Massively integrated coexpression analysis reveals transcriptional regulation, evolution and cellular implications of the noncanonical translatome

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

* corresponding author

Abstract

Background:

Recent studies uncovered pervasive transcription and translation of thousands of noncanonical open reading frames (nORFs) outside of annotated genes. The contribution of nORFs to cellular phenotypes is difficult to infer using conventional approaches because nORFs tend to be short, of recent *de novo* origins, and lowly expressed. Here we develop a dedicated coexpression analysis framework that accounts for low expression to investigate the transcriptional regulation, evolution, and potential cellular roles of nORFs in *Saccharomyces cerevisiae*.

Results:

Our results reveal that nORFs tend to be preferentially coexpressed with genes involved in cellular transport or homeostasis but rarely with genes involved in RNA processing. Mechanistically, we discover that young *de novo* nORFs located downstream of conserved genes tend to leverage their neighbors' promoters through transcription readthrough, resulting in high coexpression and high expression levels. Transcriptional piggybacking also influences the coexpression profiles of young *de novo* nORFs located upstream of genes, but to a lesser extent and without detectable impact on expression levels. Transcriptional piggybacking influences, but does not determine, the transcription profiles of *de novo* nORFs emerging nearby genes. About 40% of nORFs are not strongly coexpressed with any gene but are transcriptionally regulated nonetheless and tend to form entirely new transcription modules. We offer a web browser interface (https://carvunislab.csb.pitt.edu/shiny/coexpression/) to efficiently query, visualize and download our coexpression inferences.

Conclusions:

Our results suggest that nORF transcription is highly regulated. Our coexpression dataset serves as an unprecedented resource for unraveling how nORFs integrate into cellular networks, contribute to cellular phenotypes, and evolve.

Keywords:

Coexpression networks, de novo gene birth, noncanonical ORFs, translatome, smORFs, transcriptional regulation

Background

Eukaryotic genomes encompass thousands of open reading frames (ORFs). The vast majority are so-called “noncanonical” ORFs (nORFs) excluded from genome annotations because of their short length, lack of evolutionary conservation, and perceived irrelevance to cellular physiology [1–3]. The development of RNA sequencing (RNA-seq) [4] and ribosome profiling [5,6] has revealed genome-wide transcription and translation of nORFs across species ranging from yeast to humans [6–14]. Recent studies have characterized individual nORFs that form stable peptides and impact phenotypes, including cell growth [10,13,15], cell cycle regulation [16], muscle physiology [17–19], and immunity [20–22]. Unraveling the cellular, physiological and evolutionary implications of nORFs has become an active area of research [14,23].

Many nORFs have evolved *de novo* from previously noncoding regions [24–26]. Thus, the study of nORFs and *de novo* gene birth as evolutionary innovation carries a synergistic overlap where findings in one area could improve our understanding of the other. For instance, Sandmann et al. measured physical protein interactions for hundreds of peptides translated from nORFs and proposed that short linear motifs present in young *de novo* nORFs could mediate how nORFs impact essential cellular processes [26]. Other studies observed a gradual integration of evolutionary young ORFs into cellular networks and showed they could gain essential roles [27–29]. These studies support an evolutionary model whereby pervasive expression of nORFs generates the raw material for *de novo* gene birth [24,25].

The biological interpretation of nORF expression is complex. Some studies suggest that the transcription or translation of nORFs could be attributed to expression noise [30–32], whereby non-specific binding of RNA polymerases and ribosomes to DNA and RNA might cause promiscuous transcription or translation, respectively. How do nORFs become expressed in the first place? There are multiple hypotheses on how *de novo* ORFs gain the ability to become transcriptionally regulated [33]. One possibility is the emergence of novel regulatory regions along with or following the emergence of an ORF (ORF-first), as was shown for specific *de novo* ORFs in *Drosophila melanogaster* [34], codfish [35], human [36,37] and chimpanzee [36]. Alternatively, ORFs may emerge on actively transcribed loci such as near enhancers [38] or on long noncoding RNAs [39], as was shown for *de novo* ORFs in primates [40] and for *de novo* ORFs upstream or downstream of transcripts containing genes [37] (transcription-first) [41–43]. Transcription has a ripple effect causing coordinated activation of nearby genes [44,45]. Thus, *de novo* ORFs that emerge near established genes or regulatory regions may acquire transcriptional regulation by ‘piggybacking’ [45] on the pre-existing regulatory context [41,46]. This piggybacking could predispose *de novo* ORFs to be involved in similar cellular processes as their neighbors, which in turn would help with characterization. To date, the fraction of nORFs that are transcriptionally regulated and contribute to cellular phenotypes is unknown for any species.

An obstacle to studying nORF expression at scale is their detection, as nORF expression levels are typically low and reliant on specific conditions [24,36]. Recent studies demonstrated that the integration of omics data [14,47–49] could effectively address detection issues. For example, Wacholder et al. [14] recently discovered around 19,000 translated nORFs in *Saccharomyces cerevisiae* by massive integration of ribosome profiling data. This figure is three times larger than the number of canonical ORFs (cORFs) annotated in the yeast genome. These translated nORFs have the potential to generate peptides that affect cellular phenotypes but are almost entirely uncharacterized.

Coexpression is a well-established approach for studying transcriptional regulation through the massive integration of RNA-seq data. Coexpression refers to the similarity between transcriptional profiles of ORF pairs across numerous samples. Coexpression has been used successfully to identify new gene functions [50,51], disease-related genes [22,52,53] and for studying the conservation of the regulatory machinery [51,54] or gene modules [55] between species. Based on the assumption that genes involved in similar pathways have correlated expression patterns, coexpression can reveal relationships between genes and other transcribed genetic elements [56,57]. Most coexpression studies have focused on cORFs, but the abundance of publicly available RNA-seq data represents a tractable avenue to interrogate the transcriptional regulation of thousands of nORFs at once using coexpression approaches [47,58–61]. Indeed, RNA-seq is probe-agnostic and annotation-agnostic, thereby enabling the reuse of existing data to explore these novel ORFs. However, low expression levels can distort coexpression inferences due to statistical biases [62,63]. A coexpression analysis of translated nORFs that addresses the statistical issues arising from low expression is still lacking for any species.

Here, we developed a dedicated statistical approach that accounts for low expression levels when inferring coexpression relationships between ORFs. We applied this approach to the recently identified 19,000 translated nORFs in *S. cerevisiae* [14] and built the first high-quality coexpression network spanning the canonical and noncanonical translatome of any species. Coexpression relationships suggest that the majority of nORFs are transcriptionally regulated. While many nORFs form entirely new noncanonical transcription modules, approximately half are transcriptionally associated with genes involved in cellular homeostasis and transport. We show that *de novo* ORFs that piggyback onto their neighbors' transcription tend to have higher expression and tend to be highly coexpressed with their neighbors. We provide a web application to allow researchers to easily access this dataset to investigate the coexpression relationships and potential cellular roles for thousands of ORFs.

Results

High-quality coexpression inferences show transcriptional and regulatory relationships between nORFs and cORFs

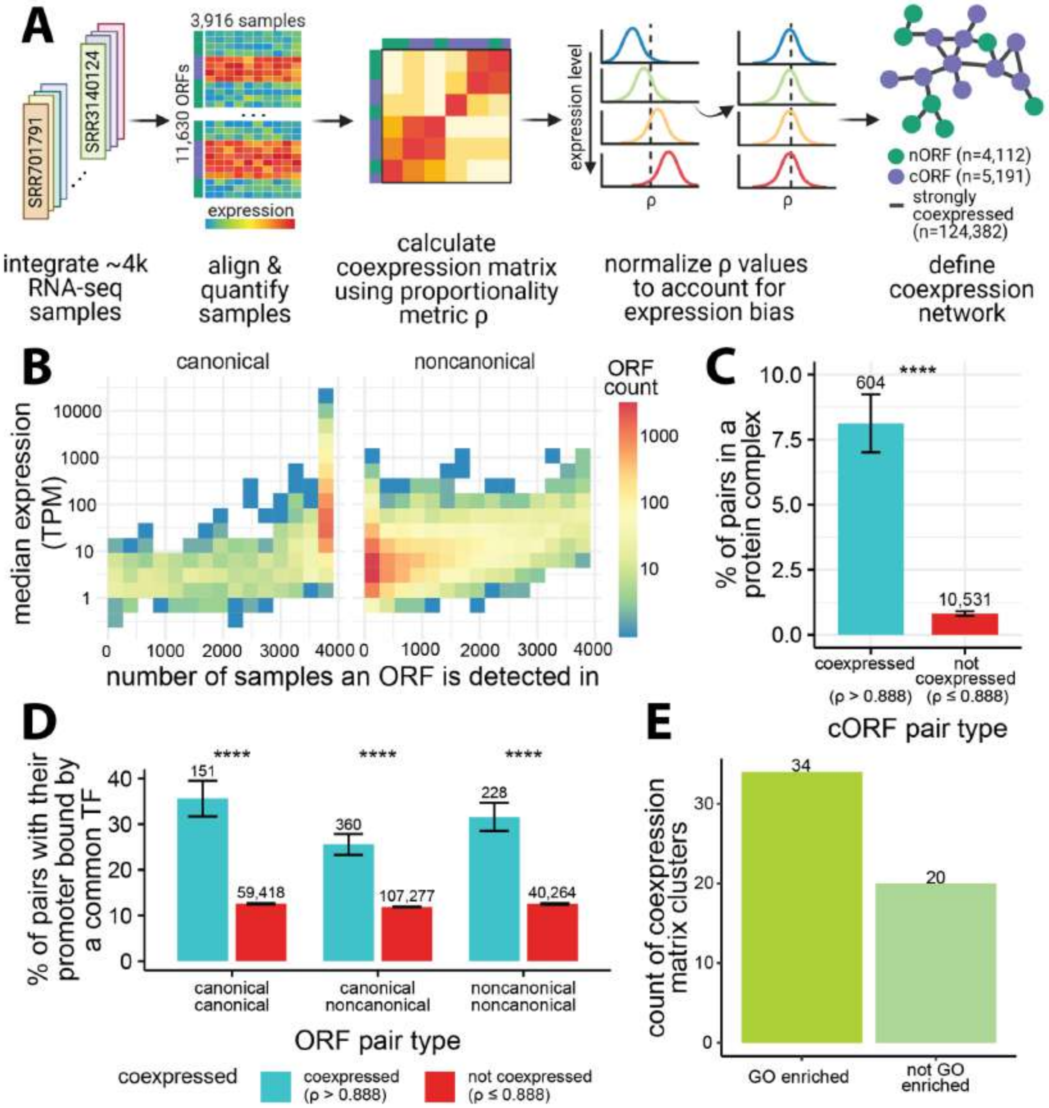


Figure 1: Overview of coexpression inference framework and properties of the dataset

A) Workflow: 3,916 samples were analyzed to create an expression matrix for 11,630 ORFs, including 5,803 cORFs and 5,827 nORFs; center log ratio transformed (clr) expression values were used to calculate the coexpression matrix using proportionality metric, ρ, followed by normalization 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 (transcript per million - TPM) and the number of samples the ORFs were detected in with at least 5 raw counts. C) Coexpressed cORF pairs (ρ > 0.888) are more likely to encode proteins that form complexes than non-coexpressed cORF pairs (Fisher's exact test p < 2.2e-16; error bars: standard error of the proportion); using annotated protein complexes from ref. [64]. D) Coexpressed ORF pairs (ρ > 0.888) are more likely to have their promoters bound by a common transcription factor (TF) than non-coexpressed ORF pairs (Fisher's exact test p < 2.2e-16; error bars: standard error of the proportion); genome-wide TF binding profiles from ref. [65] and transcription start sites (TSS) from ref. [66] were analyzed to define promoter binding (see Methods). E) Hierarchical clustering of the coexpression matrix reveals functional enrichments for most clusters that contain at least 5 cORFs; functional enrichments estimated by gene ontology (GO) enrichment analysis at false discovery rate (FDR) < 0.05 using Fisher's exact test.

To infer 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) [67], as well as all nORFs, ORFs that were either unannotated or annotated as "dubious" and "pseudogene", with evidence of translation according to Wacholder et al. [14]. To maximize detection of transcripts containing nORFs, we curated and 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 multiple 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 [68]. Thus, we used clr to transform the raw read counts and quantified coexpression relationships using the proportionality metric, ρ [69].

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). These steps resulted in an 11,630 by 11,630 coexpression matrix encompassing 5,803 cORFs and 5,827 nORFs (ORF list in Supplementary Data 2).

The combined use of clr, ρ, and sample thresholding accounted for statistical issues in estimating coexpression deriving from sparsity, but the large difference in RNA expression levels between cORFs and nORFs posed yet another challenge. Indeed, Wang et al. showed that the distribution of coexpression values is biased by expression level due to statistical artifacts [62]. We observed this artifactual bias in our dataset (Supplementary Figure 2A) and corrected for it using spatial quantile normalization (SpQN) as recommended by Wang et al. [62] (Supplementary Figure 2B). This resulted in a normalized coexpression matrix (Supplementary Data 3) with ρ values centered around 0.476.

We then created a network representation of the coexpression matrix by considering only the top 0.2% of ρ values between all ORF pairs ($\rho > 0.888$). This threshold was chosen to include

90% of cORFs (Supplementary Figure 3). Altogether, our dedicated analysis framework (Figure 1A) inferred 124,382 strong ($\rho > 0.888$) coexpression relationships between 9,303 ORFs, encompassing 4,112 nORFs and 5,191 cORFs.

To assess whether our coexpression network captures meaningful biological and regulatory relationships, we examined its overlap with orthogonal datasets. Using a curated [64] protein complex dataset for cORFs, we found that coexpressed cORF pairs are significantly more likely to encode proteins that form 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 [65] genome-wide chromatin immunoprecipitation with exonuclease digestion (ChIP-exo) dataset containing DNA-binding information for 73 sequence-specific transcription factors (TFs) and using transcript isoform sequencing (TIF-seq) [66] data to determine transcription start sites (TSSs) and promoter regions, 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 = 3.84, *canonical-noncanonical pairs*: Odds ratio = 2.55, *noncanonical-noncanonical pairs*: Odds ratio = 3.22, Fisher's exact test $p < 2.2e-16$ for all three comparisons; Figure 1D). Enrichments were robust to different coexpression cutoffs (Supplementary Figure 4-5). Using the WGCNA [70] method to cluster the coexpression matrix, we found that more than half of the clusters identified contained functionally related ORFs (gene ontology (GO) biological process enrichments at Benjamini-Hochberg (BH) adjusted false discovery rate (FDR) < 0.05; Figure 1E; Supplementary Figure 6). These analyses demonstrate the high quality of our coexpression network and confirm that it captures meaningful biological and regulatory relationships for both cORFs and nORFs.

Conventional approaches for coexpression analysis include using transcript per million (TPM) or reads per kilobase per million (RPKM) normalization, batch correction by removing top principal components, and Pearson's correlation as the similarity metric [71,56,72]. Compared to these approaches, our framework increased the proportion of coexpressed ORF pairs whose promoters are bound by a common TF specifically for pairs containing nORFs (Supplementary Figure 7), and yielded coexpression networks encompassing the largest number of nORFs at most thresholds (Supplementary Figure 8). Hence our dedicated analysis framework therefore outperforms conventional coexpression approaches for the study of nORFs. We offer an R Shiny [73] interface (https://carvunislab.csb.pitt.edu/shiny/coexpression/) to efficiently query, visualize and download the coexpression data we generated. To our knowledge, this is the most comprehensive coexpression dataset focusing on empirically translated elements, both annotated and unannotated, for any species to date.

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

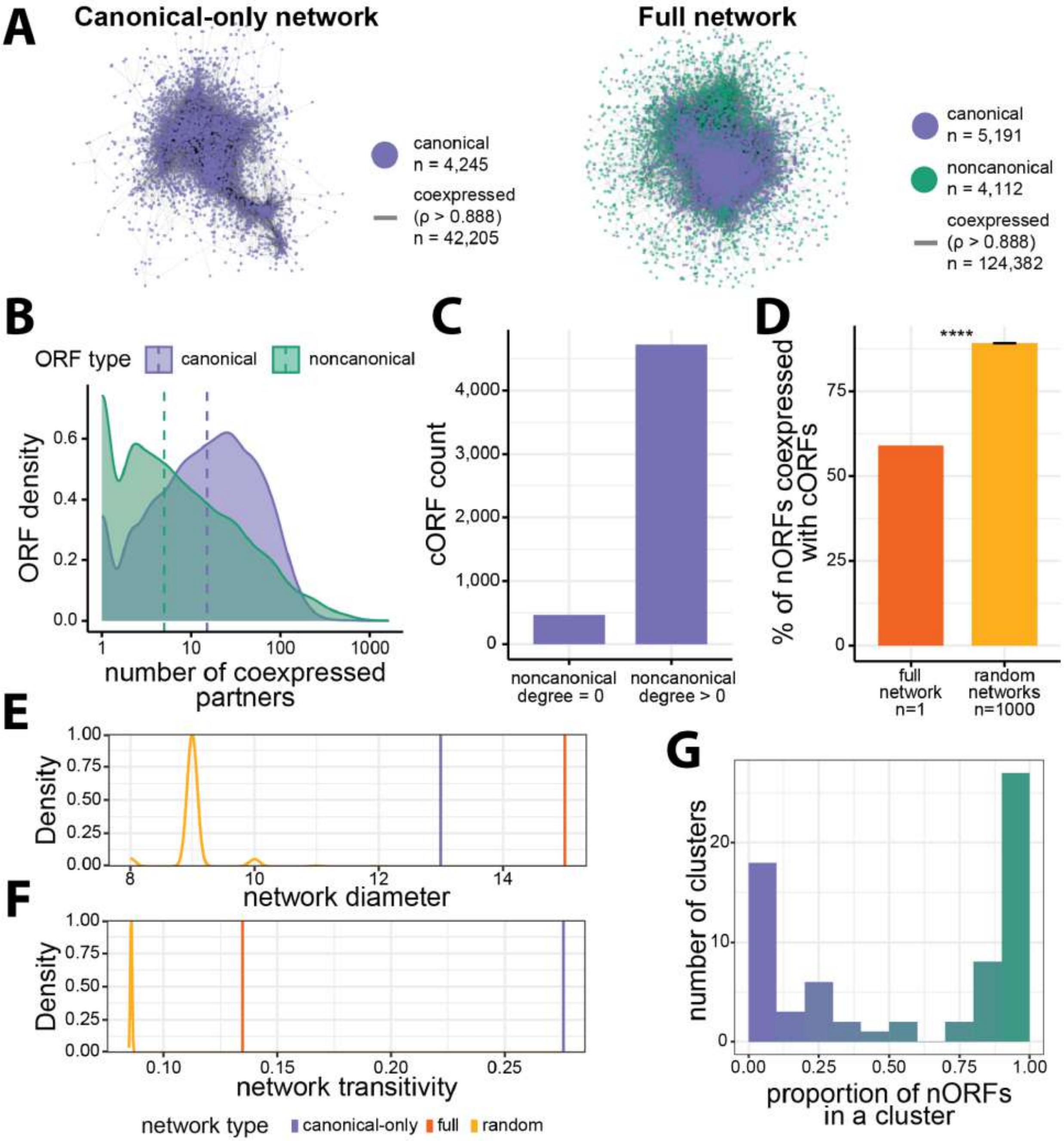


Figure 2 Topological properties of the coexpression network

A) Visualization for canonical-only and full coexpression networks using spring embedded graph layout [74]. The full 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 full network) than cORFs (Mann-Whitney U-test p < 2.2e-16). C) Most cORFs are coexpressed with at least one nORF. D) Only 59% of nORFs are coexpressed with at least one cORFs and this is less than expected by chance, on average, 89% of nORFs are coexpressed with a cORF across 1,000 randomized networks generated in a degree-preserving fashion by swapping edges of noncanonical nodes (Fisher's exact test p < 2.2e-16; error bar: standard error of the mean proportion across randomized networks). E) Addition of nORFs to the canonical-only network results in the full network being less compact, 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-only network decreases local clustering in the full network, however this is to a lesser extent than expected by chance as shown by the distribution for the 1,000 randomized networks. G) Most clusters in the coexpression matrix encompass either primarily nORFs or primarily cORFs (n= 69 clusters, *green* represents nORF majority clusters, *purple* represents cORF majority clusters).

Conventional analyses of coexpression networks have been restricted to cORFs. Our full coexpression network contains twice the number of ORFs and three times the number of strong ($\rho > 0.888$) coexpression relationships compared to the canonical-only network (Figure 2A). We sought to compare the network properties of the canonical-only and full networks. 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 9). This suggests that, while most nORFs are integrated in the full coexpression network, they also have distinct expression profiles that differ markedly from those of all cORFs and are more similar to those of other nORFs.

To investigate how these seemingly conflicting attributes impact the organization of the coexpression network, we analyzed two global network properties: 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 incorporation of nORFs in the full network led to a larger diameter relative to the canonical-only network (Figure 2E). This is in sharp contrast with the null expectation, set by 1,000 degree-preserving simulations, whereby random incorporation of nORFs decreases network diameter. The full coexpression network is thus much less compact than expected by chance, suggesting that nORFs tend to be located at the periphery of the network. Network transitivity decreased with the incorporation of nORFs compared to the canonical-only network, but to a lesser extent than expected by chance (Figure 2F). This suggests that despite their low degree and peripheral locations, the connections formed by nORFs are structured and may form noncanonical clusters.

To investigate this hypothesis, we inspected the ratio of nORFs and cORFs among the cluster assignments from WGCNA hierarchical clustering of the full coexpression matrix (Supplementary Figure 6). Strikingly, we observed a bimodal distribution of clusters, with approximately half of the clusters consisting mostly of nORFs and the other half containing mostly cORFs (Figure 2G). We conclude that nORFs exhibit a unique and non-random organization within the coexpression network, simultaneously connecting to all cORFs while also forming entirely new noncanonical transcription modules.

Coexpression profiles reveal most nORFs are transcriptionally associated with genes involved in cellular transport and homeostasis

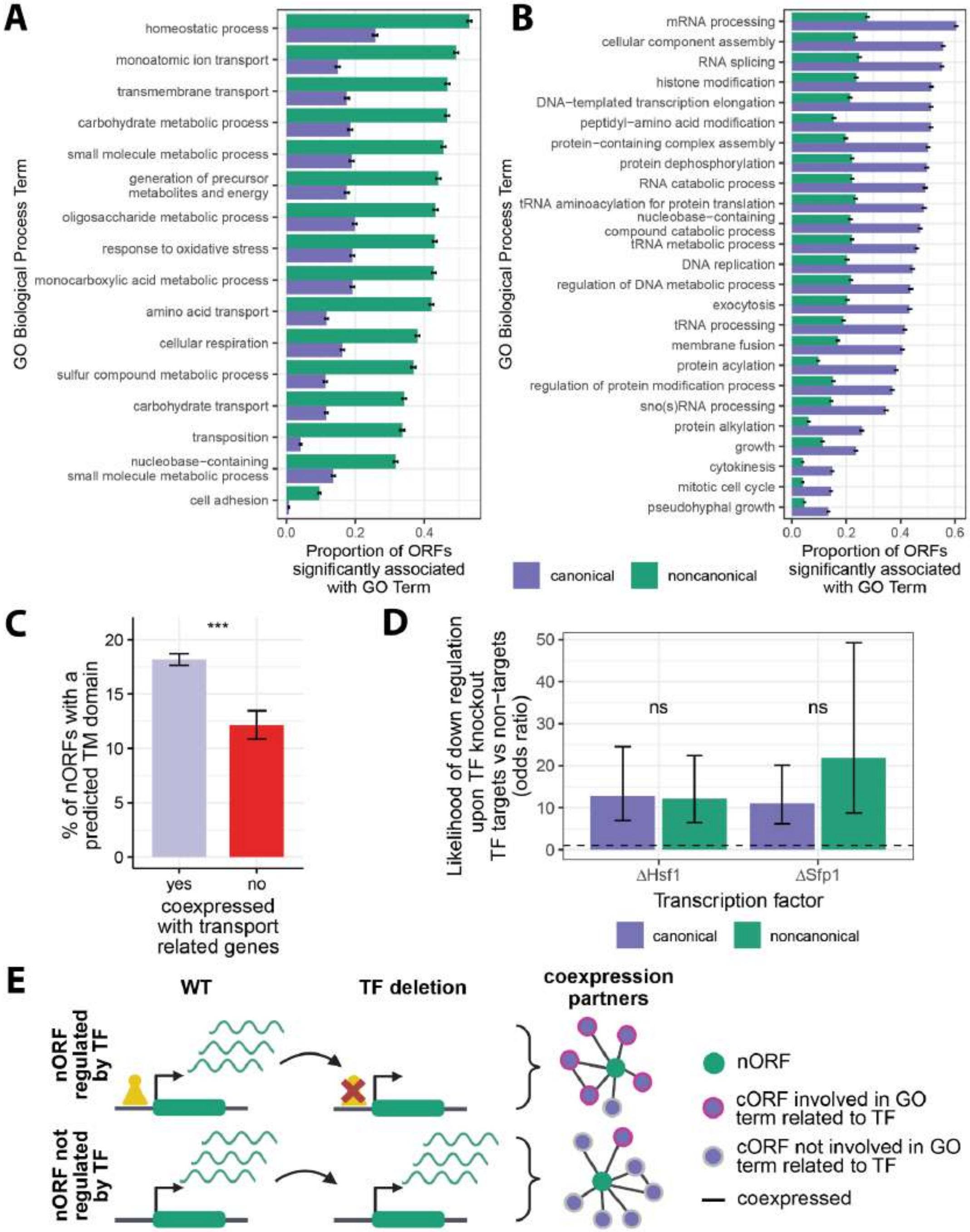


Figure 3 Biological processes associated with nORF transcriptional regulation

A-B) Biological processes that are more (A) (Odds ratio > 2, n = 16 terms) or less (B) (Odds ratio < 0.5, n = 23 terms) transcriptionally associated with nORFs than cORFs (y-axis ordered by nORF enrichment proportion from highest to lowest, BH adjusted FDR < 0.001 for all terms, Fisher's exact test, GO term enrichments were detected using gene set enrichment analyses (GSEA), error bars: standard error of the proportion). C) nORFs that are highly coexpressed with genes involved in transport are more likely to have predicted transmembrane (TM) domains as determined by TMHMM [75] compared to nORFs that are not (Odds ratio = 1.6, Fisher's exact test p = 1.3e-4; error bars: standard error of the proportion). D) nORFs and cORFs that are Sfp1 or Hsf1 targets are more likely to be downregulated when Sfp1 or Hsf1 are deleted compared to ORFs that are not targets (*Sfp1*: cORFs: p < 2.2e-16; nORFs: p = 2.8e-9; *Hsf1*: cORFs: p <2.2e-16; nORFs: p = 9.9e-13; Fisher's exact test, error bars: 95% confidence interval of the odds ratio; *dashed* line shows odds ratio of 1; RNA abundance data from SRA accession SRP159150 and SRP437124 [76] respectively). E) nORFs that are regulated by TFs are more likely to be coexpressed with genes involved in processes related to known functions of that TF.

To determine whether nORFs are transcriptionally associated with specific cellular processes, we performed gene set enrichment analyses [77] (GSEA) on their coexpression partners. GSEA takes an ordered list of genes, in this case sorted by coexpression level, and seeks to find if the higher ranked genes are preferentially annotated with specific GO terms. For each cORF and nORF, we ran GSEA to detect if their highly coexpressed partners were preferentially associated with any GO terms (Supplementary Figure 10). Almost all ORFs (99.9%), whether cORF or nORF, had at least one significant GO term associated with their coexpression partners at BH adjusted FDR < 0.01, suggesting that nORFs are engaged in coherent transcriptional programs. We then calculated, for each GO term, the number of cORFs and nORFs with GSEA enrichments in this term (Supplementary Data 4). These analyses identified specific GO terms that were significantly more (16 terms, BH adjusted FDR < 0.001, Odds ratio > 2, Fisher's exact test; Figure 3A, Supplementary Data 5) or less (23 terms, BH adjusted FDR < 0.001, Odds ratio < 2, Fisher's exact test; Figure 3B, Supplementary Data 5) prevalent among the coexpression partners of nORFs relative to those of cORFs. Most of the GO terms that were significantly enriched among the coexpression partners of nORFs were related to cellular homeostasis and transport (Figure 3A) while most of the GO terms significantly depleted among the coexpression partners of nORFs were related to DNA, RNA, and protein processing (Figure 3B). Running the same GSEA pipeline with Kyoto Encyclopedia of Genes and Genomes (KEGG) [78] annotations yielded consistent results (Supplementary Figure 11, Supplementary Data 6-7). Half of nORFs were coexpressed with genes involved in homeostasis (GO:0042592, 53%), monoatomic ion transport (GO:0006811, 49%) and transmembrane transport (GO:0055085, 47%). The nORFs transcriptionally associated with the parent term 'transport' (n = 2,718, GO:0006810, GSEA BH adjusted FDR < 0.01) were 1.6 times more likely to contain a predicted transmembrane domain than other nORFs (p = 1.3e-4, Fisher's exact test; Figure 3C), in line with potential transport-related activities. These findings reveal a strong and previously unsuspected transcriptional association between nORFs, and cellular processes related to homeostasis and transport.

Hsf1 and Sfp1 nORF targets are part of protein folding and ribosome biogenesis transcriptional programs, respectively

Overall, our analyses relating coexpression to TF binding (Figure 1D) and functional enrichments (Figure 3A-B) suggest that nORF expression is regulated rather than simply the consequence of transcriptional noise. To further investigate this hypothesis, we sought to identify regulatory relationships between specific TFs and nORFs. We reasoned that if nORFs are regulated by TFs in similar ways as cORFs, then genetic knockout of the TFs that regulate them should impact their expression levels as it does for cORFs [79]. We focused on two transcriptional activators for which both ChIP-exo [65] and knockout RNA-seq data [76] were publicly available: Sfp1, which regulates ribosome biogenesis [80] and Hsf1, which regulates heat shock and protein folding responses [81].

For both cORFs and nORFs, knockout of Sfp1 or Hsf1 was more likely to trigger a significant decrease in expression when the ORF's promoter was bound by the respective TF according to ChIP-exo evidence (Figure 3D). The statistical association between TF binding and knockout-induced downregulation was as strong for nORFs as it was for cORFs, consistent with nORFs having similar mechanisms of transcriptional activation (*Sfp1*: cORFs Odds ratio = 11.1, p < 2.2e-16; nORFs Odds ratio = 21.8, p = 2.8e-9, Fisher’s exact test; *Hsf1*: cORFs Odds ratio = 12.7, p < 2.2e-16; nORFs Odds ratio = 12.1, p = 9.9e-13, Fisher’s exact test). Therefore, the nORFs whose promoters are bound by these TFs, and whose expression levels decrease upon deletion of these TFs, are likely genuine regulatory targets of these TFs. By this stringent definition, our analyses identified 9 nORF targets of Sfp1 (and 34 cORF targets) and 19 nORF targets of Hsf1 (and 39 cORF targets). The coexpression profiles of these Sfp1 and Hsf1 nORF targets were preferentially associated with genes involved in processes directly related to the known functions of Sfp1 and Hsf1 (Supplementary Data 8). For example, the coexpression profiles of 9 Sfp1 nORF targets revealed preferential associations with genes involved in ‘ribosomal large subunit biogenesis’ and 7 Sfp1 nORF targets involved in ‘regulation of translation’ according to our GSEA pipeline (Fisher’s exact test, BH adjusted p-value < 6.7e-4 for both terms). Similarly, 13 Hsf1 nORF targets were preferentially associated with genes involved in ‘Protein Folding’ (Fisher’s exact test, BH adjusted p-value = 5.7e-9). These results show that nORF expression can be actively regulated by TFs as part of coherent transcriptional programs (Figure 3E).

de novo ORF expression and regulation are shaped by genomic location

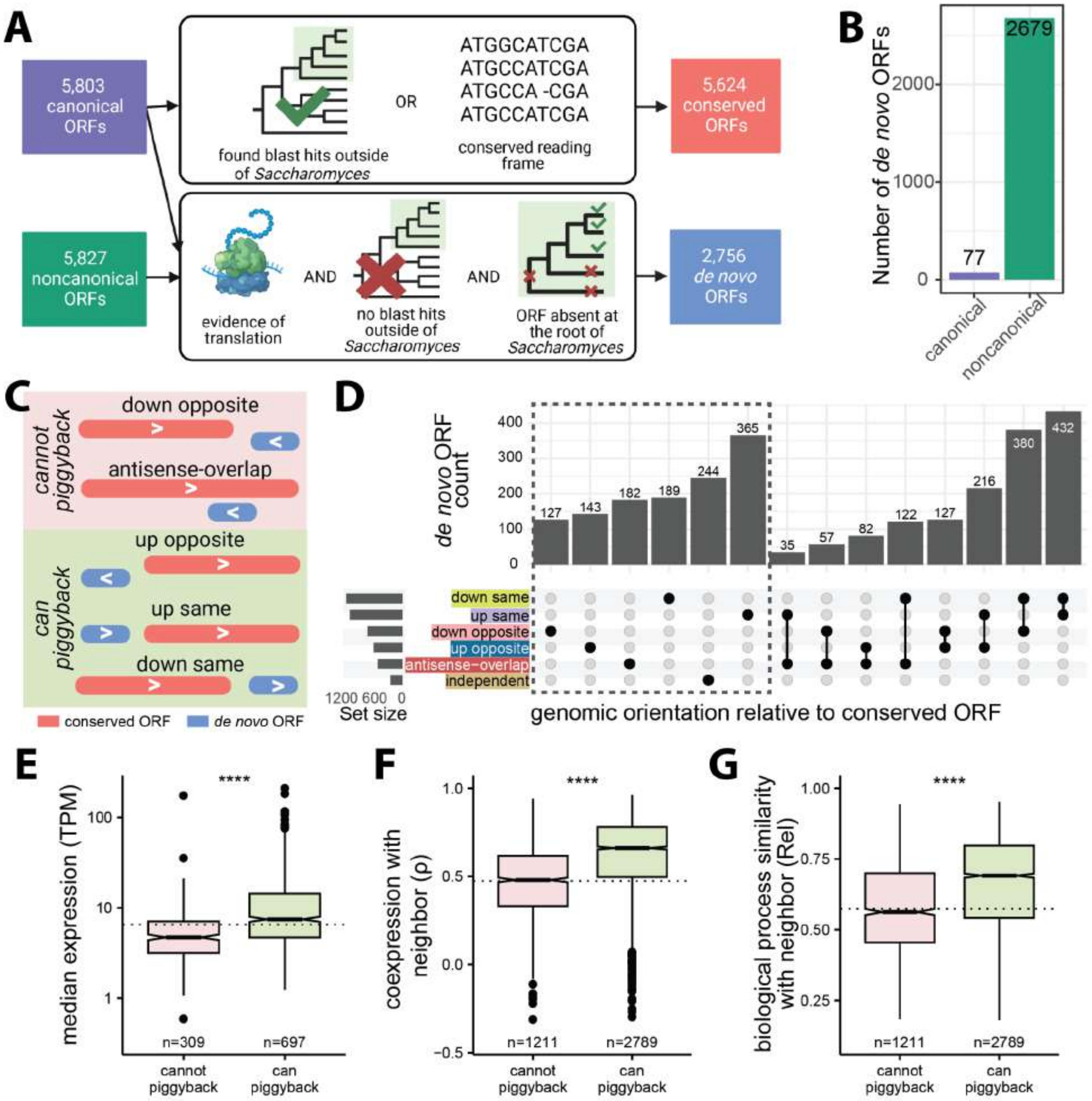


Figure 4 Expression, coexpression and biological processes similarity of *de novo* ORFs with respect to genomic orientations

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 within *Saccharomyces* (*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 (*bottom*). B) Counts of cORFs and nORFs that emerged *de novo*. C) Genomic orientations of *de novo* ORFs that cannot transcriptionally piggyback off neighboring conserved ORF (cannot share promoter with neighbor, *pink shading*) or can transcriptionally piggyback off neighboring conserved ORF (possible to share promoter with neighbor, *green shading*). D) Counts of *de novo* ORFs that are within 500 bp of a conserved ORF in different genomic orientations; ORFs further than 500bp are classified as independent. E) *De novo* ORFs in orientations that can piggyback have higher RNA expression levels than *de novo* ORFs in orientations that cannot piggyback (Cliff's Delta d = 0.4). Only *de novo* ORFs in a single orientation are considered (dashed box in panel *D*). Dashed line represents the median expression of independent *de novo* ORFs. F) *de novo* ORFs in orientations that can piggyback have higher coexpression with neighboring conserved ORFs compared to *de novo* ORFs in orientations that cannot piggyback (Cliff's Delta d = 0.43). Dashed line represents median coexpression of *de novo*-conserved ORF pairs on separate chromosomes. G) *de novo* ORFs in orientations that can piggyback are more likely to be transcriptionally associated with genes involved in the same biological processes as their neighboring conserved ORFs than *de novo* ORFs in orientations that cannot piggyback (Cliff's Delta d = 0.31). Dashed line represents median functional enrichment similarities of *de novo*-conserved ORF pairs on separate chromosomes. (For panels E-F-G: Mann-Whitney U-test, ****: $p < 2.2e-16$).

Previous literature has shown that many nORFs arise *de novo* from previously noncoding regions [24,26]. We wanted to investigate how these evolutionarily novel ORFs acquire expression and whether their locus of emergence influences this acquisition. To define which ORFs were of recent *de novo* evolutionary origins, we developed a multistep pipeline combining sequence similarity searches and syntenic alignments (Figure 4A). cORFs were considered conserved if they had homologues detectable by sequence similarity searches with BLAST in budding yeasts outside of the *Saccharomyces* genus or if their open reading frames were maintained within the *Saccharomyce*s genus [14]. cORFs and nORFs 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 criteria aimed to identify the youngest *de novo* ORFs. Overall, we identified 5,624 conserved cORFs and 2,756 *de novo* ORFs including 77 *de novo* cORFs and 2,679 *de novo* nORFs (Figure 4B). In general, the coexpression patterns of *de novo* ORFs (Supplementary Figure 12) were similar to those of nORFs (Figure 3A-B).

We hypothesized that the locus where *de novo* ORFs arise may influence their expression profiles through “piggybacking” off their neighboring conserved ORFs’ pre-existing regulatory environment. To investigate this hypothesis, we categorized *de novo* ORFs based on their positioning relative to neighboring conserved ORFs. The *de novo* ORFs further than 500 bp from all conserved ORFs were classified as independent. The remaining *de novo* ORFs were classified 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 the nearest conserved ORF (Figure 4C-D). We categorized the orientations as being able to piggyback or unable to piggyback based on their potential of sharing a promoter with neighboring conserved ORFs, with down opposite and antisense overlap as orientations that cannot piggyback and up opposite, up same, and down same as orientations that can piggyback (Figure 4C). The piggybacking hypothesis predicts that *de novo* ORFs that arise in orientations that can piggyback would be positively influenced by the regulatory environment provided by the promoters of neighboring conserved ORFs, resulting in similar transcription profiles as their neighbors and increased expression relative to *de novo* ORFs that do not benefit from a pre-existing regulatory environment.

We considered three metrics to assess piggybacking: RNA expression level, measured as median TPM over all the samples analyzed, coexpression with neighboring conserved ORF and biological process similarity with neighboring conserved ORF. To calculate biological process similarity between two ORFs, we used significant GO terms at FDR < 0.01 determined by coexpression GSEA for each ORF (Supplementary Figure 10) and calculated the similarity between these two sets of GO terms using the relevance method [82]. If two ORFs are enriched in the same specialized terms, their relevance metric would be higher than if they are enriched in different terms or in the same generic terms. We found that *de novo* ORFs in orientations that can piggyback tend to have higher expression (focusing only on ORFs that could be assigned a single orientation, dashed box in Figure 4D, Cliff's Delta d = 0.4; Figure 4E), higher coexpression with their neighbor (Cliff's Delta d = 0.43; Figure 4F), and higher biological process similarity (Cliff's Delta d = 0.31; Figure 4G), compared to ORFs in orientations that cannot piggyback (p < 2.2e-16 Mann-Whitney U-test for all). Thus, all three metrics supported the piggybacking hypothesis.

Closer examination revealed a more complex situation. First, the immediate neighbors of *de novo* ORFs in orientations that can piggyback were rarely among their strongest coexpression partners (only found in the top 10 coexpressed partners for 15% of down same, 4.5% of up same, 3% of up opposite ORFs). Therefore, emergence nearby a conserved ORF in a piggybacking orientation influences, but does not fully determine, the transcription profiles of *de novo* ORFs. Transcriptional regulation beyond that provided by the pre-existing regulatory environment may exist. Second, while ORFs in all three orientations that can piggyback displayed increased coexpression and biological process similarity with their neighbors relative to background expectations (Supplementary Figure 13A-B), only down same *de novo* ORFs displayed increased RNA expression levels (Supplementary Figure 13C). The expression levels of up same *de novo* ORFs were statistically indistinguishable from independent *de novo* ORFs, while those of up opposite *de novo* ORFs were significantly lower than those of independent *de novo* ORFs (Supplementary Figure 13C). Down same *de novo* ORFs also showed stronger coexpression and biological process similarity with their conserved neighbors than up same and up opposite *de novo* ORFs (Supplementary figure 13A-B). Therefore, the transcription of down same *de novo* ORFs appeared most susceptible to piggybacking.

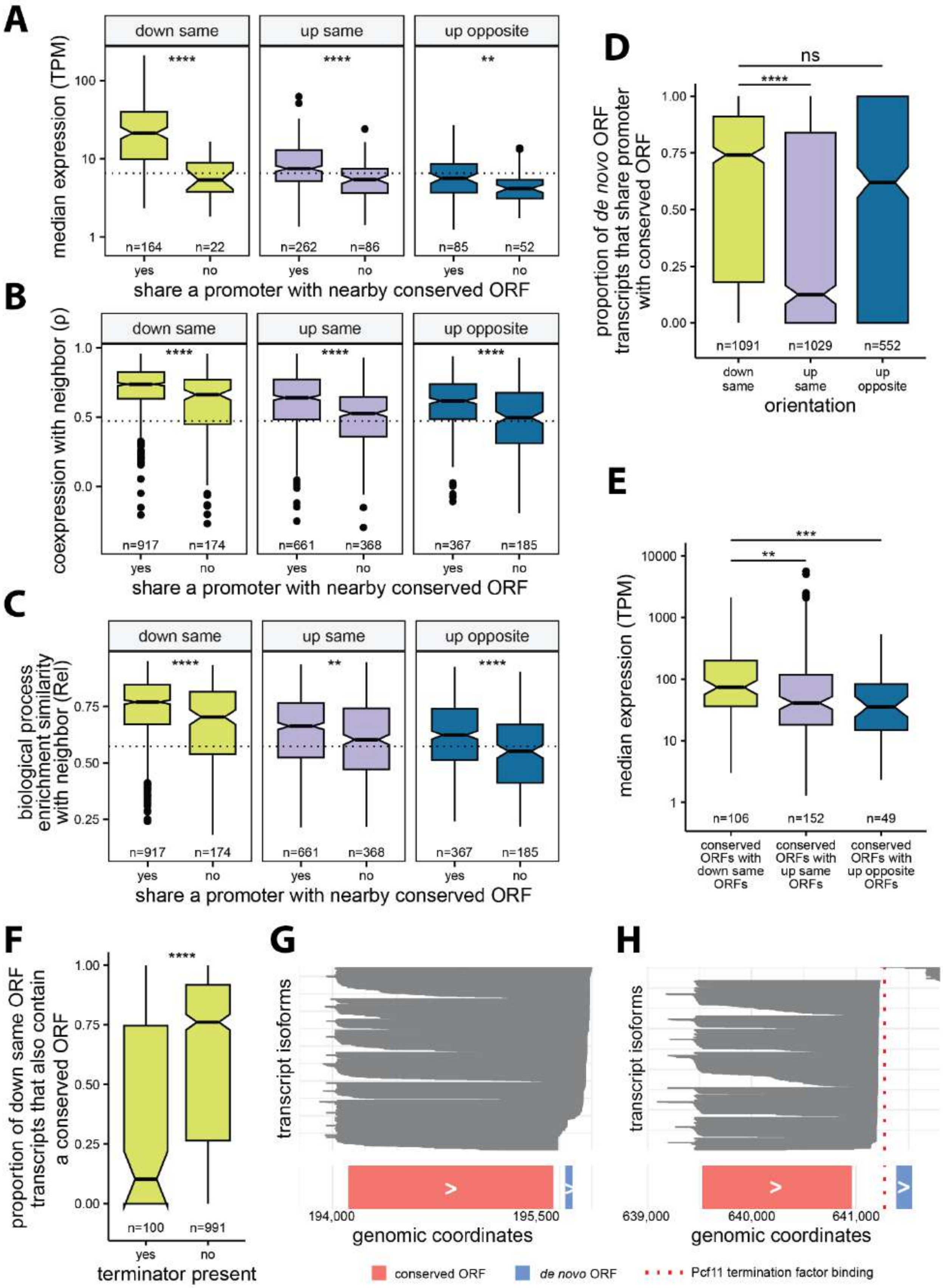

A
down same
up same
up opposite
median expression (TPM)
100
10
1
****
****
**
n=164
n=22
n=262
n=86
n=85
n=52
yes
no
share a promoter with nearby conserved ORF
B
coexpression with neighbor (ρ)
1.0
0.5
0.0
n=917
n=174
n=661
n=368
n=367
n=185
C
biological process enrichment similarity with neighbor (Rel)
0.75
0.50
0.25
D
ns
****
proportion of de novo ORF transcripts that share promoter with conserved ORF
1.00
0.75
0.50
0.25
0.00
n=1091
n=1029
n=552
down same
up same
up opposite
orientation
E
***
**
median expression (TPM)
10000
1000
100
10
1
n=106
n=152
n=49
conserved ORFs with down same ORFs
conserved ORFs with up same ORFs
conserved ORFs with up opposite ORFs
F
proportion of down same ORF transcripts that also contain a conserved ORF
1.00
0.75
0.50
0.25
0.00
n=100
n=991
terminator present
G
transcript isoforms
194,000
195,500
genomic coordinates
H
639,000
640,000
641,000
conserved ORF
de novo ORF
Pcf11 termination factor binding

Figure 5 Effects of promoter sharing on expression, coexpression and biological process similarities of *de novo* ORFs

A) *De novo* ORFs that share a promoter with neighboring conserved ORFs, as determined by TIF-seq transcript boundaries, have significantly higher expression levels than *de novo* ORFs that do not. Considering only ORFs in a single orientation. Dashed line represents the median expression of independent *de novo* ORFs. B) *De novo* ORFs that share a promoter with neighboring conserved ORFs have higher coexpression with their neighbors than *de novo* ORFs that do not share a promoter. Dashed line represents median coexpression of *de novo*-conserved ORF pairs on separate chromosomes. C) *De novo* ORFs that share a promoter have more similar functional enrichments with neighboring conserved ORFs than *de novo* ORFs that do not share a promoter. 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 promoter with neighboring conserved ORFs significantly more often than up same ORFs. E) Conserved ORFs with downstream *de novo* ORFs have a significant increase in expression compared to conserved ORFs with upstream *de novo* ORFs. F) Existence of transcription termination factors (Pcf11 or Nrd1) in between conserved ORFs and nearby downstream *de novo* ORFs leads to less shared transcripts. G) Transcript isoforms (*gray*) at an example locus where there are no transcription termination factors present between conserved ORF YBL015W (*pink*) and downstream *de novo* ORF chr2:195794-195847(+) (*blue*). H) Transcript isoforms (*gray*) at an example locus where there is Pcf11 transcription terminator present (*red line*) between conserved ORF YPR034W (*pink*) and downstream *de novo* ORF chr16:641385-641534(+) (*blue*). All detected transcript isoforms on these loci are plotted for G and F. (For all panels: ****: p ≤ 0.0001, ***: p ≤ 0.001, **: p ≤ 0.01, *: p ≤ 0.05, ns: not-significant; Mann-Whitney U-test)

To understand the molecular mechanisms leading to the differences in expression, coexpression and biological process similarity between the orientations that can piggyback, which all have the potential to share a promoter with neighboring conserved ORF, we investigated which actually do by analyzing transcript architecture. Using a publicly available TIF-seq dataset [66], we defined down same or up same ORFs as sharing a promoter with their neighbor if they mapped to the same transcript at least once. We defined up opposite ORFs as sharing a promoter with their neighbor if their respective transcripts did not have overlapping TSSs, as would be expected for divergent promoters [83]. According to these criteria, 84% of down same (n = 174), 64% of up same (n = 368), and 66% of up opposite (n = 185) *de novo* ORFs share a promoter with their neighboring conserved ORFs (Supplementary Figure 14). Among all *de novo* ORFs that arose in orientations that can piggyback, those that share promoters with neighboring conserved ORFs displayed higher expression levels than those that do not (*down same*: d = 0.75, p = 1.06e-8; *up same*: d = 0.38, p = 1.23e-7; *up opposite*: d = 0.3, p = 2.9e-3 Mann-Whitney U-test, d: Cliff's Delta; Figure 5A). We also observed a significant increase in coexpression and biological process similarity between *de novo* ORFs and their neighboring conserved ORFs when their promoters are shared compared to when they are not (coexpression: *down same*: d = 0.28, p = 2.99e-9; *up same*: d = 0.31, p < 2.2e-16; *up opposite*: d = 0.27, p = 2.1e-7; biological process similarity: *down same*: d = 0.24, p = 5.5e-7; *up same*: d = 0.108, p = 3.78e-3; *up opposite*: d = 0.24, p = 6.1e-6, d: Cliff's Delta, Mann-Whitney U-test; Figures 5B and 5C, respectively). Hence, sharing a promoter led to increases in the three piggybacking metrics for the three orientations.

Further supporting the notion that down same ORFs are particularly prone to piggybacking, the down same *de novo* ORFs that share a promoter with their conserved neighbors displayed much higher expression levels, and higher coexpression and biological process similarity with their conserved neighbor, than up same or up opposite ORFs that also share a promoter with their conserved neighbors (expression: *down same vs up same*: d = 0.58; *down same vs up opposite*: d = 0.55; coexpression: *down same vs up same:* d = 0.29, *down same vs up opposite:* d = 0.38; biological process similarity: *down same vs up same*: d = 0.37, *down same vs up opposite*: d = 0.45; d: Cliff's Delta, p < 2.2e-16 for all comparisons, Mann-Whitney U-test). This could be due to down same ORF's tendency to share promoters more often than up same ORFs, as a larger proportion of transcripts containing down same ORFs also contain a conserved ORF (*down same vs up same*: Cliff's Delta d = 0.26, Mann-Whitney U-test p < 2.2e-16; Figure 5D), or higher expression levels of conserved ORFs that have down same ORFs on their transcripts compared to conserved ORFs with up same or up opposite piggybacking ORFs (*down same vs up same*: d = 0.2, p = 5.4e-3; *down same vs up opposite*: d = 0.34, p = 6.5e-4, Mann-Whitney U-test, d: Cliff's Delta; Figure 5E).

Based on these results, we reasoned that transcriptional readthrough could be the molecular mechanism underlying the efficient transcriptional piggybacking of down same *de novo* ORFs. To investigate this hypothesis, 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 [65], we found that the presence of terminators between conserved ORFs and their downstream *de novo* ORF pairs resulted in a notably lower percentage of shared transcripts (Cliff's Delta d = -0.39, p = 1.59e-10, Mann-Whitney U-test; Figure 5F). 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 ρ = 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 (ρ = 0.1; Figure 5H). We conclude that sharing a transcript via transcriptional readthrough is the major transcriptional piggybacking mechanism for down same *de novo* ORFs.

Discussion

We explored the transcription of nORFs from multiple angles including network topology, associations with cellular processes, TF regulation, and influence of the locus of emergence on *de novo* ORF expression. Delving into network topology, we find that nORFs have distinct expression profiles that are strongly 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. The addition of nORFs to the cellular network resulted in a more clustered network than expected by chance, highlighting the previously unsuspected influence of nORFs in shaping the coexpression landscape.

Our study is the first to show a large-scale association between the expression of nORFs and cellular homeostasis and transport processes. We anticipate that future studies will follow up to test these associations experimentally. We also found nORFs to be preferentially associated with cellular processes related to metabolism, transposition and cell adhesion, but rarely with the core processes of the central dogma, DNA, RNA or protein processing. Genes involved in transport, metabolism, and stress tend to have more variable expression compared to genes in other pathways [84]. Pathways with more variable expression could be more likely to incorporate novel ORFs, possibly as a form of adaptive transcriptional response. There are several consistent observations in the literature [47,85,86]. For instance, Li et al. [47] showed that many *de novo* ORFs are upregulated in heat shock. Wilson and Masel [87] found higher translation of *de novo* ORFs under starvation conditions. Carvunis et al. [24] found *de novo* cORFs are enriched for the GO term 'response to stress'. Other studies showed examples of how specific *de novo* ORFs could be involved in stress response [35,88] or homeostasis [88,89]. For instance the *de novo* antifreeze glycoprotein AFGP allows Arctic codfish to live in colder environments [35] or *MDF1* in yeast [88,90] was found in a screen to provide resistance to certain toxins and mediates ion homeostasis [91]. Our results, combined with these previous investigations, argue that a large fraction of nORFs provide adaptation to stresses and help maintain homeostasis, perhaps through modulation of transport processes.

Recent research in yeast has revealed an enrichment of transmembrane domains [15,24,92,93] within *de novo* ORFs. Previous studies identified small nORFs and *de novo* ORFs that localize to diverse cellular membranes, such as those of the endoplasmic reticulum, Golgi, or mitochondria in different species [10,15,94–97]. These findings are consistent with the notion that *de novo* ORFs could play a role in a range of transport processes, such as ion, amino acid, or protein transport across cellular membranes. By establishing a connection between predicted transmembrane domains and increased coexpression with transport-related genes, our findings set the stage for future experimental investigations into the precise molecular mechanisms and functional roles of nORFs in diverse transport systems.

Lastly, we explored how the preexisting regulatory context influences the transcriptional profiles of *de novo* ORFs. We found that *de novo* ORFs that piggyback off their neighboring conserved

ORFs' promoters had increases in expression, coexpression and biological process similarity with their neighboring conserved ORFs. Strikingly, ORFs that emerge *de novo* downstream of conserved ORFs have the largest increases in expression, coexpression and biological process similarities with their neighbors compared to other orientations, largely due to transcriptional readthrough leading to transcript sharing. Previous studies have shown that the transcription of regions downstream of genes is functional and regulated [98]. A 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 that play a role in maintaining nuclear stability during stress [99]. 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 [100]. Lastly, a study in yeast found that genes which are preferentially expressed as bicistronic transcripts tend to contain evolutionarily younger genes compared to adjacent genes that do not share transcripts, suggesting that transcript sharing could provide a route for novel ORFs to become established genes [101]. These findings together with our results suggest that genomic regions downstream of genes may provide the most favorable environment for the transcription of *de novo* ORFs.

Our analyses show that the likelihood of a *de novo* ORF being expressed or repressed under the same conditions as the neighboring conserved ORF is influenced by the extent to which it piggybacks on the neighboring ORF's regulatory context. Therefore, in addition to the evolutionary pressure acting on the sequence of emerging ORFs, our results suggest that transcriptional regulation and genomic context also influence their functional potential. However, this influence is not entirely deterministic, and much weaker when *de novo* ORFs emerge upstream than downstream of genes. Future studies are needed to map regulatory networks controlling nORF expression and reconstruct their evolutionary histories.

There are several limitations to our study. First, while SpQN enhances the coexpression signal of lowly expressed ORFs, it comes at the cost of reducing signals in highly expressed ORFs [62]. Given our objective of studying lowly-expressed nORFs this tradeoff is deemed worthwhile. Second, our study provides evidence of associations between nORFs and cellular processes such as homeostasis and transport, but these findings are based on transcription profile similarities which do not necessarily imply cotranslation or correlated protein abundances [102]. Furthermore, our analyses were performed in the yeast *S. cerevisiae* and the generalizability of our findings to other species requires further investigation.

Conclusions

In conclusion, our study represents a significant step forward towards the characterization of nORFs. We employed advanced statistical methods to account for low expression levels and generate a high-quality coexpression network. Despite being lowly expressed, nORFs are coexpressed with almost every cORF. We find that numerous nORFs form structured, noncanonical-only transcriptional modules which could be involved in regulating novel cellular processes. We find that many nORFs are coexpressed with genes involved in homeostasis and transport related processes, suggesting that these pathways are most likely to incorporate novel ORFs. Additionally, our investigation into the influence of genomic orientation on the expression and coexpression of *de novo* ORFs showed that ORFs located downstream of conserved ORFs are most influenced by the pre-existing regulatory environment at their locus of emergence. Our findings provide a foundation for future research to further elucidate the roles of nORFs and *de novo* ORFs in cellular processes and their broader implications in adaptation and evolution.

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 downloaded from SGD [103], which included 6,600 ORFs. Second, we utilized the translated ORF list from Wacholder et al. [14] reported in their *Supplementary Table 3.* We filtered to include cORFs (Verified, Uncharacterized or Transposable element genes) as well as any nORFs 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 using bedtools [104]. We removed ORFs that were exact sequence duplicates of another ORF. This left 5,878 cORFs and 18,636 nORFs, 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 quality controlled (nucleotides with Phred score < 20 at end of reads were trimmed) and adapters were removed using TrimGalore version 0.6.4 [105]. Samples were aligned to the transcriptome GTF file containing the ORFs defined above and quantified using Salmon [106] version 0.12.0 with 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 (Supplementary Data 1). 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, Supplementary Data 2).

Coexpression calculations

The raw counts were transformed using clr. Pairwise proportionality was calculated using ρ [69] for each ORF pair. Spatial quantile normalization (SpQN) [62] of the coexpression network was performed using the mean clr expression value for each ORF as confounders to correct for mean expression bias, which resulted in similar distributions of coexpression values across varying expression levels (Supplementary Figure 2). Only ORF pairs that had at least 400 samples expressing both ORFs (at raw >5) were included. This threshold was determined empirically as detailed below.

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 pros and cons, include removing all genes that contain any zeros, imputing the zeros, or adding a pseudo count to all genes [107,108]. 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 low read count compared to an ORF with a high read count [109]. For these reasons, all raw counts below 5 were set to NA prior to clr transformation. These observations were then excluded when calculating the clr transformation and in the ρ calculations. We used clr and ρ implementations in R package *Propr* [69] and implementation of SpQN from Wang et al. [62].

To determine the minimum number of samples needed expressing both ORFs in a pair, we determined the number of samples needed for coexpression values to converge within ρ ± 0.05 or ρ ± 0.1 for 2,167 nORF-cORF pairs which have a ρ > 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 (Supplementary Figure 1).

Comparing coexpression inference approaches

To compare our approach with a batch correction approach, we used clr to transform the expression matrix, followed by removing the top principal component (PC1) of the clr expression matrix to do batch correction using the function *removePrincipalComponents* from the *WGCNA* [70] R package. We then calculated ρ values and applied SpQN normalization. Additionally, we created a coexpression matrix based on TPM as well as RPKM normalized expression values instead of clr and calculated Pearson's correlation coefficient.

Protein Complex enrichments

We retrieved a manually curated list of 408 protein complexes in *S. cerevisiae* from the CYC2008 database by Pu et al. [64]. The coexpression matrix was filtered to contain only the 1,617 cORFs found in the CYC2008 database prior to creating the contingency table. Fisher's exact test was used to calculate the significance of the association between coexpression and protein complex formation. Coexpressed was defined as the 99.8th ρ percentile (ρ > 0.888) considering all ORF pairs in the coexpression matrix (n = 62,204,406 ORF pairs) for Figure 1C.

TF binding enrichments

A ChIP-exo dataset from Rossi et al. [65] containing DNA-binding information for 73 sequence-specific TFs across the whole genome was used. For each ORF we identified which TFs had binding within 200 bp upstream of the ORF's TSS. The TSSs for all ORFs in the coexpression matrix was determined by the median 5' transcript isoform (TIF) start positions using TIF-seq [66] dataset. Only ORFs found in the TIF-seq dataset were considered (n = 5,334 cORFs and 5,362 nORFs). To calculate the enrichments reported in Figures 1D, Supplementary Figure 5 and Supplementary Figure 7, 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 = 973 cORFs and 936 nORFs). Fisher's exact test was used to calculate the association between coexpression and having their promoters bound by a common TF. Coexpressed was defined as the 99.8th ρ percentile (ρ > 0.888) considering all ORF pairs in the coexpression matrix (n = 62,204,406 ORF pairs) for Figure 1D.

Coexpression matrix clustering

We used the weighted gene coexpression network analysis (*WGCNA*) package [70] in R to cluster our coexpression matrix. To do this, we first transformed our coexpression matrix into a weighted adjacency matrix by applying a soft thresholding which involved raising the coexpression matrix to the power of 12. This 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 full coexpression network. ORFs in the other four clusters were not included in the network as they did not pass the ρ threshold.

GO analysis of clusters

We downloaded GO trees (file: go-basic.obo) and annotations (files: sgd.gaf) from ref. [110]. We used the Python package, *GOATools* [111], 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 FDR with BH multiple testing correction [112] 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 cORFs in the cluster (n=54).

Network randomization and topology analyses

To create random networks while preserving the same degree distribution, we used an edge swapping method (Supplementary Figure 9). 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* [113] and networks were plotted using spring embedded layout [74] in Python package *networkx* [114].

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) calculates enrichments of an ordered list of genes given a biological annotation such as GO or KEGG. For each ORF in our dataset, we used ρ values to order annotated ORFs and provided this sorted set to *fgsea* [115] . We used the GO slim file downloaded from SGD [103] for GO annotations. We used *clusterProfiler* [116] R package to download KEGG annotations using KEGG REST API [78] on 1 April 2023 and then used *fgseaMultilevel* function in *fgsea* R package to calculate enrichments for both annotations individually. To calculate GO or KEGG terms that are enriched or depleted for nORFs compared to cORFs, we calculated the number of cORFs and nORFs that had GSEA enrichments at BH adjusted FDR < 0.01. Using these counts we calculated the proportion of nORFs and cORFs associated with a GO or KEGG 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 BH correction. Odds ratios were calculated by dividing proportion of nORFs to proportion of cORFs. Proportions for the GO terms with BH adjusted FDR < 0.001 and Odds ratio greater than 2 or less than 0.5 are plotted in Figures 3A-B and are reported in Supplementary Data 5 and proportions for KEGG terms are plotted in Supplementary Figure 11 and reported in Supplementary Data 6.

Transmembrane domain enrichment

Transmembrane domains were predicted using TMHMM 2.0 [75] for all nORFs. An ORF was classified as having a transmembrane domain if it was predicted to have at least one transmembrane domain. nORFs were classified as "coexpressed with transport-related genes" if the ORF had a GSEA enrichment at FDR < 0.01 with any of the 15 GO slim transport terms: transport, ion transport, amino acid transport, lipid transport, carbohydrate transport, regulation of transport, transmembrane transport, vacuolar transport, vesicle-mediated transport, endosomal transport, nucleobase-containing compound transport, Golgi vesicle transport, nucleocytoplasmic transport, nuclear transport, or cytoskeleton-dependent intracellular transport. Fisher's exact test was used to calculate the significance of association between transport-related processes and transmembrane domain.

Differential expression analysis for TF deletion and overrepresentation tests

For Hsf1 analysis, RNA-seq samples were from Ciccarelli et al. (SRA accession SRP437124) [76]. Hsf1 deletion strains were compared to wild type (WT) strains when exposed to heat shock conditions. For Sfp1 analysis, RNA-seq samples were from SRA accession SRP159150. In both cases, deletion strains were compared to WT strains. Differential expression was calculated using R package *DESeq2* [117], and ORFs were defined as differentially expressed if the log fold change (FC) in RNA expression between WT and control strains was greater than or less than 0.5 i.e. log(FC) > 0.5 or log(FC) < -0.5 and BH adjusted p-value < 0.05. ChIP-exo data for Hsf1 and Sfp1 binding was taken from Rossi et al. [65] and an ORF was labeled as having Hsf1 or Sfp1 binding if the TF was found within 200 bp upstream of the ORF's TSS. Fisher's exact test was performed to see if there is an association between an nORF in a GO biological process and being regulated by the TF. We define an nORF to be "in" a GO term if it has a GSEA enrichment for that GO term at FDR < 0.01. We defined an nORF as regulated by a TF if the nORF had evidence of the TF binding within 200 bp of the nORF's TSS in ChIP-exo and has significantly downregulated expression in the TF deletion RNA-seq samples compared to the

WT samples. BH p-value correction was performed for all GO terms tested. Significant GO terms and the associated regulated nORFs are reported in Supplementary Data 8.

Detection of homologs using BLAST

We obtained the genomes of 332 budding yeasts from Shen et al. [118]. To investigate the homology of each non overlapping ORF in our dataset, we used TBLASTN and BLASTP [119] 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., 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. [14]. 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 cORF 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 branches were less than 0.6, suggesting the absence of a homologous ORF in those two species. We identified conserved ORFs if a nonoverlapping cORF has an average RFC > 0.8 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 with e-value < 0.0001. Then for two overlapping ORFs, if one had RFC > 0.8 and the other had 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.

Calculation of GO term similarities

GO term similarities were calculated using the Relevance method developed in Schlicker et al. [82]. 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* [120] package in R that implements these similarity measures.

Termination factor binding analysis

ChIP-exo data for Pcf11 and Nrd1 termination factor binding sites are taken from Rossi et al. [65]. 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.

Determining shared promoters

To determine whether two ORFs shared a promoter, we reused the TIF-Seq dataset from Pelechano et al. [66]. TIF-Seq is a sequencing method that detects the boundaries of 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 for down same and up same pairs, as well as found TIFs with non-overlapping TSSs for up opposite *de novo*-conserved ORF pairs. 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 not share a promoter.

Web application

We utilized R language [121] and the shiny framework [73] to develop a web application which allows querying of ORFs in our dataset for information about their coexpression with other ORFs, network visualization, and GSEA enrichments. It can be accessed through a web browser and is available at https://carvunislab.csb.pitt.edu/shiny/coexpression/.

Acknowledgments

Figures 1A, 4A, 4C, and Supplementary Figure 10 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, Alistair Turcan, Dr. Nikolaos Vakirlis, and Dr. Maria Chikina 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. All authors approved the final version of the manuscript.

Funding

This work was supported by: the National Science Foundation 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.

Source code

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

Ethics Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

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

Supplementary Data

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

Supplementary Data 1: RNA-seq studies and samples used in this study. (CSV)

Supplementary Data 2: ORFs included in the coexpression matrix. (CSV)

Supplementary Data 3: Coexpression matrix generated in this study. (CSV)

Supplementary Data 4: GSEA analysis results for each ORF using GO BP annotations. (CSV)

Supplementary Data 5: List of GO BP terms that are more associated with nORFs than cORFs and statistics. (CSV)

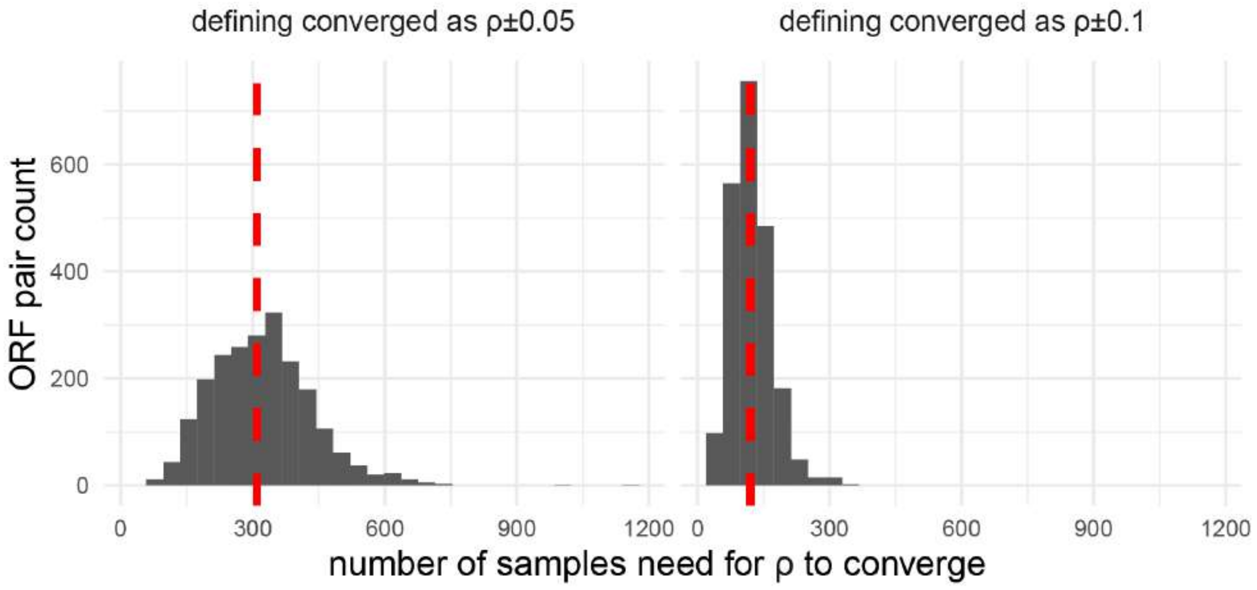
Supplementary Data 6: List of KEGG terms that are more associated with nORFs than cORFs and statistics. (CSV)

Supplementary Data 7: GSEA analysis results for each ORF using KEGG annotations. (CSV)

Supplementary Data 8: GO BP terms where nORFs are regulated by either Hsf1 or Sfp1 in GO BP terms are overrepresented. (CSV)

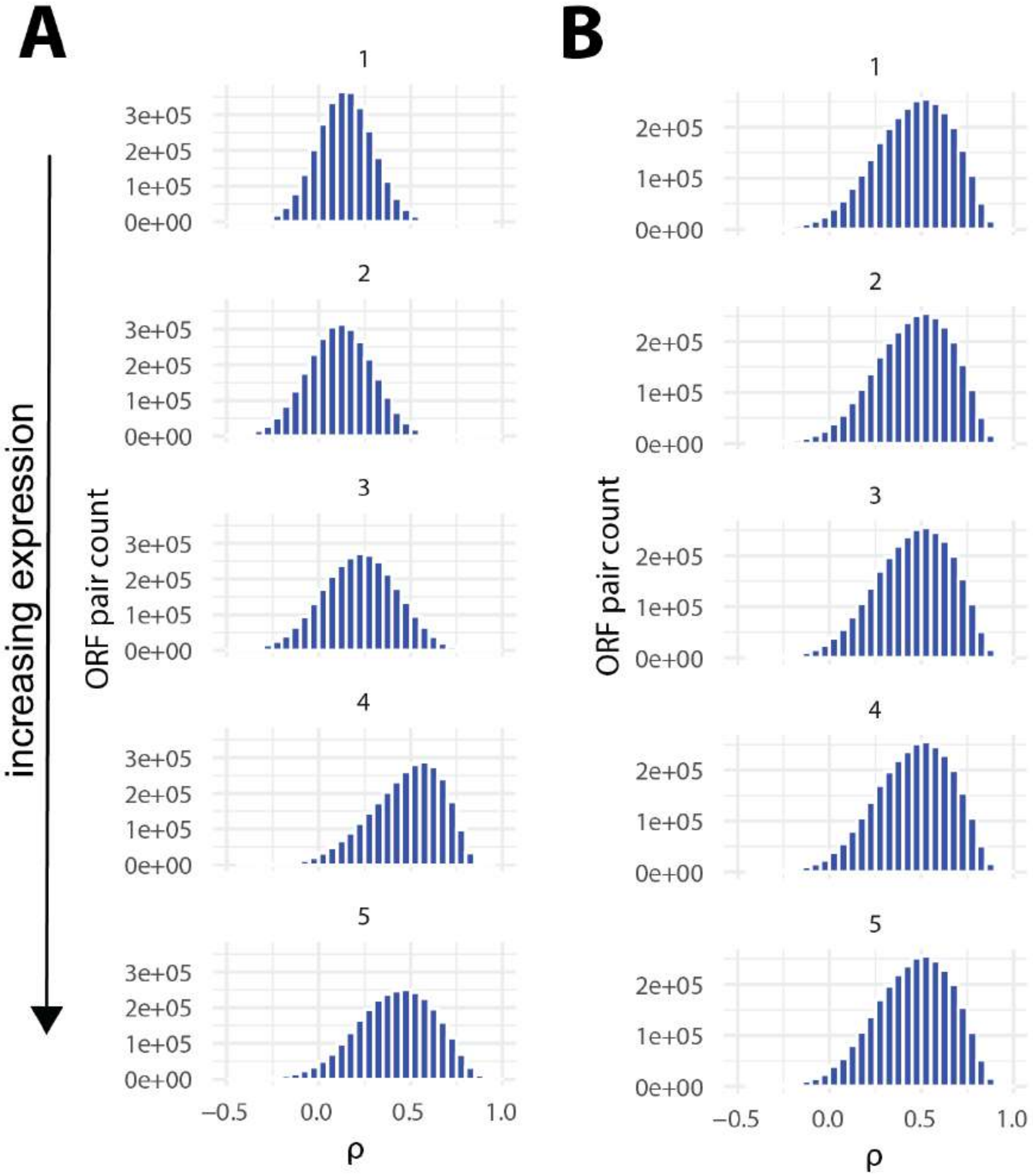
Supplementary Figures

Supplementary Figure 1



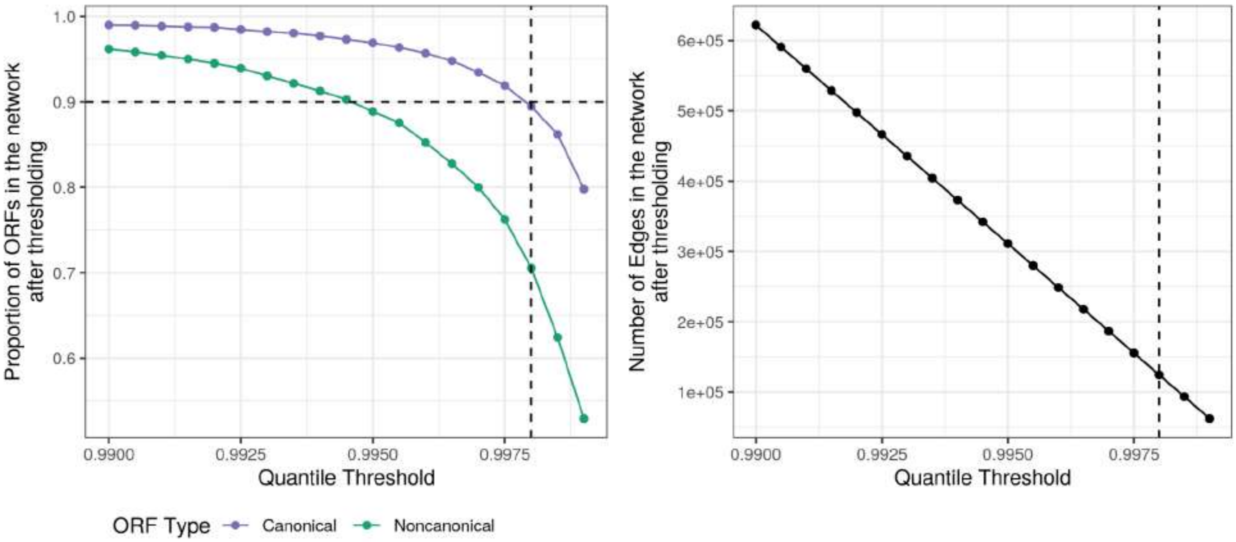
Supplementary Figure 1 To understand the effect of sample size on coexpression values and to determine how many samples is sufficient for ρ to converge, we recalculated coexpression for a given ORF pair using n = 2 samples through n = all samples. Fluctuations were calculated as max(ρ)-min(ρ) within 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



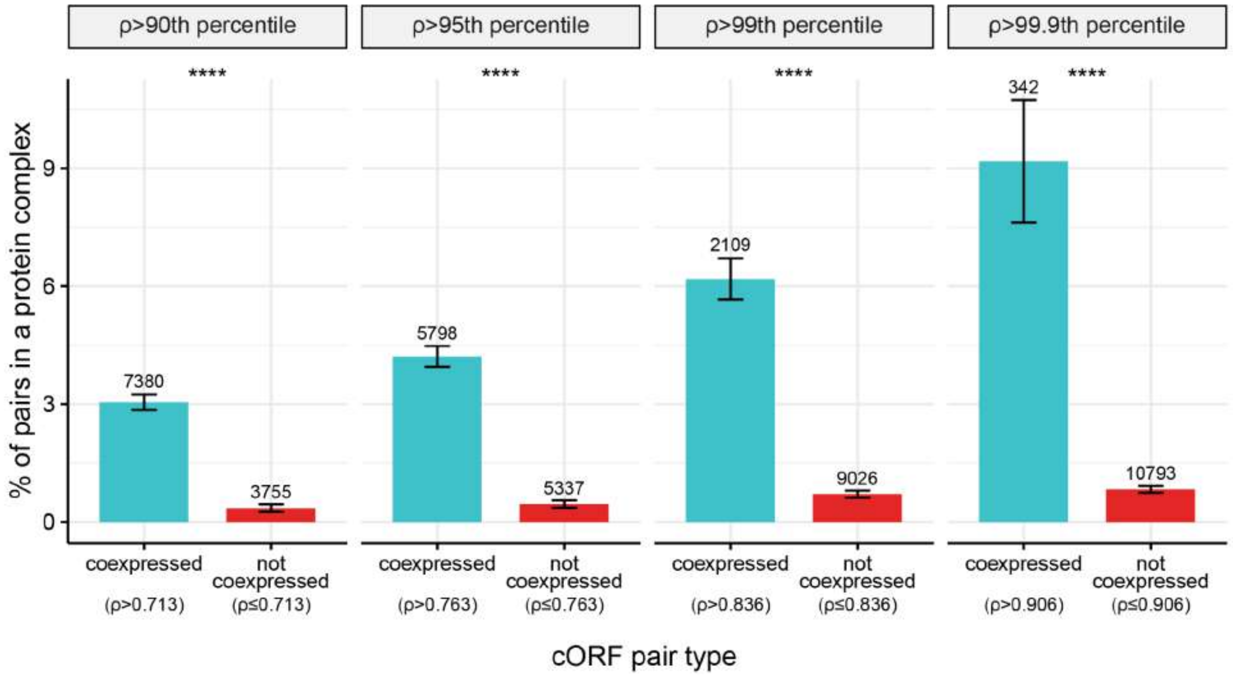
Supplementary Figure 2 Distribution of coexpression values (ρ) for ORF pairs binned by expression level, from lowly expressed pairs *top* to highly expressed pairs *bottom*, A) before spatial quantile normalization (SpQN) and B) after SpQN, which normalizes the coexpression values so that the distribution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

Supplementary Figure 3



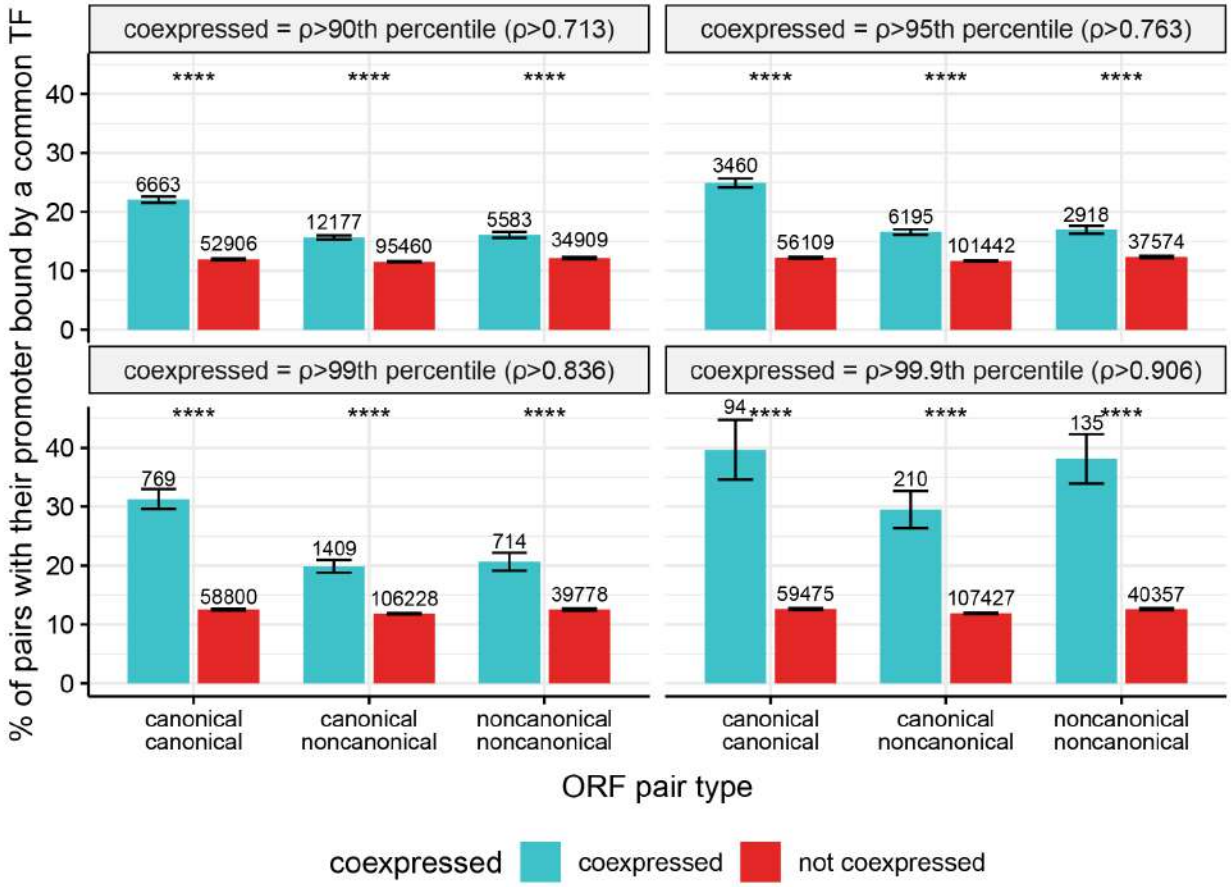
Supplementary Figure 3 Network threshold affects cORFs and nORFs differently. *Left* shows the proportion of cORFs or nORFs in the network at each quantile threshold and the *right* shows the number of connections in the network. Dashed line represents 0.9998 quantile which was chosen for creating the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

Supplementary Figure 4



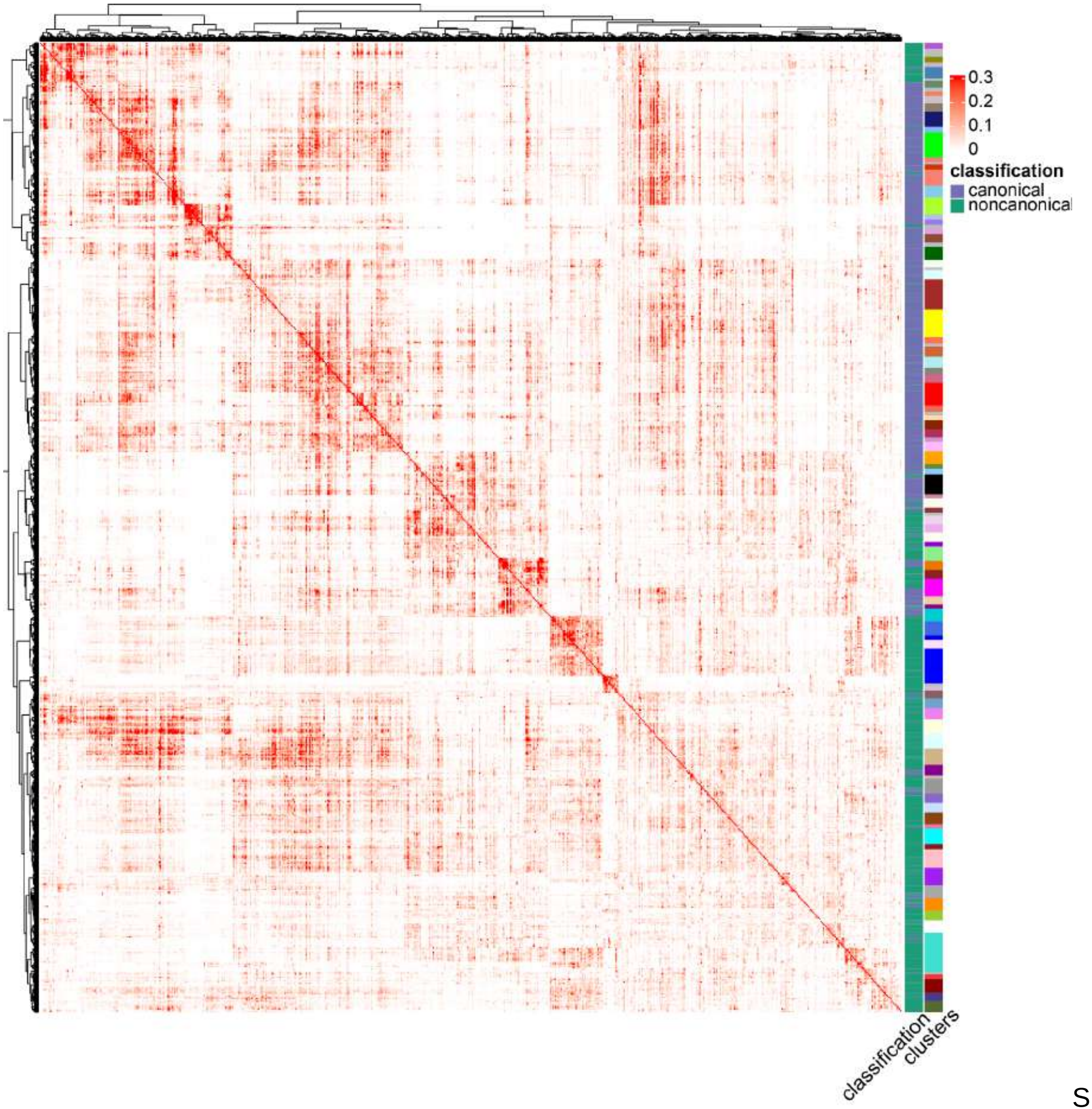
Supplementary Figure 4 Coexpressed cORFs pairs are more likely to encode proteins that form protein complexes than non-coexpressed cORF pairs, and this is consistent across different coexpression cutoffs. Coexpression was defined using the top 90th, 95th, 99th, and 99.9th percentile of all ORF pairs in the network (n = 62,204,406 ORF pairs). 90th percentile ($\rho >$ 0.713) Odds ratio = 8.89; 95th percentile ($\rho >$ 0.763) Odds ratio = 9.59; 99th percentile ($\rho >$ 0.836) Odds ratio = 9.23; 99.9th percentile ($\rho >$ 0.906) Odds ratio = 12.1; Fisher's exact test $p <$ 2.2e-16 for all comparisons. Numbers above bars represent the number of ORF pairs in each category. Error bars represent the standard error of the proportion. A list of 408 protein complexes were retrieved from Pu et al. CYC2008 database [64]. Enrichments were calculated using only the 1,617 cORFs found in the CYC2008 database.

Supplementary Figure 5



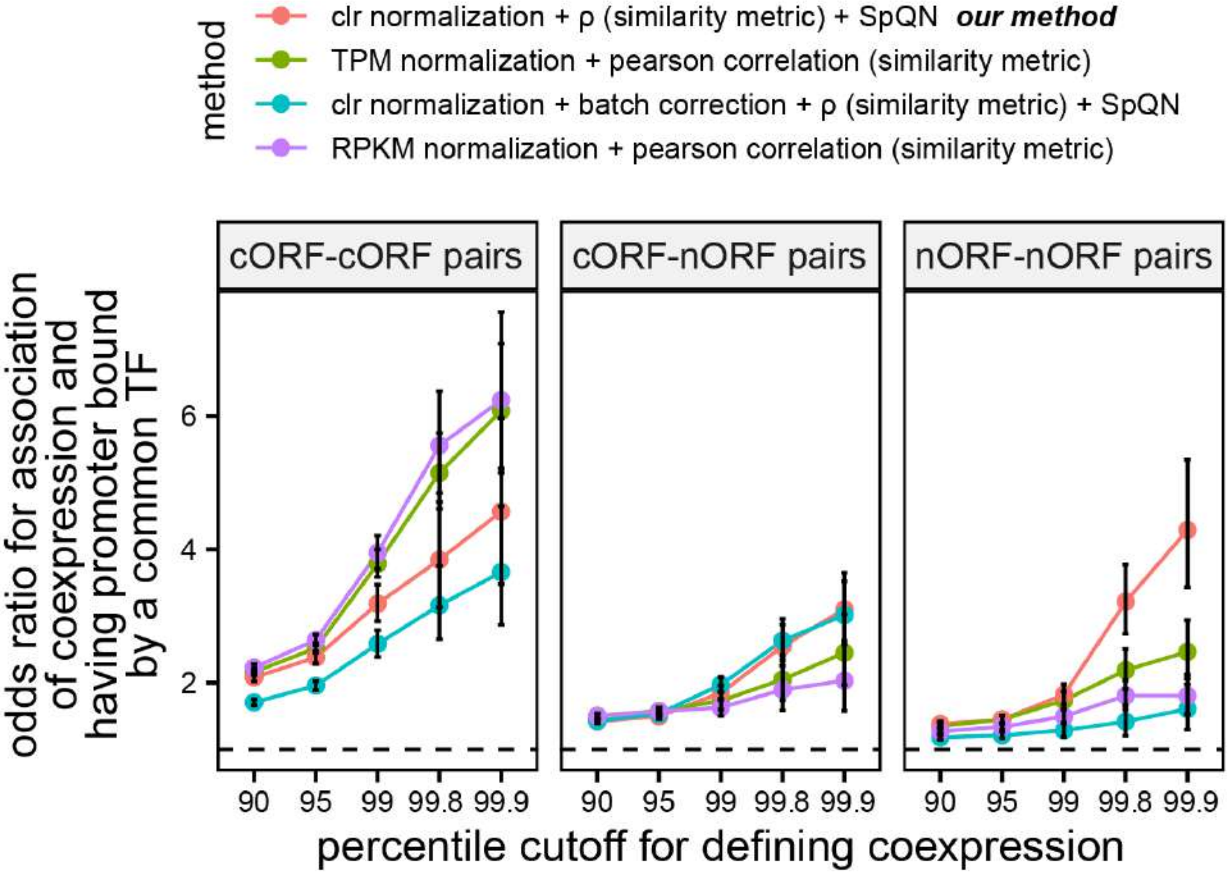
Supplementary Figure 5 Coexpressed ORF pairs are more likely to have their promoters bound by a common TF than non-coexpressed ORF pairs, and this is true across different coexpression cutoffs and for canonical-canonical (cc), canonical-noncanonical (cn) and noncanonical-noncanonical (nn) ORF pairs. Coexpression was defined using the top 90th, 95th, 99th, and 99.9th percentile of all ORF pairs in the network (n = 62,204,406 ORF pairs). 90th percentile (ρ > 0.713): cc Odds ratio = 2.08, cn Odds ratio = 1.42, nn Odds ratio = 1.38; 95th percentile (ρ > 0.763): cc Odds ratio = 2.38, cn Odds ratio = 1.50, nn Odds ratio = 1.45; 99th percentile (ρ > 0.836): cc Odds ratio = 3.19, cn Odds ratio = 1.85, nn Odds ratio = 1.82; 99.9th percentile (ρ > 0.906): cc Odds ratio = 4.57, cn Odds ratio = 3.10, nn Odds ratio = 4.29; ****: Fisher's exact test p < 2.2e-16 for all comparisons. Error bars represent the standard error of the proportion. Using a ChIP-exo dataset from Rossi et al. [65] containing DNA-binding information for 73 sequence-specific TFs, TF binding was defined as a ChIP-exo peak within 200 bp upstream of the ORF's TSS. Only ORFs whose promoter was bound by at least one TF were considered. Numbers above bars represent the number of ORF pairs in each category.

Supplementary Figure 6



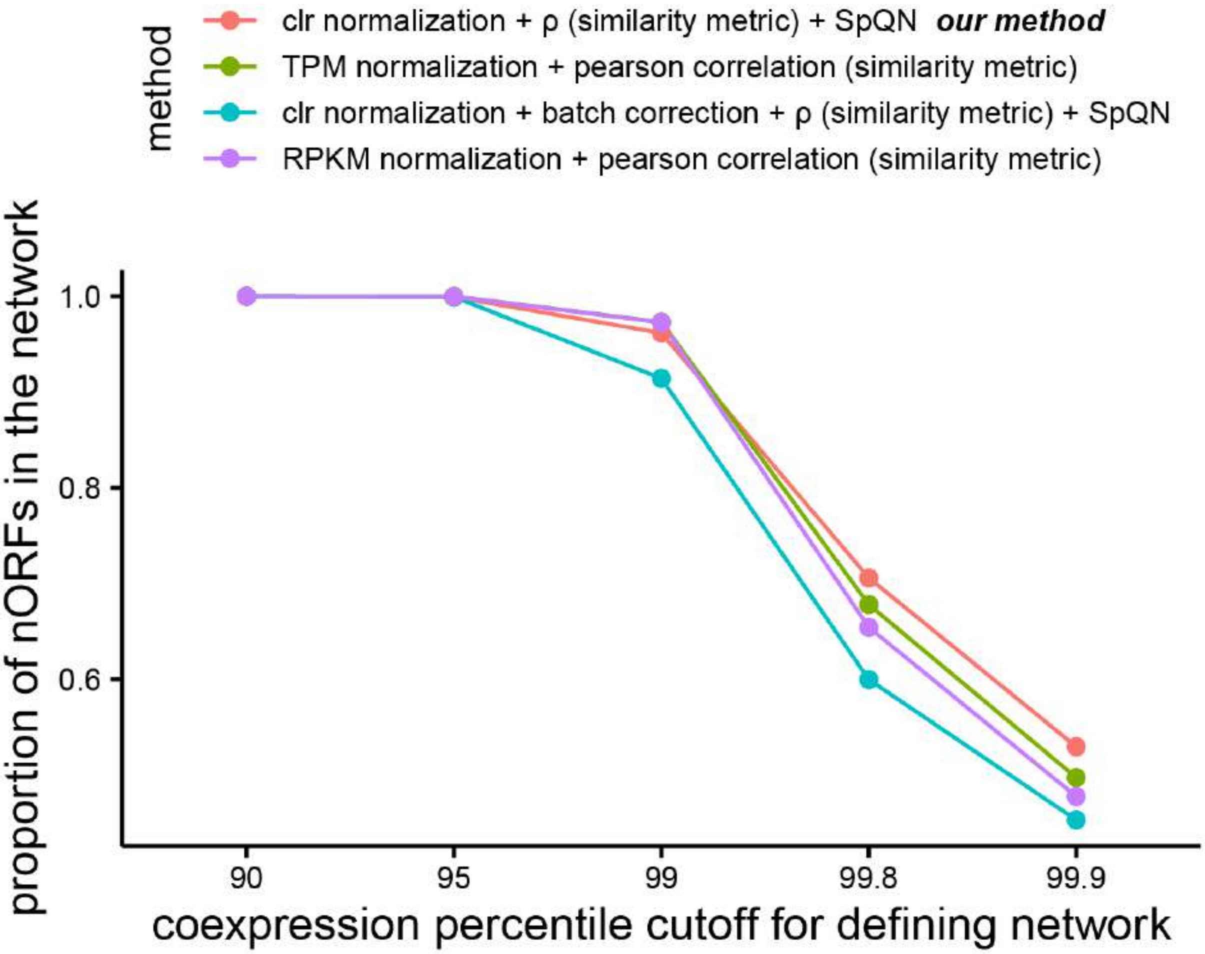
Supplementary Figure 6 Clustered matrix heatmap. Coexpression values are first transformed by taking power of 12 and then WGCNA pipeline [70] is applied. Clusters are determined by cutting dendrograms (see methods for details). Colors on 'clusters' section represent the different clusters. Values of 0.3 and above are represented by red to show the structure of the heatmap.

Supplementary Figure 7



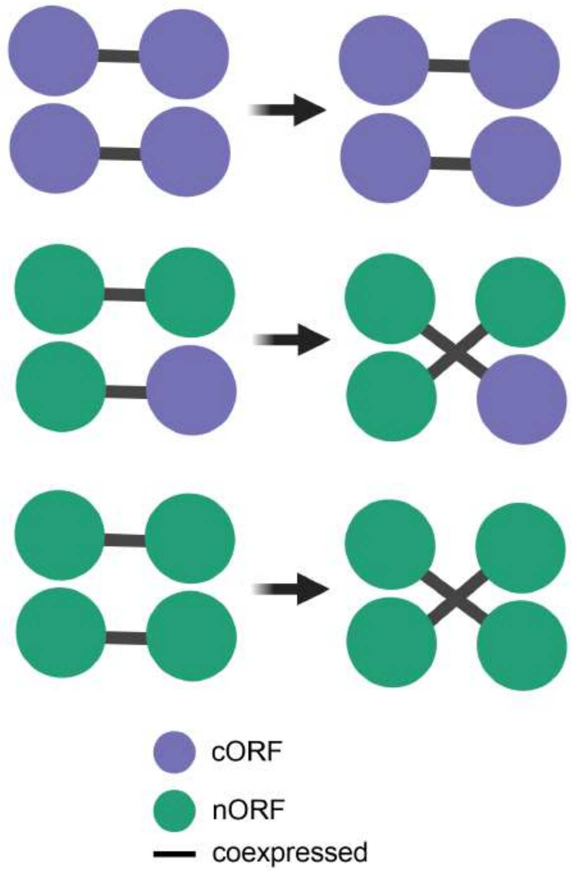
Supplementary Figure 7 Using clr normalization, ρ similarity metric and SpQN normalization leads to the highest odds ratios for nORF-nORF coexpressed pairs to also have their promoters bound by common TFs. Our method (*pink*) uses clr to transform the expression matrix, uses proportionality metric ρ to calculate coexpression and SpQN to normalize the coexpression matrix. Method TPM + pearson (*green*) uses TPM to normalize the expression matrix followed by Pearson correlation to calculate coexpression. Method clr + batch correction + rho + SpQN (*blue*) uses clr to transform the expression matrix, followed by removing the top principal component of the clr expression matrix to do batch correction, followed by calculating coexpression using proportionality metric ρ and SpQN normalization of the coexpression matrix. Method RPKM + pearson correlation (*purple*) uses RPKM to normalize the expression matrix followed by Pearson correlation to calculate coexpression. Coexpression percentiles were determined using all ORF pairs (n = 62,204,406 ORF pairs). All odds ratios are significant at p < 2.15e-5, Fisher exact test. Batch correction performed by removing the top principal component on the clr transformed expression matrix. Error bars represent the 95% confidence interval of the odds ratio. Dashed line shows an odds ratio of 1.

Supplementary Figure 8



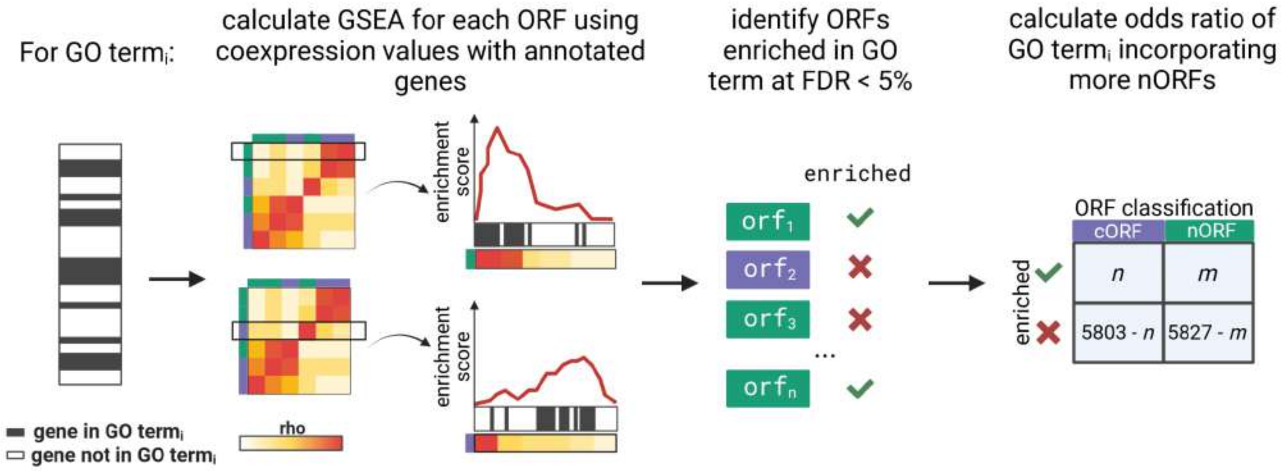
Supplementary Figure 8 Proportion of nORFs defined as coexpressed (and therefore included in the coexpression network) at various coexpression percentile cutoffs using four different methods. Our method (*pink*) uses clr to transform the expression matrix, uses proportionality metric ρ to calculate coexpression and SpQN to normalize the coexpression matrix. Method TPM + Pearson (*green*) uses TPM to normalize the expression matrix followed by Pearson correlation to calculate coexpression. Method ρ + batch correction (*blue*) uses clr to transform the expression matrix, followed by removing the top principal component of the clr expression matrix to do batch correction, followed by calculating coexpression using proportionality metric ρ and SpQN normalization of the coexpression matrix. Method RPKM + pearson correlation (*purple*) uses RPKM to normalize the expression matrix followed by Pearson correlation to calculate coexpression. Coexpression percentiles were determined using all ORF pairs (n = 62,204,406 ORF pairs).

Supplementary Figure 9



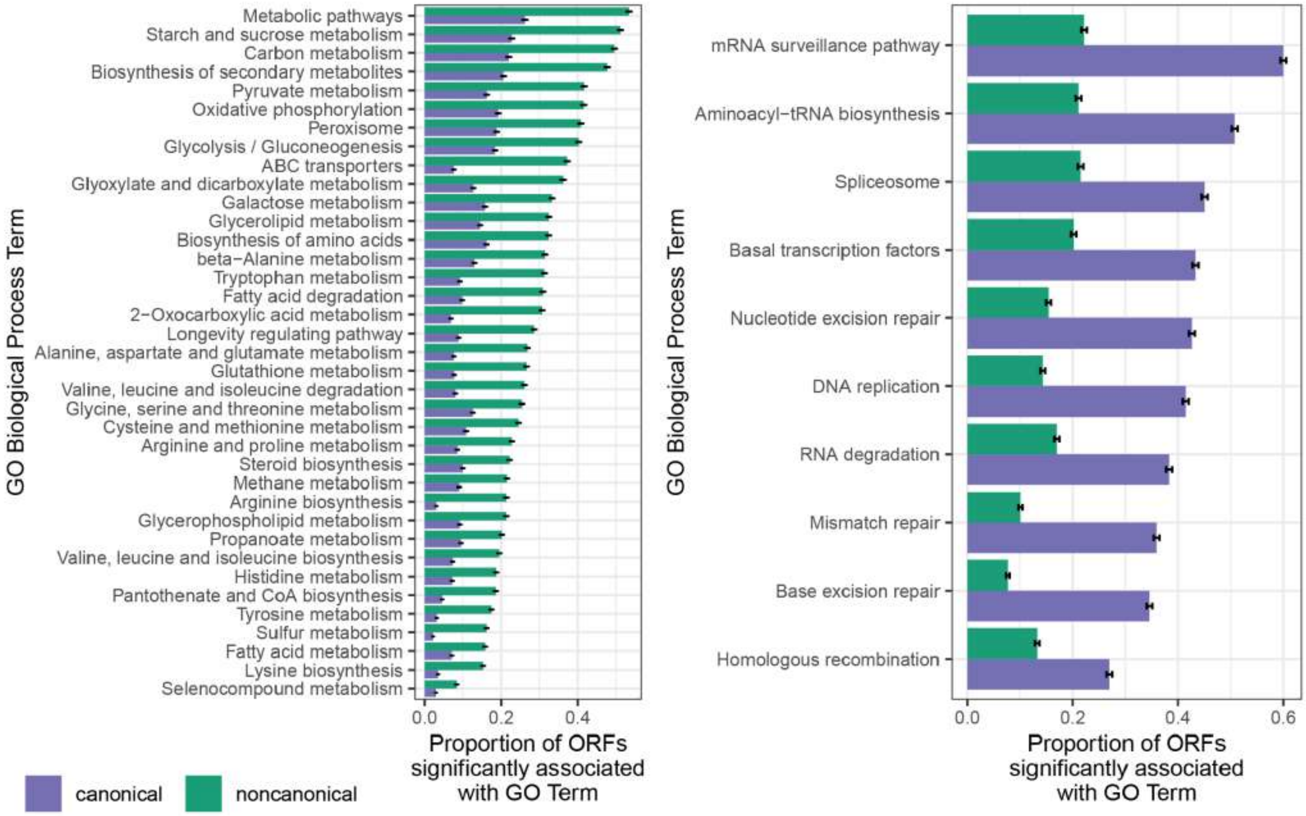
Supplementary Figure 9 Strategy for generating randomized networks. Edges between cORF-nORF and nORF-nORF pairs were swapped in a pairwise manner such that the degree of each node stayed the same. Edges between cORF-cORF pairs were not randomized.

Supplementary Figure 10



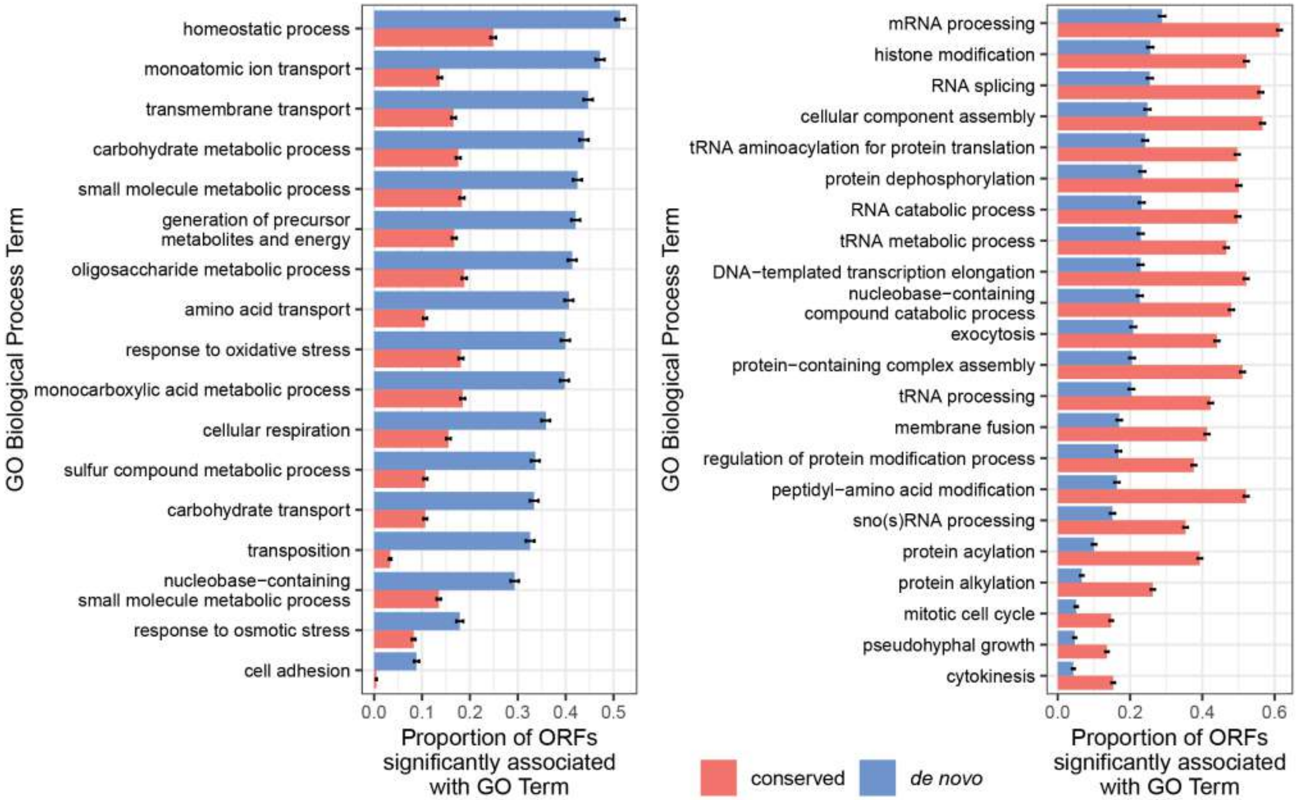
Supplementary Figure 10 GSEA pipeline using coexpression profiles to find GO terms that are more likely to incorporate nORFs.

Supplementary Figure 11



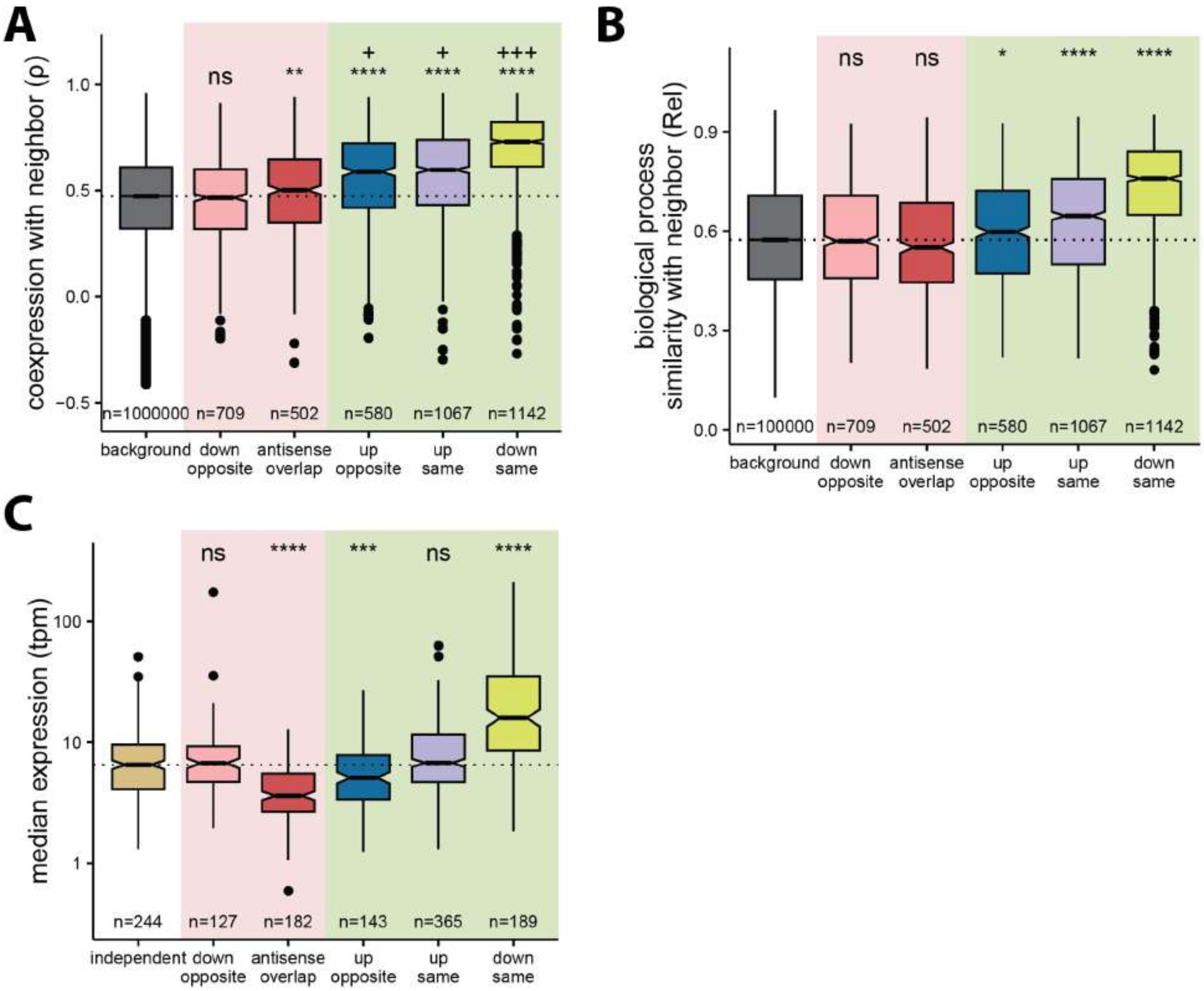
Supplementary Figure 11 KEGG pathways that proportionally have more (*left*) (Odds ratio > 2, n = 37 terms) or less (*right*) (Odds ratio < 0.5, n = 10 terms) GSEA enrichments with nORFs compared to cORFs (y-axis ordered by nORF enrichment proportion from highest to lowest, BH adjusted FDR < 0.001 for all terms, Fisher’s exact test). Error bars represent the standard error of the proportion.

Supplementary Figure 12



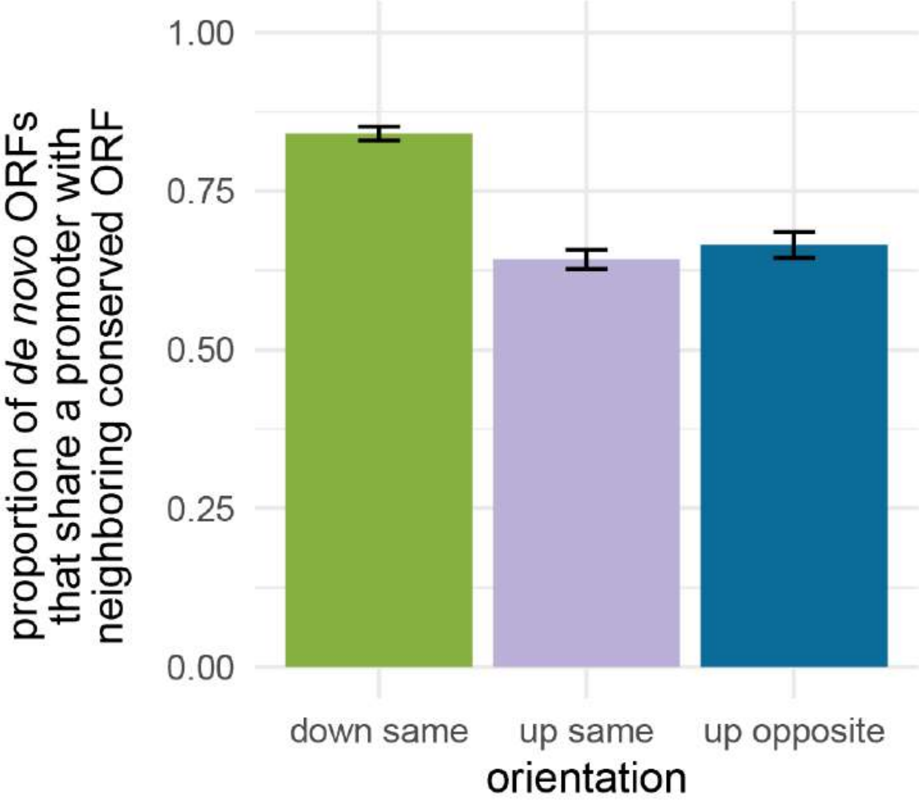
Supplementary Figure 12 GO terms that proportionally have more (*left*) (Odds ratio > 2, n = 35 terms) or less (*right*) (Odds ratio < 0.5, n = 11 terms) GSEA enrichments with *de novo* ORFs compared to conserved ORFs (y-axis ordered by *de novo* ORF enrichment proportion from highest to lowest, BH adjusted FDR < 0.001 for all terms, Fisher's exact test). Error bars represent the standard error of the proportion.

Supplementary Figure 13



Supplementary Figure 13 A) Coexpression (y-axis) of *de novo* ORFs with neighboring conserved ORFs per orientation (x-axis). Down same *de novo* ORFs tend to be highly coexpressed with their neighbors; background: *de novo*-conserved ORF pairs located on different chromosomes. B) Biological process similarity (y-axis) of *de novo* ORFs with neighboring conserved ORFs per orientation (x-axis). 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); background: *de novo*-conserved ORF pairs located on different chromosomes. C) Median expression of *de novo* ORFs (y-axis) per orientation (x-axis). *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 4D; independent: *de novo* ORFs located further than 500 bp from all conserved ORFs). For panels A-B-C: Mann-Whitney U-test, ****: p ≤ 0.0001, ***: p ≤ 0.001, **: p ≤ 0.01, *: p ≤ 0.05, ns: not-significant, +: small effect size (Cliff's d < 0.33), ++: medium effect size (Cliff's d < 0.474), +++: large effect size (Cliff's d ≥ 0.474); all orientations are compared to either background pairs (A, B) or to independent ORFs (C).

Supplementary Figure 14



Supplementary Figure 14 Proportion of *de novo* ORFs that share a promoter with their neighboring conserved ORF. To determine if ORFs shared a promoter with neighbors we used a publicly available TIF-seq dataset from Pelechano et al [65]. We defined down same or up same ORFs as sharing a promoter if they mapped to the same transcript at least once, and defined up opposite ORFs as sharing a promoter if their respective transcripts did not have overlapping TSSs. We found that 84% of down same (n = 174), 64% of up same (n = 368), and 66% of up opposite (n = 185) *de novo* ORFs share a promoter with their neighboring conserved ORF. Error bars represent the standard error of the proportion.

References

[1] Dujon B. The yeast genome project: what did we learn? Trends Genet TIG 1996;12:263–70. https://doi.org/10.1016/0168-9525(96)10027-5.

[2] Fisk DG, Ball CA, Dolinski K, Engel SR, Hong EL, Issel-Tarver L, et al. Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. Yeast Chichester Engl 2006;23:857–65. https://doi.org/10.1002/yea.1400.

[3] Basrai MA, Hieter P, Boeke JD. Small Open Reading Frames: Beautiful Needles in the Haystack. Genome Res 1997;7:768–71. https://doi.org/10.1101/gr.7.8.768.

[4] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 2008;320:1344–9. https://doi.org/10.1126/science.1158441.

[5] Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 2009;324:218–23. https://doi.org/10.1126/science.1168978.

[6] Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep 2014;8:1365–79. https://doi.org/10.1016/j.celrep.2014.07.045.

[7] Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and

evolutionary conservation. EMBO J 2014;33:981–93. https://doi.org/10.1002/embj.201488411.

[8] Couso J-P, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol 2017;18:575–89. https://doi.org/10.1038/nrm.2017.58.

[9] Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, et al. A hidden human proteome encoded by ‘non-coding’ genes. Nucleic Acids Res 2019;47:8111–25. https://doi.org/10.1093/nar/gkz646.

[10] Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science 2020;367:1140–6. https://doi.org/10.1126/science.aay0262.

[11] Orr MW, Mao Y, Storz G, Qian S-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res 2020;48:1029–42. https://doi.org/10.1093/nar/gkz734.

[12] Vitorino R, Guedes S, Amado F, Santos M, Akimitsu N. The role of micropeptides in biology. Cell Mol Life Sci 2021;78:3285–98. https://doi.org/10.1007/s00018-020-03740-3.

[13] Prensner JR, Enache OM, Luria V, Krug K, Clauser KR, Dempster JM, et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol 2021;39:697–704. https://doi.org/10.1038/s41587-020-00806-2.

[14] Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, et al. A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Syst 2023;14:363-381.e8. https://doi.org/10.1016/j.cels.2023.04.002.

[15] Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun 2020;11:781. https://doi.org/10.1038/s41467-020-14500-z.

[16] Arnoult N, Correia A, Ma J, Merlo A, Garcia-Gomez S, Maric M, et al. Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. Nature 2017;549:548–52. https://doi.org/10.1038/nature24023.

[17] Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, et al. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. Cell 2015;160:595–606. https://doi.org/10.1016/j.cell.2015.01.009.

[18] Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, et al. Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. Science 2013;341:1116–20. https://doi.org/10.1126/science.1238802.

[19] Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature 2017;541:228–32. https://doi.org/10.1038/nature21034.

[20] Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, et al. The translation of non-canonical open reading frames controls mucosal immunity. Nature 2018;564:434–8. https://doi.org/10.1038/s41586-018-0794-7.

[21] Bhatta A, Atianand M, Jiang Z, Crabtree J, Blin J, Fitzgerald KA. A Mitochondrial Micropeptide Is Required for Activation of the Nlrp3 Inflammasome. J Immunol 2020;204:428–37. https://doi.org/10.4049/jimmunol.1900791.

[22] Niu X, Zhang J, Zhang L, Hou Y, Pu S, Chu A, et al. Weighted Gene Co-Expression Network Analysis Identifies Critical Genes in the Development of Heart Failure After Acute Myocardial Infarction. Front Genet 2019;10. https://doi.org/10.3389/fgene.2019.01214.

[23] Wright BW, Yi Z, Weissman JS, Chen J. The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol 2021. https://doi.org/10.1016/j.tcb.2021.10.010.

[24] Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and *de novo* gene birth. Nature 2012;487:370–4. https://doi.org/10.1038/nature11184.

[25] Van Oss SB, Carvunis A-R. De novo gene birth. PLOS Genet 2019;15:e1008160. https://doi.org/10.1371/journal.pgen.1008160.

[26] Sandmann C-L, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell 2023;83:994-1011.e18. https://doi.org/10.1016/j.molcel.2023.01.023.

[27] Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol 2015;16:202. https://doi.org/10.1186/s13059-015-0772-4.

[28] Abrusán G. Integration of New Genes into Cellular Networks, and Their Structural Maturation. Genetics 2013;195:1407–17. https://doi.org/10.1534/genetics.113.152256.

[29] Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. Genome Biol 2010;11:R127. https://doi.org/10.1186/gb-2010-11-12-r127.

[30] 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 2016;1859:31–40. https://doi.org/10.1016/j.bbagrm.2015.07.017.

[31] Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang Y-C, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol 2018;19:208. https://doi.org/10.1186/s13059-018-1590-2.

[32] Xu H, Li C, Xu C, Zhang J. Chance promoter activities illuminate the origins of eukaryotic intergenic transcriptions. Nat Commun 2023;14:1826. https://doi.org/10.1038/s41467-023-37610-w.

[33] Schlötterer C. Genes from scratch – the evolutionary fate of de novo genes. Trends Genet 2015;31:215–9. https://doi.org/10.1016/j.tig.2015.02.007.

[34] Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science 2014;343:769–72. https://doi.org/10.1126/science.1248286.

[35] Zhuang X, Yang C, Murphy KR, Cheng C-HC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc Natl Acad Sci 2019;116:4400–5. https://doi.org/10.1073/pnas.1817138116.

[36] Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. PLOS Genet 2015;11:e1005721. https://doi.org/10.1371/journal.pgen.1005721.

[37] Vakirlis N, Vance Z, Duggan KM, McLysaght A. De novo birth of functional microproteins in the human lineage. Cell Rep 2022;41:111808. https://doi.org/10.1016/j.celrep.2022.111808.

[38] Majic P, Payne JL. Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. Mol Biol Evol 2020;37:1165–78. https://doi.org/10.1093/molbev/msz300.

[39] Ruiz-Orera J, Villanueva-Cañas JL, Albà MM. Evolution of new proteins from translated sORFs in long non-coding RNAs. Exp Cell Res 2020;391:111940. https://doi.org/10.1016/j.yexcr.2020.111940.

[40] Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, et al. Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral

LncRNAs in Primates. PLOS Genet 2015;11:e1005391. https://doi.org/10.1371/journal.pgen.1005391.

[41] Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A Molecular Portrait of De Novo Genes in Yeasts. Mol Biol Evol 2018;35:631–45. https://doi.org/10.1093/molbev/msx315.

[42] Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. ELife 2016;5:e09977. https://doi.org/10.7554/eLife.09977.

[43] Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res 2009;19:1752–9. https://doi.org/10.1101/gr.095026.109.

[44] Ebisuya M, Yamamoto T, Nakajima M, Nishida E. Ripples from neighbouring transcription. Nat Cell Biol 2008;10:1106–13. https://doi.org/10.1038/ncb1771.

[45] Ghanbarian AT, Hurst LD. Neighboring Genes Show Correlated Evolution in Gene Expression. Mol Biol Evol 2015;32:1748–66. https://doi.org/10.1093/molbev/msv053.

[46] 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 2015;4:e08890. https://doi.org/10.7554/eLife.08890.

[47] Li J, Singh U, Arendsee Z, Wurtele ES. Landscape of the Dark Transcriptome Revealed Through Re-mining Massive RNA-Seq Data. Front Genet 2021;12.

[48] O’Meara TR, O’Meara MJ. DeORFanizing Candida albicans Genes using Coexpression. MSphere 2021;6:e01245-20. https://doi.org/10.1128/mSphere.01245-20.

[49] Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Mol Cell 2022;82:2885-2899.e8. https://doi.org/10.1016/j.molcel.2022.06.023.

[50] Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, et al. A Gene Expression Map for *Caenorhabditis elegans*. Science 2001;293:2087–92. https://doi.org/10.1126/science.1061603.

[51] Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 2003;302:249–55. https://doi.org/10.1126/science.1087447.

[52] Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun 2014;5:3231. https://doi.org/10.1038/ncomms4231.

[53] Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 2011;474:380–4. https://doi.org/10.1038/nature10110.

[54] Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 2013;500:593–7. https://doi.org/10.1038/nature12364.

[55] 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 2020;48:W566–71. https://doi.org/10.1093/nar/gkaa348.

[56] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2018;19:575–92. https://doi.org/10.1093/bib/bbw139.

[57] Yin W, Mendoza L, Monzon-Sandoval J, Urrutia AO, Gutierrez H. Emergence of co-expression in gene regulatory networks. PLOS ONE 2021;16:e0247671. https://doi.org/10.1371/journal.pone.0247671.

[58] Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, et al. Small open reading frames associated with morphogenesis are hidden in plant

genomes. Proc Natl Acad Sci 2013;110:2395–400. https://doi.org/10.1073/pnas.1213958110.

[59] Bashir K, Hanada K, Shimizu M, Seki M, Nakanishi H, Nishizawa NK. Transcriptomic analysis of rice in response to iron deficiency and excess. Rice 2014;7:18. https://doi.org/10.1186/s12284-014-0018-1.

[60] Stiens J, Tan YY, Joyce R, Arnvig KB, Kendall SL, Nobeli I. Using a Whole Genome Co-expression Network to Inform the Functional Characterisation of Predicted Genomic Elements from Mycobacterium tuberculosis Transcriptomic Data 2022:2022.06.22.497203. https://doi.org/10.1101/2022.06.22.497203.

[61] Li H, Xiao L, Zhang L, Wu J, Wei B, Sun N, et al. FSPP: A Tool for Genome-Wide Prediction of smORF-Encoded Peptides and Their Functions. Front Genet 2018;9. https://doi.org/10.3389/fgene.2018.00096.

[62] Wang Y, Hicks SC, Hansen KD. Addressing the mean-correlation relationship in co-expression analysis. PLOS Comput Biol 2022;18:e1009954. https://doi.org/10.1371/journal.pcbi.1009954.

[63] Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Exploiting single-cell expression to characterize co-expression replicability. Genome Biol 2016;17:101. https://doi.org/10.1186/s13059-016-0964-6.

[64] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 2009;37:825–31. https://doi.org/10.1093/nar/gkn1005.

[65] Rossi MJ, Kuntala PK, Lai WKM, Yamada N, Badjatia N, Mittal C, et al. A high-resolution protein architecture of the budding yeast genome. Nature 2021;592:309–14. https://doi.org/10.1038/s41586-021-03314-8.

[66] Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 2013;497:127–31. https://doi.org/10.1038/nature12121.

[67] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 2012;40:D700–5. https://doi.org/10.1093/nar/gkr1029.

[68] Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. Nat Methods 2019;16:381–6. https://doi.org/10.1038/s41592-019-0372-4.

[69] Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci Rep 2017;7:16252. https://doi.org/10.1038/s41598-017-16520-0.

[70] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

[71] Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics 2015;31:2123–30. https://doi.org/10.1093/bioinformatics/btv118.

[72] Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks. Genome Biol 2019;20:94. https://doi.org/10.1186/s13059-019-1700-9.

[73] Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. shiny: Web application framework for R. 2023.

[74] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw Pract Exp 1991;21:1129–64. https://doi.org/10.1002/spe.4380211102.

[75] Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 2001;305:567–80. https://doi.org/10.1006/jmbi.2000.4315.

[76] Ciccarelli M, Masser AE, Kaimal JM, Planells J, Andréasson C. Genetic inactivation of essential HSF1 reveals an isolated transcriptional stress response selectively induced by protein misfolding 2023:2023.05.05.539545. https://doi.org/10.1101/2023.05.05.539545.

[77] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.

[78] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62. https://doi.org/10.1093/nar/gkv1070.

[79] Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet 2007;39:683–7. https://doi.org/10.1038/ng2012.

[80] Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, et al. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci 2004;101:14315–22. https://doi.org/10.1073/pnas.0405353101.

[81] Masser AE, Kang W, Roy J, Mohanakrishnan Kaimal J, Quintana-Cordero J, Friedländer MR, et al. Cytoplasmic protein misfolding titrates Hsp70 to activate nuclear Hsf1. ELife 2019;8:e47791. https://doi.org/10.7554/eLife.47791.

[82] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 2006;7:302. https://doi.org/10.1186/1471-2105-7-302.

[83] Wei W, Pelechano V, Järvelin AI, Steinmetz LM. Functional consequences of bidirectional promoters. Trends Genet 2011;27:267–76. https://doi.org/10.1016/j.tig.2011.04.002.

[84] Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun 2020;11:6141. https://doi.org/10.1038/s41467-020-19921-4.

[85] Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, et al. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat Commun 2021;12:604. https://doi.org/10.1038/s41467-021-20911-3.

[86] Khitun A, Ness TJ, Slavoff SA. Small open reading frames and cellular stress responses. Mol Omics 2019;15:108–16. https://doi.org/10.1039/C8MO00283E.

[87] Wilson BA, Masel J. Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. Genome Biol Evol 2011;3:1245–52. https://doi.org/10.1093/gbe/evr099.

[88] Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the de novo-originated gene MDF1. Sci Rep 2014;4. https://doi.org/10.1038/srep07280.

[89] Frumkin I, Laub MT. Selection of a de novo gene that can promote survival of E. coli by modulating protein homeostasis pathways 2023:2023.02.07.527531. https://doi.org/10.1101/2023.02.07.527531.

[90] Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res 2010;20:408–20. https://doi.org/10.1038/cr.2010.31.

[91] Pagé N, Gérard-Vincent M, Ménard P, Beaulieu M, Azuma M, Dijkgraaf GJP, et al. A Saccharomyces cerevisiae Genome-Wide Mutant Screen for Altered Sensitivity to K1 Killer Toxin. Genetics 2003;163:875–94. https://doi.org/10.1093/genetics/163.3.875.

[92] Tassios E, Nikolaou C, Vakirlis N. Intergenic Regions of Saccharomycotina Yeasts are Enriched in Potential to Encode Transmembrane Domains. Mol Biol Evol 2023;40:msad059. https://doi.org/10.1093/molbev/msad059.

[93] Peng J, Zhao L. The origin and structural evolution of de novo genes in Drosophila 2023:2023.03.13.532420. https://doi.org/10.1101/2023.03.13.532420.

[94] Kesner JS, Chen Z, Aparicio AA, Wu X. A unified model for the surveillance of translation in diverse noncoding sequences 2022:2022.07.20.500724. https://doi.org/10.1101/2022.07.20.500724.

[95] Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nat Chem Biol 2013;9:59–64. https://doi.org/10.1038/nchembio.1120.

[96] Zhang S, Reljić B, Liang C, Kerouanton B, Francisco JC, Peh JH, et al. Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat Commun 2020;11:1312. https://doi.org/10.1038/s41467-020-14999-2.

[97] Leong AZ-X, Lee PY, Mohtar MA, Syafruddin SE, Pung Y-F, Low TY. Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. J Biomed Sci 2022;29:19. https://doi.org/10.1186/s12929-022-00802-5.

[98] Mayr C. What Are 3′ UTRs Doing? Cold Spring Harb Perspect Biol 2019;11:a034728. https://doi.org/10.1101/cshperspect.a034728.

[99] Vilborg A, Passarelli MC, Yario TA, Tycowski KT, Steitz JA. Widespread Inducible Transcription Downstream of Human Genes. Mol Cell 2015;59:449–61. https://doi.org/10.1016/j.molcel.2015.06.016.

[100] Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N, Bazzini AA. Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J 2020;39:e104763. https://doi.org/10.15252/embj.2020104763.

[101] Wu B, Cox MP. Characterization of Bicistronic Transcription in Budding Yeast. MSystems 2021;6:e01002-20. https://doi.org/10.1128/mSystems.01002-20.

[102] Kustatscher G, Grabowski P, Rappsilber J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol 2017;13:937. https://doi.org/10.15252/msb.20177548.

[103] Saccharomyces Genome Database | SGD n.d. https://www.yeastgenome.org/ (accessed January 20, 2021).

[104] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

[105] Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, et al. FelixKrueger/TrimGalore 2023. https://doi.org/10.5281/zenodo.7598955.

[106] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 2017;14:417–9. https://doi.org/10.1038/nmeth.4197.

[107] Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 2017;18:59. https://doi.org/10.1186/s13059-017-1188-0.

[108] L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016;17:75. https://doi.org/10.1186/s13059-016-0947-7.

[109] Lovell DR, Chua X-Y, McGrath A. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. NAR Genomics Bioinforma 2020;2:lqaa040. https://doi.org/10.1093/nargab/lqaa040.

[110] Gene Ontology Resource. Gene Ontol Resour n.d. http://geneontology.org/ (accessed March 10, 2022).

[111] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 2018;8:1–17. https://doi.org/10.1038/s41598-018-28948-z.

[112] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 1995;57:289–300.

[113] Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. InterJournal 2005;Complex Systems:1695.

[114] Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proc. 7th Python Sci. Conf., Pasadena, CA USA: 2008, p. 11–5.

[115] Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis 2021:060012. https://doi.org/10.1101/060012.

[116] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141.

[117] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.

[118] Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 2018;175:1533-1545.e20. https://doi.org/10.1016/j.cell.2018.10.023.

[119] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.

[120] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26:976–8. https://doi.org/10.1093/bioinformatics/btq064.

[121] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.