.17.452746 Preprint at https://doi.org/10.1101/2021.07.17.452746 (2023).
8. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
9. Aspden, J. L. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**, e03528 (2014).
10. Crappé, J. *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **43**, e29 (2015).
11. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147**, 789–802 (2011).

12. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
13. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
14. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
15. Wright, B. W., Yi, Z., Weissman, J. S. & Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* (2021) doi:10.1016/j.tcb.2021.10.010.
16. Carvunis, A.-R. *et al.* Proto-genes and *de novo* gene birth. *Nature* **487**, 370–374 (2012).
17. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genet.* **11**, e1005721 (2015).
18. Li, J., Singh, U., Arendsee, Z. & Wurtele, E. S. Landscape of the Dark Transcriptome Revealed Through Re-mining Massive RNA-Seq Data. *Front. Genet.* **12**, (2021).
19. O'Meara, T. R. & O'Meara, M. J. DeORFanizing Candida albicans Genes using Coexpression. *mSphere* **6**, e01245-20 (2021).
20. Housman, G. & Ulitsky, I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1859**, 31–40 (2016).
21. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
22. Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
23. Arnoult, N. *et al.* Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. *Nature* **549**, 548–552 (2017).

24. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**, 595–606 (2015).
25. Jackson, R. *et al.* The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434–438 (2018).
26. Van Oss, S. B. & Carvunis, A.-R. De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019).
27. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
28. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**, 769–772 (2014).
29. Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci.* **116**, 4400–4405 (2019).
30. Vakirlis, N., Vance, Z., Duggan, K. M. & McLysaght, A. De novo birth of functional microproteins in the human lineage. *Cell Rep.* **41**, 111808 (2022).
31. Chen, J.-Y. *et al.* Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. *PLOS Genet.* **11**, e1005391 (2015).
32. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).
33. Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* **5**, e09977 (2016).
34. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
35. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10**, 1106–1113 (2008).

36. Ghanbarian, A. T. & Hurst, L. D. Neighboring Genes Show Correlated Evolution in Gene Expression. *Mol. Biol. Evol.* **32**, 1748–1766 (2015).

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

38. Blevins, W. R. *et al.* Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat. Commun.* **12**, 604 (2021).

39. Kim, S. K. *et al.* A Gene Expression Map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).

40. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**, 249–255 (2003).

41. Niu, X. *et al.* Weighted Gene Co-Expression Network Analysis Identifies Critical Genes in the Development of Heart Failure After Acute Myocardial Infarction. *Front. Genet.* **10**, (2019).

42. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5**, 3231 (2014).

43. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).

44. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).

45. Lee, J., Shah, M., Ballouz, S., Crow, M. & Gillis, J. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Res.* **48**, W566–W571 (2020).

46. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **19**, 575–592 (2018).

47. Yin, W., Mendoza, L., Monzon-Sandoval, J., Urrutia, A. O. & Gutierrez, H. Emergence of co-expression in gene regulatory networks. *PLOS ONE* **16**, e0247671 (2021).

48. Stiens, J. *et al.* Using a Whole Genome Co-expression Network to Inform the Functional Characterisation of Predicted Genomic Elements from Mycobacterium tuberculosis Transcriptomic Data. 2022.06.22.497203 Preprint at https://doi.org/10.1101/2022.06.22.497203 (2022).
49. Bashir, K. *et al.* Transcriptomic analysis of rice in response to iron deficiency and excess. *Rice* **7**, 18 (2014).
50. Hanada, K. *et al.* Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci.* **110**, 2395–2400 (2013).
51. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37**, 825–831 (2009).
52. Rossi, M. J. *et al.* A high-resolution protein architecture of the budding yeast genome. *Nature* **592**, 309–314 (2021).
53. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
54. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
55. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.* **7**, 16252 (2017).
56. Wang, Y., Hicks, S. C. & Hansen, K. D. Addressing the mean-correlation relationship in co-expression analysis. *PLOS Comput. Biol.* **18**, e1009954 (2022).
57. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
58. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

59. Yofe, I. *et al.* One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat. Methods* **13**, 371–378 (2016).
60. Schlicker, A., Domingues, F. S., Rahnenführer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**, 302 (2006).
61. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127–131 (2013).
62. Khitun, A., Ness, T. J. & Slavoff, S. A. Small open reading frames and cellular stress responses. *Mol. Omics* **15**, 108–116 (2019).
63. Wilson, B. A. & Masel, J. Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome Biol. Evol.* **3**, 1245–1252 (2011).
64. Li, D., Yan, Z., Lu, L., Jiang, H. & Wang, W. Pleiotropy of the de novo-originated gene MDF1. *Sci. Rep.* **4**, (2014).
65. Frumkin, I. & Laub, M. T. Selection of a de novo gene that can promote survival of E. coli by modulating protein homeostasis pathways. 2023.02.07.527531 Preprint at https://doi.org/10.1101/2023.02.07.527531 (2023).
66. Pagé, N. *et al.* A Saccharomyces cerevisiae Genome-Wide Mutant Screen for Altered Sensitivity to K1 Killer Toxin. *Genetics* **163**, 875–894 (2003).
67. Kesner, J. S., Chen, Z., Aparicio, A. A. & Wu, X. A unified model for the surveillance of translation in diverse noncoding sequences. 2022.07.20.500724 Preprint at https://doi.org/10.1101/2022.07.20.500724 (2022).
68. Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T. & Steitz, J. A. Widespread Inducible Transcription Downstream of Human Genes. *Mol. Cell* **59**, 449–461 (2015).
69. Wu, Q. *et al.* Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J.* **39**, e104763 (2020).

70. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* **14**, 417–419 (2017).

71. Lin, P., Troup, M. & Ho, J. W. K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).

72. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

73. Lovell, D. R., Chua, X.-Y. & McGrath, A. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. *NAR Genomics Bioinforma.* **2**, lqaa040 (2020).

74. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).

75. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

76. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems**, 1695 (2005).

77. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in *Proceedings of the 7th python in science conference* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (2008).

78. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

79. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

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

81. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).