Landscape of the Dark Transcriptome Revealed through Re-mining Massive RNA-Seq Data

Jing Li[1,3,4], Urminder Singh[2,3,4], Zebulun Arendsee[2,3,4] and Eve Syrkin Wurtele[1,2,3,4*]

[1] *Genetics and Genomics Graduate Program, Iowa State University, Ames, 541004, USA*

[2] *Bioinformatics and Computational Biology Program, Iowa State University, Ames, 50014, USA*

[3] *Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, 50014, USA*

[4] *Center for Metabolic Biology, Iowa State University, Ames, 50014, USA*

[*] Corresponding author.

E-mail: evewurtele@gmail.com (Wurtele ES).

Jing Li: jingli@iastate.edu, https://orcid.org/0000-0003-0761-2977

Urminder Singh: usingh@iastate.edu, https://orcid.org/0000-0003-3703-0820

Zebulun Arendsee: zbwrnz@gmail.com, https://orcid.org/0000-0002-5833-798X

Eve S Wurtele: evewurtele@gmail.com, https://orcid.org/0000-0003-1552-9495

Running title: *Li J et al / Landscape of the Dark Transcriptome Revealed through Re-mining Massive RNA-Seq Data*

Total words: 5643, figures: 9, table:1, 15 supplementary figures and 5 supplementary tables in a single supplementary file. Links to other large supplementary data and code.

Abstract

The "dark transcriptome" can be considered the multitude of sequences that are transcribed but not annotated as genes. We evaluated expression of 6,692 annotated genes and 29,354 unannotated ORFs in the *Saccharomyces cerevisiae* genome across diverse environmental, genetic and developmental conditions (3,457 RNA-Seq samples). Over 48% of the transcribed ORFs have translation evidence. Phylostratigraphic analysis infers most of these transcribed ORFs would encode species-specific proteins ("orphan-ORFs"); hundreds have mean expression comparable to annotated genes. These data reveal unannotated ORFs most likely to be protein-coding genes. We partitioned a co-expression matrix by Markov Chain Clustering; the resultant clusters contain 2,468 orphan-ORFs. We provide the aggregated RNA-Seq yeast data with extensive metadata as a project in MetaOmGraph, a tool designed for interactive analysis and visualization. This approach enables reuse of public RNA-Seq data for exploratory discovery, providing a rich context for experimentalists to make novel, experimentally-testable hypotheses about candidate genes.

KEYWORDS: orphan gene, *de novo*, RNA-Seq, Ribo-Seq, gene function, cluster analysis

Introduction

Pervasive transcription of unannotated genome sequence in eukaryotic species is evidenced in multiple RNA-Seq studies. [1–5]. Indeed, transcription and translation has been described for non-genic regions of genomes in diverse species [6–15]. Many studies have dismissed this expression as transcriptional "noise" [4, 16–18]. However, several functional genes have been identified from the so-called ”noise” [19, 20]. This mass of unannotated transcripts, often ignored and little understood, we refer to as the “dark transcriptome” (Figure 1.A).

Each organism contains species-specific genes (denoted here as "orphan genes"). The challenge of distinguishing orphan genes in genomes and predicting their functions is immense, resulting in an under-appreciation of their importance. The emergence of novel protein coding genes specific to a single species (orphans) is a vital mechanism that allows organisms to survive a changing environment [21, 7, 22–25]. Over generations, those orphan genes that continue to provide a survival advantage will be maintained. Orphan genes can be identified from within a list of genes by phylostratigraphy, the classification of each gene according to its inferred age of emergence [21, 22]. Two general mechanisms enable orphan gene emergence:1) *de novo* evolution and 2) divergence.

Orphan genes can evolve *de novo* from non-coding sequence in regions of the genome lacking genes entirely or as new reading frames within existing genes [11, 22, 26, 27]. Indeed, transcriptional and translational "noise" has been suggested as a mechanism that facilitates novel gene emergence [28–33]. This hypothesis is borne out by *in vitro* and *in vivo* synthetic biology research demonstrating that novel peptides are often able to bind small molecules (e.g., ATP, and metals) [34] and induce beneficial phenotypes when expressed [34, 35]. If information on the expression of the dark transcriptome was more easily accessible, the potential roles of expressed transcripts could be better considered.

Orphan genes can also evolve from existing proteins by divergence of protein coding sequences (CDSs) beyond recognition [22, 28, 29, 31, 33, 36–39]. We estimate from the phylostratigraphic data on yeast genes that this process would require ultra-rapid sequence divergence relative to that of the average protein. Evolution of orphan genes from existing protein-coding genes has been estimated to account for about 18% (human), 25% (Drosophila), and 45% (yeast) of annotated taxonomically-restricted genes [39]. (This estimate considers only the ~50%

of yeast genes that can be compared across species, i.e., those that are located within syntenic intervals of related genomes [40].)

A systematic analysis of current computational methods for genome annotation indicates many orphan genes may be missed in annotation projects [41]. This is because genes are often identified from sequenced genomes by combining evidence based on homology with other species [42, 43] with *ab initio* machine-learning predictions by detecting canonical sequence motifs (e.g., splice junctions) [44, 45]. However, homology and *ab initio* approaches can be problematic in predicting orphan genes. First, orphan genes cannot be identified by homology to genes of other species, since they have none. Secondly, to the extent that an orphan has not yet evolved canonical motifs, *ab initio* prediction may be ineffective. For example, compared to the gold-standard annotations in the curated TAIR community database [46], the popular *ab initio* pipeline *MAKER* [44] predicted as few as 11% of the annotated *Arabidopsis* orphan genes, depending on the RNA-Seq evidence supplied [41].

Enhancing *ab initio* pipelines by other sequence-based information (e.g., motif/domain information, cellular location predictions, predicted isoelectric point (pI), genomics context) can improve gene predictions [47–49]. However, because it is not a given that newly evolved genes have canonical features, direct alignment of transcriptomic and/or proteomic data to the genome is critical for annotating orphan genes, as well as non-coding transcripts (lncRNAs, etc.) [3, 5, 7, 10, 32, 41, 48, 50].

Here, we reuse and re-mine aggregated RNA-Seq data to discover new potential gene candidates. The study comprehensively evaluates transcription and ribosomal binding of all open reading frames (ORFs) in the yeast genome over a wide variety of conditions, in the context of annotated genes. The research extends the results of previous studies, in that it globally represents ORFs in the *S. cerevisiae* genome across thousands of samples. Furthermore, we provide these data and extensive metadata via a biologist-friendly platform, MetaOmGraph (MOG [51], https://github.com/urmi-21/MetaOmGraph), which provides interactive, exploratory analysis [52] and visualization of expression levels, expression conditions, and co-expressed genes for the ORF-containing transcripts. This approach enables experimentalists to prioritize ORFs for functional characterization, and to logically define experimental parameters for these characterizations [51].

Results

Identifying potential cryptic orphan genes in *S. cerevisiae*

S. cerevisiae has the most extensively sequenced and annotated genome within the *Saccharomyces* genus, or perhaps across eukaryotes. However, despite the large body of research on *S. cerevisiae*, this genome expresses many transcripts not annotated as genes [3, 7, 9, 50, 53, 54], some of *de novo* origin [7, 26, 27, 39, 55], some supported with translational evidence [27, 50]. Our overall goal was to generate a comprehensive overview of expression of ORFs, and make this available in a format that can be readily explored. For this study, we classified all unannotated ORFs (>150 nt) and *Saccharomyces* genome database (SGD)-annotated genes in the *S. cerevisiae* genome according to phylostrata, transcription and translation evidence, and genomic context. We also included yeast ORFs < 150nt with transcription and/or translation evidence that had been characterized in two previous publications: smORFs [7] and txORFs [3]. Figure 1.B defines our terminology and lists the numbers of genes and ORFs in each category.

We inferred the oldest phylostratum (PS [56]) to which each *S. cerevisiae* protein (or candidate protein) could be traced, using the customizable *phylostratr* package [40] (Figure S1). Similarity to proteins of cellular organisms (i.e., proteins tracing back to prokaryotes) was designated as PS=1; no similarity to any protein outside of *S. cerevisiae* was designated as PS=15. (See supplementary file, *S.cerevisiae_RNA-seq_3457_27.mog* for full PS assignments by transcript). This analysis infers that fewer than 4% of SGD-annotated genes are orphans. In contrast, 54% of unannotated ORFs are orphans ("orphan-ORFs"), 40% are genus-specific (PS=10-14), and only 6% are more highly conserved (PS=1-9) (Figure 1.B).

In fungi, plants, and animals, the mean lengths of CDSs of annotated genes increase during evolution, with CDSs of orphan genes being the shortest [23, 27, 40, 57, 58] (Figure S2.A). The ORFs of yeast also follow a similar trend: average lengths of orphan-ORFs are shorter and average length of ORFs increases with increasing phylostrata (Figure S2). Consistent with the finding of Basile [59], the mean GC content for SGD-annotated orphan genes in *S. cerevisiae* is slightly lower (though not statistically significant) than that of more conserved genes. Like the SGD-annotated orphan genes, the Q3-transcribed orphan-ORFs (ORFs in top quartile of mean expression, see Figure 1.B) have a slightly lower mean GC content than ORFs of other phylostratum levels (Figure

S2.B). Vakirlis [55] reported a higher mean GC content among those orphan genes that have a confirmed *de novo* origin.

Transcriptional landscape of genes and ORFs

Expression of many annotated orphan genes is developmentally localized, up-regulated under environmental stress, or associated with species-specific traits [23, 60–63]. For example, more yeast orphans are ribosomally-bound under starvation conditions than control conditions [6, 7]. We anticipated that sparse-expression would be a characteristic of many of those orphan-ORFs that are actually orphan *genes* that have escaped annotation. To capture expression of these orphan-ORFs, we deemed it essential to use RNA-Seq samples comprising diverse developmental, genetic, and environmental conditions.

RNA-Seq samples drawn from a wide range of conditions have an added benefit. Because orphans have no homologs in other species, and no recognizable functional domains, these characteristics cannot be used to provide a clue as to function [23], rendering functional inference a particular challenge. The assumption that genes with similar patterns of expression are likely to encode proteins involved in a common process provides a powerful approach to infer experimentally-testable functions for genes of unknown function. Therefore, using datasets incorporating the diverse conditions in which orphans-ORFs or orphan genes might be expressed is key to functional inference and to determine the conditions that induce their expression.

To gather RNA-Seq data from diverse conditions, we collected raw sequence reads and metadata of 3,457 RNA-Seq samples from 177 studies in The National Center for Biotechnology Information-Sequence Read Archive (NCBI-SRA). (See *S.cerevisiae_RNA-seq_3457_27.mog* for metadata and counts). The experimental variables across these samples include a wide variety of mutants, chemical treatments, stresses, nutrition deprivations, and growth stages. We quantified the expression of all 29,354 ORFs and 6,692 SGD-annotated genes of *S. cerevisiae* across the 3,457 RNA-Seq samples.

Figure 2 shows a heatmap for expression of SGD-annotated genes, smORFs (sequences encoding small orphan proteins with ribosomal evidence of translation [7]), and transcribed orphan-ORFs (> 150 nt) across the 3,457 RNA-Seq samples. (See Figure S3 for expression plot of all genes and ORFs). The mean expression across all samples for SGD-annotated genes is 38 cpm, whereas the mean expression for the Q3-transcribed ORFs is 18 cpm (Table S1). Many SGD-annotated genes are expressed in most of the samples. In contrast, as we anticipated based on the erratic pattern of expression of annotated orphan genes, most of the orphan-ORFs show very low expression in most RNA-Seq samples, but accumulate highly in a few samples. This sporadic expression contributes significantly to the observed lower mean expression of the orphans. It also demonstrates how many transcribed sequences might be missed if smaller, less diverse datasets are analysed.

Ninety-nine percent of the 3,457 RNA-Seq samples have transcription evidence for at least one of the Q3-transcribed ORFs (Figure 3). Some samples are particularly rich in Q3-transcribed ORFs. For example, 50 samples have transcription evidence for >1,200 of the Q3-transcribed ORFs; 47 of these samples are from wild type strains, many grown under conditions of nutritional or chemical stress.

The conserved SGD-annotated genes have higher mean expression than either the orphan SGD-annotated genes, the Q3-transcribed orphan-ORFs, or the Q3-transcribed conserved-ORFs (Kolmogorov-Smirnov Test, p-values < 0.001; Figure 4). However, over 600 orphan-ORFs have a higher mean expression than 10% of conserved SGD-annotated genes, 289 orphan-ORFs have a mean expression higher than 25% of the conserved SGD-annotated genes, and 36 orphan-ORFs have a mean expression higher than 90% of conserved SGD-annotated genes (Figure 4 and Table S1. A).

Genomic context of the ORFs.

We surveyed the genomic location of each ORF relative to the nearest SGD-annotated gene (Figure S6). A recent study using on two experimental conditions [50] reported that a high proportion of expressed but unannotated transcripts in yeast overlap known CDSs but are transcribed from the opposite strand. Consistent with [50], 30% of the transcribed ORFs overlap CDSs and are transcribed from the opposite strand (reverse orientation) in our study. Furthermore, regardless of orientation (same *versus* reverse), ORFs that overlap an annotated CDS have a median of mean expression *5-fold higher* than ORFs located within or outside a CDS (Wilcoxon rank-sum test, p-value < 0.001, Figure S6). We have no explanation for this phenomenon. The orientation in which overlapping ORFs are transcribed relative to the associated CDSs does not significantly alter the mean expression level of the ORFs (Figure S6).

Of the 289 orphan-ORFs with the highest transcription (Figure 4), 49% overlap an annotated CDS, rather than being within an annotated CDS or outside an annotated CDS (Figure S7). This is significantly higher (Fisher's exact test, p-value< 0.001) than the proportion among all ORFs (15%) and all orphan-ORFs (14%) that overlap CDS (Figures S6, S7). Most of the 289 orphan-ORFs that overlap a CDS have a reverse orientation (convergent or divergent) relative to the SGD-annotated gene they overlap (Figure S7).

Many RNAs in fungi and humans that have been annotated as "lncRNAs" are associated with ribosomes, and/or have proteomics evidence, indicating some of them may function as protein-coding genes [2, 6, 11, 32, 64]. To examine translation evidence in our study, we globally evaluated translation evidence, mapping raw reads from 302 ribosomal profiling RNA-Seq (Ribo-Seq) samples in SRA to the unannotated ORFs and SGD-annotated genes of *S. cerevisiae.* (See supplementary file *Ribo-Seq_counts.csv* and *Ribo-Seq_metadata.xlsx* for raw counts and metadata). About 61% of Q3-transcribed conserved-ORFs, 40% of genus-specific-ORFs, and 51% of orphan-ORFs have translational evidence among these Ribo-Seq samples (Figure 1.B). This compares to 97% of the conserved SGD-annotated genes, 45% of genus-specific SGD-annotated genes, and 38% of orphan SGD-annotated genes. The mean Ribo-Seq raw counts were significantly different (t-test p-value < 0.001) among classes of transcripts, depending on whether they were orphan, genus-specific, or conserved (Figure 5.A). The mean Ribo-Seq raw counts for the low-transcribed ORFs are significantly lower than for the Q3-transcribed ORFs, and the mean Ribo-Seq raw counts for the ORFs with no transcription evidence are 0 or near 0 (Figure S8).

The proportions of Q3-transcribed ORFs with translation evidence located within, overlapping, or between annotated CDSs are significantly different among orphan-ORFs, genus-specific-ORFs, and conserved-ORFs (Chi-square test, p-value < 0.001) (Figure 5.B). Notably, 54% of Q3-transcribed orphan-ORFs with translation evidence are located in the intervals between annotated CDSs, compared to only 12% of the genus-specific ORFs and 9% of the conserved ORFs (Figure 5.B).

Since yeast was the first model eukaryotic genome [65], and has been reannotated over time, it would be expected that most conserved genes are already annotated. However, some genus-specific-genes might have been missed because homology is a major criterion used for genome annotation. Orphan genes, which have no homologs in other species, sparser expression, and likely fewer canonical features [41], are yet less likely to have been annotated. In total, 1,477 Q3-transcribed genus-specific-ORFs and 1,850 Q3-transcribed orphan-ORFs have ribosomal binding evidence. These transcribed, translated ORFs are candidates as protein-coding genes.

Five hundred and thirty of the 858 Q3-transcribed conserved-ORFs also have translation evidence. There are several possible explanations for why a transcript with homologs in other species are not annotated as genes. Some of these conserved-ORFs may be pseudogenes that retain some homology and expression, but have lost functional capacity. Other conserved-ORFs might encode active proteins, by because they are expressed only under limited conditions they might not have been sampled when SGD annotations were made. Other conserved-ORFs may have been ignored because their ORF codes for a shorter protein than the canonical gene family member. (On average, a Q3-transcribed conserved-ORF is significantly shorter than the homologous SGD-annotated gene (t-test, p-value < 0.001)). However, it not a given that because an ORF encodes a shorter protein it is non-functional. Shorter homologs of proteins with known function may play a biological role in regulating signal transduction, modulating enzyme activity, and/or affecting protein complexes, potentially competing with their “full-length” homolog [66, 67]. Translation of a short conserved-ORF also might be regulatory, in that it limits translation of a nearby active protein [68].

Network inference and co-expression analysis

To analyse the expression patterns of the ORFs in the context of annotated genes, we optimized correlation and network parameters for the RNA-Seq expression data (see Methods, Figures S10, S11, and Table S3), and focused our subsequent interactive co-expression analysis and visualization on a dataset ("SGD+ORF" dataset) composed of 14,885 transcripts (all SGD-annotated genes; the 7,054 Q3-transcribed ORFs; and all 1139 smORFs) across 3,457 RNA-Seq samples.

We then computed the Pearson pairwise correlation (PCC) matrix for the SGD+ORF dataset, and partitioned the resultant PCC matrix by Markov chain graph clustering (MCL) [69] into 544 clusters (Table S4 for overview; genes and ORFs with cluster designations at supplementary file *S.cerevisiae_RNA-seq.mog*). Forty-six percent of the 273 SGD-annotated orphan genes and 59% of the 3,899 Q3-transcribed orphan-ORFs are members of clusters containing more than five genes and include genes of known function, thus providing potential for functional inference.

It was possible that ORF expression might be correlated with that of adjacent or overlapping SGD-annotated genes, i.e., that ORFs are expressed due to a physical proximity to transcribed SGD-annotated genes. We used two approaches to evaluate the extent to which such "piggybacking" might occur. In the first approach, we focused on the 390 ORFs that are located completely within UTRs of SGD-annotated genes (88% are orphan-ORFs). About 80% of these ORFs have a PCC less than 0.6 (0.6 is the correlation cut-off we used for MCL) with the encompassing SGD-annotated genes, however, about 2% (eight) ORFs have a correlation higher than 0.9. In the second approach, we calculated how many ORFs are in the same cluster as nearby annotated genes. To do this, we randomly selected 366 ORFs that were members of clusters, and made test clusters of the same sizes, each cluster containing randomly-selected SGD-annotated genes and the identical ORFs as in the experimental data. Then, we calculated the distance of each ORF to each SGD-annotated gene in the randomly-created and the experimental clusters. The distances were not statistically different in the experimental versus the random clusters (p-value=0.16 in a t-test for difference). These tests indicate that the expression of ORFs is not generally associated with the ORFs being within or near to an SGD-annotated gene, and co-expressed with it. However, there is strong support for such a relationship in specific cases (e.g., Figure 9, and as reported in [55]).

About 65% of the Q3-transcribed ORFs are assigned to clusters in the co-expression matrix. Regardless of whether they are protein-coding, they could play a biological role. The highly transcribed ORFs with translational activity provide an evidence-based cadre of *candidate* protein-coding genes that could be experimentally tested.

GO enrichment analysis for co-expressed clusters

In order to evaluate the significance of the clustering results, we compared the extent of enrichment of Gene Ontology (GO) terms in the set of clusters obtained from MCL-partitioning experimental data to that of 100 randomly-generated sets of clusters. For each randomly-generated set, the number of clusters and the number of genes per cluster were held the same as the set of clusters from the experimental data; however, the genes assigned to each cluster were changed by random permutation. The best adjusted p-value for enriched GO terms was recorded for each cluster and averaged across all clusters to obtain a mean best p-value [70] (Figure 6). Distribution of the p-values for GO terms in the 100 sets of randomized clusters was compared to that of the experimental data (red arrows in Figure 6). For each GO ontology category (Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)), the best mean p-values for the experimental data are 0.019, 0.023, and 0.027, respectively. These values are significantly better than those of any of the randomly-obtained cluster sets, indicating that the MCL gene clusters derived from the experimental data is not random. Co-expressed genes are implicated as being involved in a similar process [71, 72]. That this study is based on over 3,000 biological conditions further strengthens the likelihood that genes in each cluster might share a related biological process.

Exploring Gene Function: Case study, Cluster 112

The co-expression clusters are often composed of genes and ORFs distributed across spatially diverse regions of the genome (For a list of all genes and ORFs as partitioned into clusters by MCL, see supplementary file *S.cerevisiae_RNA-seq_3457_27.mog).* For example, MCL Cluster 112 (Figure 7) contains 20 SGD-annotated genes and 21 unannotated ORFs dispersed on 14 chromosomes. Twelve of the genes are in the seripauperin (*PAU*) family. The molecular function of the *PAU* genes is not known. However, *PAU*-rich co-expressed gene clusters have been identified in independent microarray studies [73, 74]. Many *PAU*s are induced by low temperature and anaerobic conditions, and repressed by heme (Rachidi, Martinez, Barre & Blondin, 2000) and individual *PAU* proteins confer resistances to biotic and abiotic stresses [76]. *YER011W* and *YJR150C*, also in Cluster 112, are localized to the same cellular compartments as *PAU*s and are also induced under anaerobic conditions [77–80]. The other SGD-annotated genes in this cluster have no functional description. GO enrichment analysis identified eight GO terms as significantly-over-represented in Cluster 112 (Table 1). Figure 8 represents a case study of an approach to develop a meaningful hypothesis. The example shows the expression of the genes and ORFs in Cluster 112 across all 3545 samples of the RNA-Seq SGD+ORF dataset, and highlights the two studies that evaluate oxygen content as an experimental variable. Study SRP067275 compares four growth stages of the stress-tolerant yeast strain GLBRCY22-3 grown in YPDX and ACSH media, with and without oxygen [81] (Figure 8, top left); the expression of the genes and ORFs in Cluster 112 is higher under anaerobic conditions, irrespective of media or growth stage. Study SRP098655 compares *OLE1*-repressible strains growing under anaerobic and aerobic conditions [82] (Figure 8, top right); expression of genes and ORFs in Cluster 112 is induced in cells grown under anaerobic conditions. These expression patterns indicate the genes and the ORFs in this cluster might be sensitive to anoxia, or might play a role in cellular response to this stress.

Exploring Gene Function: Case study, *smORF247301*

Though rare, some transcribed ORFs that are located near or in an existing gene share a similar transcription pattern. An example is *smORF247301*, one of the most highly expressed smORFs, which is 77 nt upstream of *YPL223C* (Figure 9). MOG analysis indicates *smORF247301* and the SGD-annotated gene *YPL223C* have a PCC of 0.95 across the 3,457 RNA-Seq samples. *smORF247301* is located on the "+" strand of chromosome 16, while *YPL223C* is on the "-" strand of the same chromosome. The CDS of *YPL223C* is 507 nt, while *smORF247301* is 33 nt. *YPL223C* is more highly expressed than *smORF247301*. *YPL223C*, a hydrophilin gene that is essential in surviving desiccation-rehydration, is regulated by the high-osmolarity glycerol (HOG) pathway [83], and induced by osmotic, ionic, oxidative, heat shock and heavy metals stresses. Analysis using MOG shows *smORF247301* and *YPL223C* have increased expression in response to osmotic, heat, and desiccation stresses in three independent studies (Figure 9 B-D). *smORF247301* has translation evidence ([7] and this study).

It is possible that the transcription and translation of *smORF247301* is "noise" (Eling, Morgan & Marioni, 2019) associated with the expression of the nearby *YPL223C*. A second possibility is that *smORF247301* is a young, not-yet-annotated gene. It might be "piggybacking" on the expression apparatus of *YPL223C*. However, *smORF247301* and *YPL223C* are transcribed in a *convergent* orientation (-> <-, Figure 5B); thus, the process, described by [55], whereby two transcripts in divergent orientation (<- ->, Figure 5B) are co-expressed via a common bidirectional promoter would not apply in the case of *smORF247301* and *YPL223C*. A different "piggybacking" mechanism might apply: perhaps, due to its location in open chromatin, *smORF247301* is provided with a ready-made exposure to transcription factors when gene *YPL223C* is transcribed. If a transcript (e.g., *smORF247301*) conferred a survival advantage under the same conditions as did its established neighbouring gene (e.g., *YPL223C*), it could emerge as a new, co-expressed, gene by this mechanism.

Five hundred and thirty-seven orphan-ORFs with transcription and translation evidence are in physical proximity to an SGD-annotated gene and are transcribed in a *divergent* orientation (see supplementary file, *divergent_pairs.csv*). Of these pairs, 12 are co-expressed (PCC > 0.6); these 12 ORFs are potentially co-expressed by a bidirectional promoter (e.g., as described by [55]) The 525 orphan-ORFs that are not co-expressed, might still be controlled by a bidirectional promoter, because yeast ORFs can be transcribed by a bidirectional promoter, but not be correlated in expression because they are influenced by different transcription factors [85].

Future studies

The SGD+ORF dataset we provide can be reanalysed by different approaches. Each combination of network inference and partitioning approaches can supply complementary information. For example, networks can be inferred by correlation, mutual information [86], or relatedness approaches [87]. Pearson correlation, used here, is highly sensitive at extracting genes whose expression is linearly correlated across multiple conditions, but misses non-linear co-expression. Likewise, networks can be partitioned by several methods, e.g., MCL (as in this study), Modularity [88], or a promising new approach, Reduced Network Extreme Ensemble Learning (RenEEL) [89]. There has been little investigation into the strengths and weaknesses of the various inference and partitioning methods for extracting different types of biological information.

Moreover, we focus here on protein-coding transcripts; similar investigations using diverse RNA-Seq data could center on non-coding RNAs or transcript-encoding very small proteins.

The information resulting from such studies can easily be incorporated into a new MOG project to enable interactive analysis and visualization.

Conclusion

In this study we have globally assessed the accumulation of transcripts representing 36,046 annotated genes and unannotated ORFs of *S. cerevisiae* across 3,457 public RNA-Seq samples derived from diverse biological conditions. Ninety-five per cent of the transcribed ORFs are orphans or genus-specific. Despite a strong tendency to be transcribed only under restricted conditions, 269 orphan-ORFs had mean levels of transcription greater than 25% of SGD-annotated genes. Over 2,000 transcribed ORFs with translation evidence are members of co-expression clusters, providing additional clues as to a potential function.

The proportion of transcribed and translated ORFs that are functional is completely unknown. The SGD+ORF dataset assembled herein represents expression of SGD-annotated genes and unannotated ORFs under multiple conditions; it is delivered in a readily explorable, user-friendly format via the MOG platform. Combining this network-informed view of aggregate RNA-Seq data with text-mining of sample and gene metadata creates a powerful approach to develop novel, experimentally-testable hypotheses on the potential functions of as-yet-unannotated transcripts.

Materials and methods

Extracting ORFs and delineating orphan-ORFs in *S. cerevisiae*

ORFs (>150 nt) that were not annotated in SGD as CDS, were extracted from the yeast genome (version: R64-1-1) by bedtools2 [90], and translated by emboss [91], yielding 24,912 ORFs. To these ORFs we added two sets of ORFs <150 nt identified in other studies: the 1,139 small translated sequences (smORFs) identified by ribosome profiling [7] and the 3,303 of ORFs identified by TIF-Seq (txCDS) [3] that were less than 150 nt (thus, not included in the bedtools2 extraction). These 29,354 ORFs, together with the 6,692 protein-coding genes annotated in SGD, were subjected to phylostratigraphic analysis.

We inferred the phylostratum for 29,354 ORFs and 6,692 SGD-annotated protein-coding genes via the R package, *phylostratr* [40]. The analysis compared the proteins predicted from the *S. cerevisiae* ORFs to proteins of 123 target species distributed across phylostrata: 117 species identified by the *phylostratr* algorithm, supplemented with six manually-selected species in the Saccharomyces genus (*S. paradoxus, S. mikatae, S. kudriavzevii, S. arboricola, S. eubayanus*, and *S. uvarum*). To minimize false positives when identifying orphan ORFs and CDS from *S. cerevisiae*, we took advantage of the customization capabilities of *phylostratr* and included the predicted translation products from all ORFs (>150 nt) from each of the six *Saccharomyces* genomes, in addition to all SGD-annotated proteins of these species. (See *Supplementary_Material.pdf,* Figure S1 for workflow, Section 13 for full species list, and *phylostratr_heatmap.pdf* for gene by gene (and ORF by ORF) heatmap). Each gene was assigned to the most evolutionarily-distant phylostratum that contains an inferred homolog. A gene or ORF is inferred to be an orphan if its encoded protein is assigned the phylostratum level *S. cerevisiae*.

A BLASTP for each ORF and CDS in *S. cerevisiae* against *Saccharomyces* spp ORFs and CDS gave identical results to those of *phylostratr* in identifying the orphan genes and ORFs.

Raw read processing and network optimization

Our RNA-Seq data analysis pipeline is shown in Figure S9. We selected all samples with *S. cerevisiae* taxon ID 4932, Illumina platform, and paired layout from NCBI-SRA and then filtered out samples with miRNA-Seq, ncRNA-Seq, or RIP-Seq library strategies. In total, we collected raw reads data (FASTQ format) and metadata from 3,457 RNA-Seq samples (177 studies). A transcriptome was created from SGD-annotated cDNA and unannotated ORFs, and then expression levels of annotated genes and ORFs over the 3,457 RNA-Seq samples were quantified by *kallisto* [92] (See supplementary file *S.cerevisiae_RNA-seq.mog* for RNA-Seq metadata and normalized cpm data; all data including raw counts is accessible at DataHub (https://datahub.io/lijing28101/yeast_supplementary)).

We evaluated the performance of two diverse normalization methods for the raw count data (Section 8 in *Supplementary_Material.pdf*). We normalized raw counts by *edgeR* [93] based on the evaluation of [94]. We also normalized the same data by a single cell RNA-Seq normalization approach *SCnorm* [95]. This method examines sequence information from individual cells with the aim to provide a higher resolution of cellular differences. We tested this method because two features of single cell RNA-Seq data are similar to the orphan-ORF-focused multi-study data assembled for our study: 1) raw counts contain an abundance of zero-expression values, 2) technical variability among samples is high. After normalization, only ORFs with mean expression values in the upper quantile of mean expression (Q3-transcribed) were retained. We generated two datasets: 1) all SGD-annotated genes (SGD dataset); and 2) all Q3-transcribed ORFs, smORFs, and SGD-annotated genes (SGD+ORF dataset). For each normalization approach and dataset, we calculated pairwise Pearson correlation matrices among all 3,457 RNA-Seq samples.

Three PCC cutoffs (0.6, 0.7 and 0.8) were used to create networks of different densities from the matrices (Section 7 in *Supplementary_Material.pdf*). We then applied MCL to partition each network using our in-house Java Spark implementation (GitHub: https://github.com/lijing28101/SPARK_MCL) designed to optimize efficiency. All data analysis in this work, except for MCL clustering and RNA-Seq expression visualization, were performed in R software.

Cluster evaluation by GO term enrichment analysis

Clusters resulting from each of the eight MCL analyses obtained from the different normalization methods and PCCs were evaluated by GO enrichment analysis using *clusterProfiler* [96]; in this evaluation, only clusters with over five genes were considered (Section 7 in *Supplementary_Material.pdf*). The GO term enrichment of each experimental result was compared to that of 100 random sets of clusters, which were obtained by permuting gene IDs. For these permutations, the same number of clusters of the same size as those from the experimental result were assigned to each random set using the method of [70]. The best adjusted p-value (p_{min}, smallest adjusted p-value) was recorded for the enriched GO terms in each cluster. Each random cluster set was assigned a score Si, which is the average p_{min} across all clusters in the set.

$$S_i = \frac{\sum_{j=1}^{n} p_{\min j}}{n} \qquad (1)$$

where n indicates the number of clusters. The distribution of S values for GO classes, Biological Process, Cellular Component, and Molecular Function, for random sets were compared to the respective values for the real experimental data. In each ontology, the experimental score was less than any of the random scores, indicating that experimental data have biological significance (permutation test, p-value=0). Based on the GO enrichment results we chose *edgeR* normalization (Section 8 in *Supplementary_Material.pdf*) and a PCC of 0.6 (Section 7 in *Supplementary_Material.pdf*) for future analyses.

Ribo-Seq analysis

To investigate the translational activity of unannotated ORFs, we analysed 302 samples (23 studies) of yeast Ribo-Seq data; this represented about half of the available Ribo-Seq in the SRA database. Raw reads (SRA-formatted) were downloaded, and the SRA toolkit was used to convert the raw reads to a FASTQ format. *BBDuk* was used to find and remove adapter sequences from the 3’ end of reads, and rRNA reads were identified and removed using *BBMap* [97]. The cleaned Ribo-Seq reads were aligned to the reference genome by *HISAT2* [98]. The actively translating ORFs were detected and quantified by *Ribotricer,* which considers the periodicity of ORF profiles and provides multiple options for customization (we used the recommended parameters for yeast) [15]. The gene/ORF with mean counts across 302 Ribo-Seq samples higher than 0.3 was consider to have translation evidence.

Visualization and gene function exploration

As proof-of-concept for the utility of these data, we used the MOG platform [51] to provide examples of co-expression and functional inference. We first created a MOG project that combined: 1) the levels of expression of each gene and ORF in the SGD+ORF dataset across 3,457 conditions, 2) gene and ORF metadata, and 3) sample metadata. The gene and ORF metadata includes: functional annotations (from SGD); MCL cluster memberships with GO enrichment analysis; mean expression levels for RNA-Seq and ribosomal profiling; ribosomal binding evidence; genome location relative to UTRs and CDSs; GC content; length; genomic positional coordinates, orientation; and phylostratal assignment. We then added metadata to the MOG project about each sample and study from NCBI-SRA, including: study ID, title, summary, reference, design description, library construction protocol, sequencing apparatus; sample title, experimental attributes, number of replicates; replicate name, sequencing depth, base coverage.

Authors' contributions

JL and EW conceived of the project and drafted the manuscript. All authors contributed to the manuscript. JL carried out the design of the study and performed the statistical analysis. US participated in the visualization on MOG and provided Ribo-Seq analysis code and guidance. ZA contributed to the phylostrata analysis. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

We thank Kevin Bassler and Pramesh Singh at University of Houston for insightful discussion about clustering analysis, and Manhoi Hur and Nishanth Sivakumar for developing the MCL Spark pipeline used in this study. We are grateful to Arun Seetharam and Priyanka Bhandary for helpful discussion on RNA-Seq alignment. This work was enabled by an award to ESW, AS and US from Extreme Science and Engineering Discovery Environment (XSEDE) supported by National Science Foundation ACI-1548562, and by Condo Cluster at Iowa State University.

Funding

This work is funded in part by the National Science Foundation grant NSF-IOS 1546858, *Orphan Genes: An Untapped Genetic Reservoir of Novel Traits.*

References

[1] Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 2007; 14: 103.

[2] Hangauer MJ, Vaughn IW, McManus MT. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet*; 9. Epub ahead of print 20 June 2013. DOI: 10.1371/journal.pgen.1003569.

[3] Lu T-C, Leu J-Y, Lin W-C. A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Mol Biol Evol* 2017; 34: 2823–2838.

[4] Pertea M, Shumate A, Pertea G, et al. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 2018; 332825.

[5] Wu B, Knudson A. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio* 2018; 9: e01024-18.

[6] Wilson BA, Masel J. Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome Biol Evol* 2011; 3: 1245–1252.

[7] Carvunis A-R, Rolland T, Wapinski I, et al. Proto-genes and *de novo* gene birth. *Nature* 2012; 487: 370–374.

[8] Chew G-L, Pauli A, Rinn JL, et al. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Dev Camb Engl* 2013; 140: 2828–2834.

[9] Smith JE, Alvarez-Dominguez JR, Kline N, et al. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Rep* 2014; 7: 1858–1866.

[10] Ruiz-Orera J, Messeguer X, Subirana JA, et al. Long non-coding RNAs as a source of new peptides. *eLife*; 3. Epub ahead of print 2014. DOI: 10.7554/eLife.03523.

[11] Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet* 2015; 11: e1005721.

[12] Hsu PY, Calviello L, Wu H-YL, et al. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc Natl Acad Sci* 2016; 113: E7126–E7135.

[13] Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics*; 17. Epub ahead of print 31 May 2016. DOI: 10.1186/s12859-016-1102-x.

[14] Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *nar* 2017; 46: D497–D502.

[15] Choudhary S, Li W, Smith AD. Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinforma Oxf Engl*. Epub ahead of print 21 November 2019. DOI: 10.1093/bioinformatics/btz878.

[16] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489: 57–74.

[17] Lloréns-Rico V, Cano J, Kamminga T, et al. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* 2016; 2: e1501363.

[18] Barroso GV, Puzovic N, Dutheil JY. The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics* 2018; 208: 173–189.

[19] Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 2014; 15: 193.

[20] Ji Z, Song R, Regev A, et al. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 2015; 4: e08890.

[21] Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* 2007; 23: 533–539.

[22] Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat Rev Genet* 2011; 12: 692–702.

[23] Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. *Trends Plant Sci* 2014; 19: 698–708.

[24] McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet* 2016; 17: 567–578.

[25] Schlötterer C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet* 2015; 31: 215–219.

[26] Arendsee Z, Li J, Singh U, et al. fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics* 2019; 20: 440.

[27] Oss SBV, Carvunis A-R. De novo gene birth. *PLOS Genet* 2019; 15: e1008160.

[28] Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. *Nat Rev Genet* 2013; 14: 645–660.

[29] Landry CR, Zhong X, Nielly-Thibault L, et al. Found in translation: functions and evolution of a recently discovered alternative proteome. *Curr Opin Struct Biol* 2015; 32: 74–80.

[30] Hoen DR, Bureau TE. Discovery of Novel Genes Derived from Transposable Elements Using Integrative Genomic Analysis. *Mol Biol Evol* 2015; 32: 1487–1506.

[31] Gubala AM, Schmitz JF, Kearns MJ, et al. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. *Mol Biol Evol* 2017; 34: 1066–1082.

[32] Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, et al. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* 2018; 2: 890–896.

[33] Xie C, Bekpen C, Künzel S, et al. Studying the dawn of de novo gene emergence in mice reveals fast integration of new genes into functional networks. *bioRxiv* 2019; 510214.

[34] Bao Z, Clancy MA, Carvalho RF, et al. Identification of Novel Growth Regulators in Plant Populations Expressing Random Peptides. *Plant Physiol* 2017; 175: 619–627.

[35] Neme R, Amador C, Yildirim B, et al. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat Ecol Evol*; 1. Epub ahead of print June 2017. DOI: 10.1038/s41559-017-0127.

[36] Menschaert G, Van Criekinge W, Notelaers T, et al. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics MCP* 2013; 12: 1780–1790.

[37] Vanderperre B, Lucier J-F, Bissonnette C, et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS One* 2013; 8: e70698.

[38] Khalturin K, Hemmrich G, Fraune S, et al. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 2009; 25: 404–413.

[39] Vakirlis N, Carvunis A-R, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the dominant source of orphan genes. *bioRxiv* 2019; 735175.

[40] Arendsee Z, Li J, Singh U, et al. phylostratr: a framework for phylostratigraphy. *Bioinformatics* 2019; 35: 3617–3627.

[41] Seetharam A, Arendsee Z, Wurtele E. Maximizing prediction of orphan genes in assembled genomes. *bioRxiv*.

[42] Meyer IM, Durbin R. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res* 2004; 32: 776–783.

[43] Proux-Wéra E, Armisén D, Byrne KP, et al. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* 2012; 13: 237.

[44] Cantarel BL, Korf I, Robb SMC, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 2008; 18: 188–196.

[45] Hoff KJ, Lange S, Lomsadze A, et al. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 2016; 32: 767–769.

[46] Berardini TZ, Reiser L, Li D, et al. The Arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome. *Genes N Y N 2000* 2015; 53: 474–485.

[47] Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. *G3 Genes Genomes Genet* 2015; 5: 1323–1333.

[48] González C, Lazcano M, Valdés J, et al. Bioinformatic Analyses of Unique (Orphan) Core Genes of the Genus Acidithiobacillus: Functional Inferences and Use As Molecular Probes for Genomic and Metagenomic/Transcriptomic Interrogation. *Front Microbiol* 2016; 7: 2035–2035.

[49] Werner MS, Sieriebriennikov B, Prabh N, et al. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res* 2018; 28: 1675–1687.

[50] Blevins WR, Ruiz-Orera J, Messeguer X, et al. Frequent birth of de novo genes in the compact yeast genome. *bioRxiv* 2019; 575837.

[51] Singh U, Hur M, Dorman K, et al. MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets. *bioRxiv*. Epub ahead of print 2019. DOI: 10.1101/698969.

[52] Tukey JW. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[53] Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 2013; 497: 127–131.

[54] Wu B, Knudson A. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio* 2018; 9: e01024-18.

[55] Vakirlis N, Hebert AS, Opulente DA, et al. A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol* 2018; 35: 631–645.

[56] Šestak MS, Domazet-Lošo T. Phylostratigraphic Profiles in Zebrafish Uncover Chordate Origins of the Vertebrate Brain. *Mol Biol Evol* 2015; 32: 299–312.

[57] Toll-Riera M, Bosch N, Bellora N, et al. Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Mol Biol Evol* 2009; 26: 603–612.

[58] Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. *eLife* 2014; 3: e01311.

[59] Basile W, Elofsson A. The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes. *bioRxiv*. Epub ahead of print 7 September 2017. DOI: 10.1101/185983.

[60] Guo W-J, Li P, Ling J, et al. *Significant Comparative Characteristics between Orphan and Nonorphan Genes in the Rice (Oryza sativa L.) Genome*. Epub ahead of print 2007. DOI: 10.1155/2007/21676.

[61] Li L, Foster CM, Gan Q, et al. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J* 2009; 58: 485–498.

[62] Colbourne JK, Pfrender ME, Gilbert D, et al. The Ecoresponsive Genome of Daphnia pulex. *Science* 2011; 331: 555–561.

[63] Bhandary P, Seetharam AS, Arendsee ZW, et al. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. *Plant Sci* 2018; 267: 32–47.

[64] Wu D-D, Irwin DM, Zhang Y-P. De Novo Origin of Human Protein-Coding Genes. *PLOS Genet* 2011; 7: e1002379.

[65] Hawthorne DC, Mortimer RK. Chromosome Mapping in Saccharomyces: Centromere-Linked Genes. *Genetics* 1960; 45: 1085–1110.

[66] Frith MC, Forrest AR, Nourbakhsh E, et al. The abundance of short proteins in the mammalian proteome. *PLoS Genet* 2006; 2: e52.

[67] Storz G, Wolf YI, Ramamurthi KS. Small Proteins Can No Longer Be Ignored. *Annu Rev Biochem* 2014; 83: 753–777.

[68] Wu H-YL, Song G, Walley JW, et al. The Tomato Translational Landscape Revealed by Transcriptome Assembly and Ribosome Profiling. *Plant Physiol* 2019; 181: 367–380.

[69] Dongen V, Marinus S. Graph clustering by flow simulation. *Dr Diss* 2000; 1.

[70] Mentzen WI, Wurtele ES. Regulon organization of Arabidopsis. *BMC Plant Biol* 2008; 8: 99.

[71] Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* 1998; 95: 14863–14868.

[72] Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol Biol Cell* 1998; 9: 3273–3297.

[73] Orellana M, Aceituno FF, Slater AW, et al. Metabolic and transcriptomic response of the wine yeast Saccharomyces cerevisiae strain EC1118 after an oxygen impulse under carbon-sufficient, nitrogen-limited fermentative conditions. *FEMS Yeast Res* 2014; 14: 412–424.

[74] Magwene PM, Kim J. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol* 2004; 5: R100.

[75] Rachidi N, Martinez M-J, Barre P, et al. Saccharomyces cerevisiae PAU genes are induced by anaerobiosis. *Mol Microbiol* 2000; 35: 1421–1430.

[76] Rivero D, Berná L, Stefanini I, et al. Hsp12p and PAU genes are involved in ecological interactions between natural yeast strains. *Environ Microbiol* 2015; 17: 3069–3081.

[77] Kowalski LR, Kondo K, Inouye M. Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae. *Mol Microbiol* 1995; 15: 341–353.

[78] Kitagaki H, Shimoi H, Itoh K. Identification and Analysis of a Static Culture-Specific Cell Wall Protein, Tir1p/Srp1p in Saccharomyces Cerevisiae. *Eur J Biochem* 1997; 249: 343–349.

[79] Sertil O, Cohen BD, Davies KJ, et al. The DAN1 gene of S. cerevisiae is regulated in parallel with the hypoxic genes, but by a different mechanism. *Gene* 1997; 192: 199–205.

[80] Cohen BD, Sertil O, Abramova NE, et al. Induction and repression of DAN1 and the family of anaerobic mannoprotein genes in Saccharomyces cerevisiae occurs through a complex array of regulatory sites. *Nucleic Acids Res* 2001; 29: 799–808.

[81] McIlwain SJ, Peris D, Sardi M, et al. Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research. *G3 GenesGenomesGenetics* 2016; 6: 1757–1766.

[82] Degreif D, de Rond T, Bertl A, et al. Lipid engineering reveals regulatory roles for membrane fluidity in yeast flocculation and oxygen-limited growth. *Metab Eng* 2017; 41: 46–56.

[83] Garay-Arroyo A, Colmenero-Flores JM, Garciarrubio A, et al. Highly Hydrophilic Proteins in Prokaryotes and Eukaryotes Are Common during Conditions of Water Deficit. *J Biol Chem* 2000; 275: 5668–5674.

[84] Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. *Nat Rev Genet* 2019; 1.

[85] Xu Z, Wei W, Gagneur J, et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* 2009; 457: 1033–1037.

[86] Zhang X, Zhao X-M, He K, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 2012; 28: 98–104.

[87] Netotea S, Sundell D, Street NR, et al. ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics* 2014; 15: 106.

[88] Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 2006; 103: 8577–8582.

[89] Guo J, Singh P, Bassler KE. Reduced network extremal ensemble learning (RenEEL) scheme for community detection in complex networks. *Sci Rep* 2019; 9: 1–11.

[90] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; 26: 841–842.

[91] Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 2000; 16: 276–277.

[92] Weijers SR, de Jonge J, van Zanten O, et al. KALLISTO: cost effective and integrated optimization of the urban wastewater system Eindhoven. *Water Pract Technol*; 7. Epub ahead of print 1 June 2012. DOI: 10.2166/wpt.2012.036.

[93] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139–140.

[94] Dillies M-A, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013; 14: 671–683.

[95] Bacher R, Chu L-F, Leng N, et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* 2017; 14: 584–586.

[96] Yu G, Wang L-G, Han Y, et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J Integr Biol* 2012; 16: 284–287.

[97] Bushnell B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. LBNL-7065E, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner (17 March 2014, accessed 16 December 2019).

[98] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015; 12: 357–360.

[99] Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal, p. 1695.

Figure legends

Figure 1 Quantification of SGD annotated genes and dark transcriptome.

A. Definition of Dark transcriptome. Pervasive transcription of unannotated sequences has been found in many species. Some of these might be protein coding genes that have escaped annotation. Most of these unannotated coding genes are orphan (species-specific) genes, which have no homolog to other species, and are hard to predict using current gene prediction tools. These orphan genes could emerge by rapid divergence from ancient genes or could evolve *de novo*. Other transcribed but unannotated sequences might be non-coding genes. Although many studies have explored the function and classification of the non-coding transcripts, many transcribed sequences are still unclassified.

B. Classification and numbers of expressed transcripts for SGD-annotated genes (green boxes) and ORFS (yellow boxes). Orphan-ORFs, unique to *Saccharomyces cerevisiae* (phylostrata (PS)=15); genus-specific-ORFs, unique to *Saccharomyces* spp. (PS=10-14); conserved-ORFs, homologs in older species (PS=1-9). Q3-transcribed, ORFs with mean transcription across the 3,457 samples ranking in the upper (Q3) quantile of the unannotated transcripts. Low-transcribed ORFs, ORFs with mean transcription across the 3,457 samples ranking in the lower 75% of the unannotated transcripts. Non-transcribed orphan-ORFs (Figure S8). Red font, number of genes/ORFs with translation evidence according to Ribo-Seq analysis. (For full PS designations and transcription expression, see supplementary file, *S. cerevisiae_RNA-seq_3457_27.mog*; for translation per transcript, see supplementary file, *Ribo-Seq_rawcounts.csv.*)

Figure 2 RNA-Seq expression heatmap across 3,457 samples for orphan-ORFs and SGD- annotated genes.

Top panel, SGD-annotated genes (6,692); middle panel, smORFs (Carvones et al., 2012) (1,139); bottom panel, orphan-ORFs (15,805). (See Figure S3 for all transcript classes). Each row represents a transcript. Within a panel, each transcript is ordered by its mean cpm. Within each row, the 3,457 samples are sorted independently by highest expression of the transcript. The restricted conditions of expression of many orphan-ORFs is visually apparent.

Figure 3 Counts of highly-transcribed (Q3) ORFs in each RNA-Seq sample

The black bars show distribution of the counts of the 8193 Q3-transcribed ORFs. X-axis, 3,457 RNA-Seq samples, sorted by counts. The grey bar inset details the 50 RNA-Seq samples with the largest number of Q3-transcribed ORFs; each of these samples contains over 1200 Q3-transcribed ORFs.

Figure 4 Density plot of mean expression level of transcripts across 3,457 samples for SGD-annotated genes and Q3-transcribed ORFs

X-axis, *edgeR*-normalized mean expression of genes and ORFs. Y-axis, number of transcripts. The area under the curve of the density function represents the probability of a range of mean cpm. The bimodal curve of all orphan-ORFs is attributable to the low mean expression of the smORFs (see Figure S4, S5). About half of the Q3-transcribed orphan-ORFs have higher mean expression than orphan SGD-annotated genes. Over 600 orphan-ORFs have a higher mean expression than 10% of conserved SGD-annotated genes; 289 orphan-ORFs (gray hatched area) have a higher mean expression than 25% of conserved SGD-annotated genes; and, 36 orphan-ORFs have a mean expression higher than 90% of conserved SGD-annotated genes (See also Table S2)

Figure 5 Mean expression and numbers of genes and ORFs with translational evidence, partitioned by phylostrata and genomic context.

Ribo-Seq data were analysed for genes and ORFs across 302 samples using *ribotricer* [15]. **A.** Mean raw Ribo-Seq counts/transcript for all genes and ORFs. X-axis, genes and ORFs as classified by phylostrata. Y-axis, mean raw counts. The letters above each bar indicate significance in each group according to a t-test (p-value cutoff is 0.01). Similar to mean RNA-Seq counts, the conserved genes and conserved-ORFs have more total mean Ribo-Seq counts. **B.** The 3,857 Q3-transcribed ORFs that had Ribo-Seq translation evidence were divided into groups according to their relationship to annotated CDS (see also Figure S6), and the numbers of genes and ORFs with translational evidence was determined. The gene/ORF with mean counts across 302 Ribo-Seq samples higher than 0.3 was consider to have translation evidence. X-axis, groups of genes and ORFs, classified by phylostrata. Y-axis, number of ORFs in each group. The proportions of ORFs are significantly different among three groups according to a chi-square test (p-value<0.001). Over half the orphan-ORFs with translation evidence are located between CDSs.

Figure 6 GO enrichment analysis of experimental data and random test distribution

A Pearson correlation matrix of the SGD+ORF dataset was partitioned into clusters by MCL. Best p-values (mean of the lowest adjusted p-values for GO terms) were determined across all clusters of the experimental data and all clusters of random permutations, similar to [70] (Section 7 in *Supplementary_Material.pdf*). Red arrow, experimental data. Black bars, best p-value of 100 randomly-obtained permutations with size and number of clusters identical to experimental data. BP, biological process; CC, cellular component; MF, molecular function. The clustering result is significantly better for experimental data than any random permutation.

Figure 7 Network view of genes and ORFs in Cluster 112

A Pearson correlation matrix of the SGD+ORF dataset was partitioned into clusters by MCL. Cluster 112 is an example of a cluster containing SGD-annotated genes and ORFs, including orphans. Edge colors, Pearson correlations of 0.6 to 1.0. Visualization by *igraph* in R [99].

Figure 8 The 41 genes and ORFs in Cluster 112 respond to anoxia

A Pearson correlation matrix of the SGD+ORF dataset was partitioned into clusters by MCL. The 41 genes and ORFs in Cluster 112 are co-expressed across multiple conditions. X-axis, 3,457-samples, sorted by study. Y-axis, expression values. Each line represents the expression pattern of a single gene or ORF. Top left inset, zoom-in to visualize Study SRP067275. RNA-Seq samples sorted by: aerobic or anaerobic condition, ACSH or YPDX media, and growth phase. ACSH, Ammonia Fiber Expansion-(AFEX-) pretreated corn stover hydrolysate. YPDX, YP media containing 60 g/L and 30 g/L xylose. Top right inset, zoom-in to visualize Study SRP098655. The genes and ORFs are up-regulated in response to anoxia, regardless of changes in growth media. No ORF in Cluster 112 is located near an SGD-annotated gene in Cluster 112. Visualizations and co-expression calculations by MOG (Singh et al., 2020).

Figure 9 Expression patterns of smORF247301 and YPL223C

smORF247301 and *YPL223C* are located on adjacent regions of chromosome 16 and are transcribed in convergent orientation. **A.** Expression patterns are similar (Pearson correlation, 0.95) across 3,457 samples. **B-D.** Expression patterns for smORF247301, YPL223C in three studies. X-axis, 3 samples per treatment. Purple bar on right side of panels, mean expression level of all SGD-annotated genes; Green bar on right side of panels, mean expression level of all SGD-annotated genes plus ORFs. Visualizations and co-expression calculations produced by MOG (Singh et al., 2020).

Tables

Table 1 Significantly-enriched GO terms in Cluster 112

Based on the GO terms assigned to the gene members of known function, Cluster 112 is enriched in the GO terms shown in Table 1. The results indicate a possible role in stress response related to the cell wall for the ORF members of Cluster 112. (See *S. cerevisiae_RNA-seq_3457_27.mog* for complete clustering and ontology results).

Ontology	GO name	Adjust p-value
MF	structural constituent of cell wall	1.90E-27
BP	response to stress	3.51E-27
CC	fungal-type cell wall	1.62E-20
BP	fungal-type cell wall organization	1.34E-19
CC	fungal-type vacuole	3.60E-05
CC	cell wall	1.99E-03
CC	extracellular region	6.18E-03
CC	anchored component of membrane	1.87E-02

Supplementary material

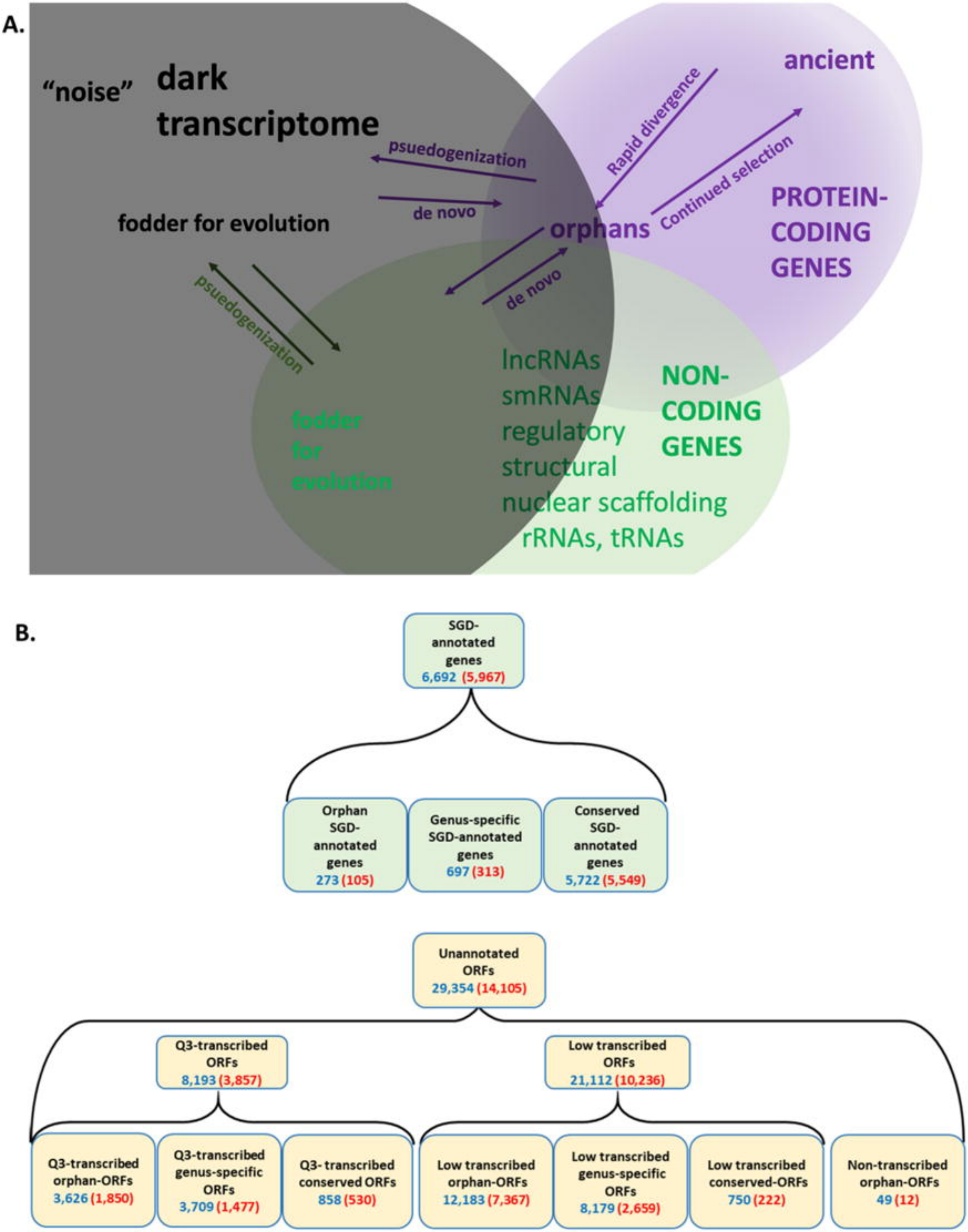
Supplementary Materials.pdf: include all supplementary figures, tables and description.

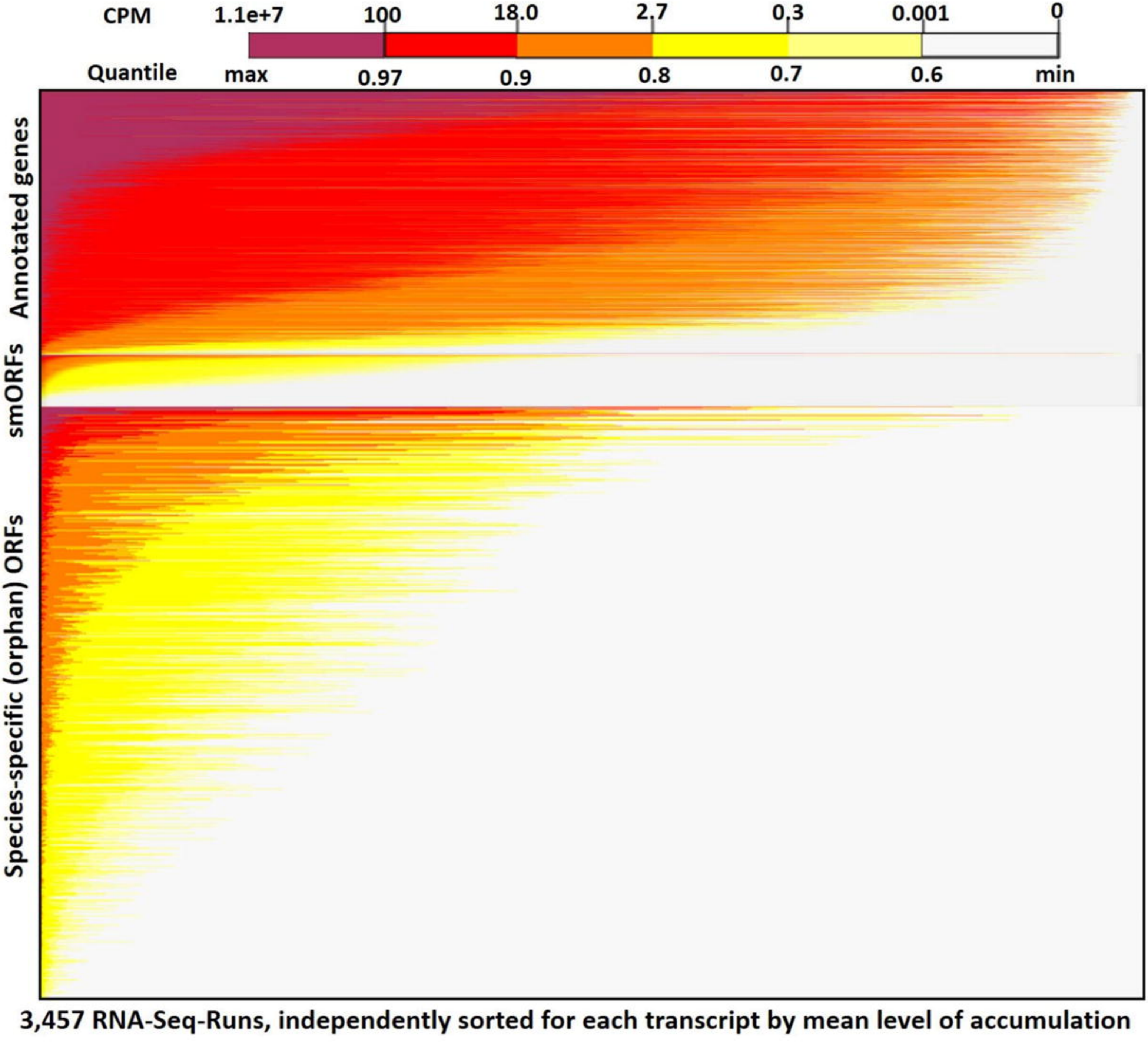
All supplementary data (including MOG files, raw count data, cluster information, UTR results, phylostratr heatmap, Ribo-Seq metadata and results) are available at https://datahub.io/lijing28101/yeast_supplementary

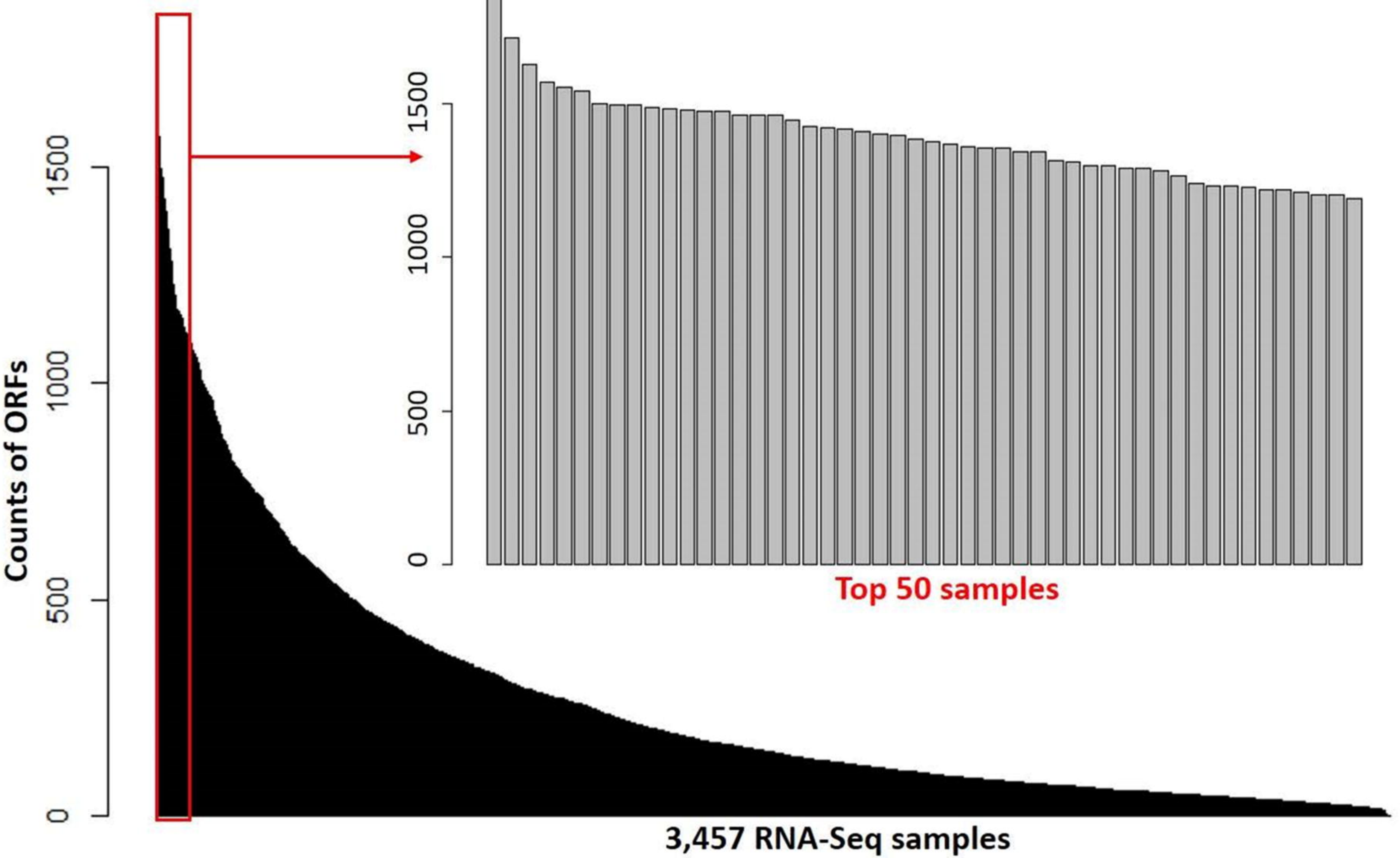
MOG file of *S. cerevisiae* RNA-Seq expression (*S.cerevisiae_RNA-seq_3457_27.mog*): http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm

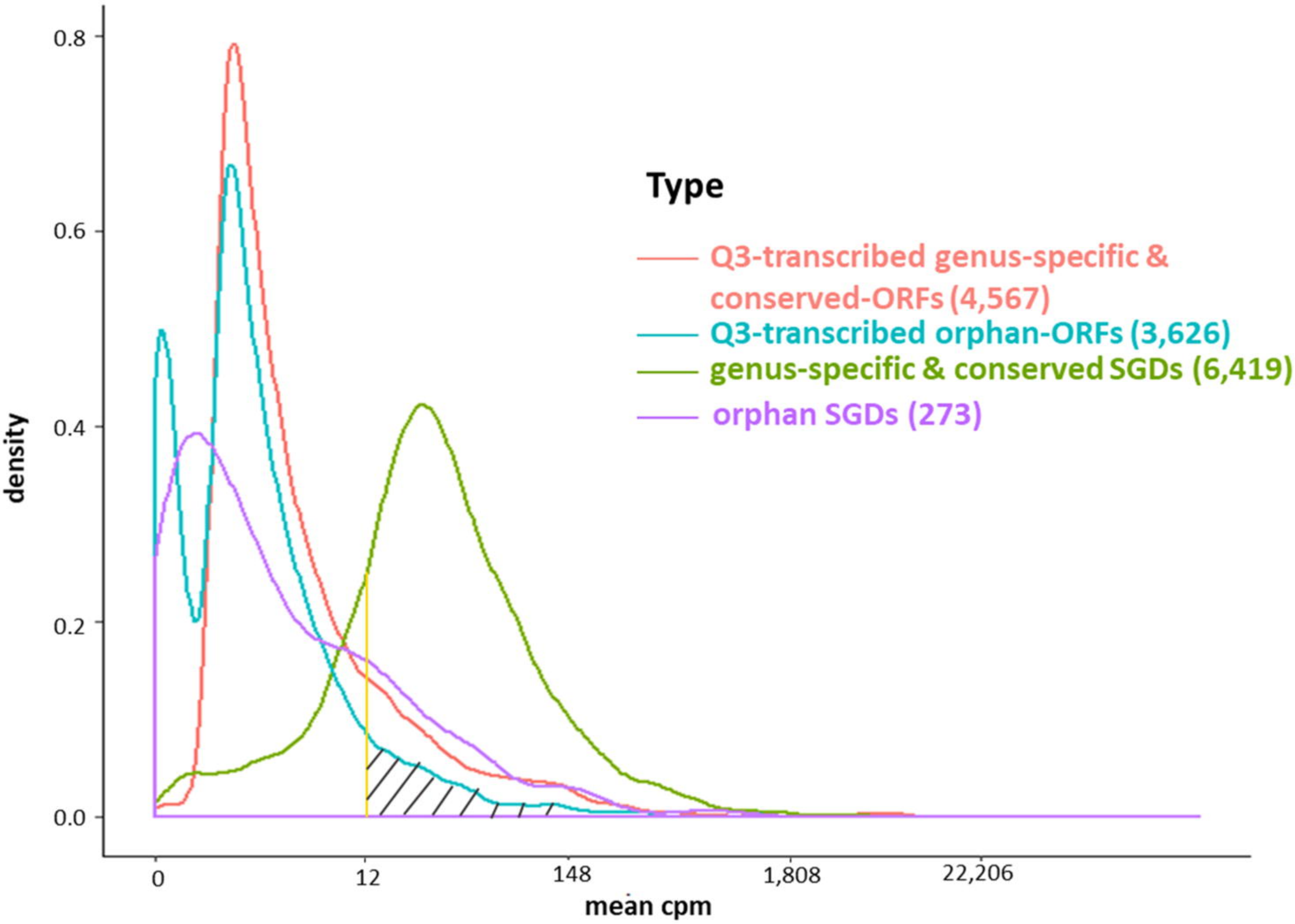
MetaOmGraph software: https://github.com/urmi-21/MetaOmGraph

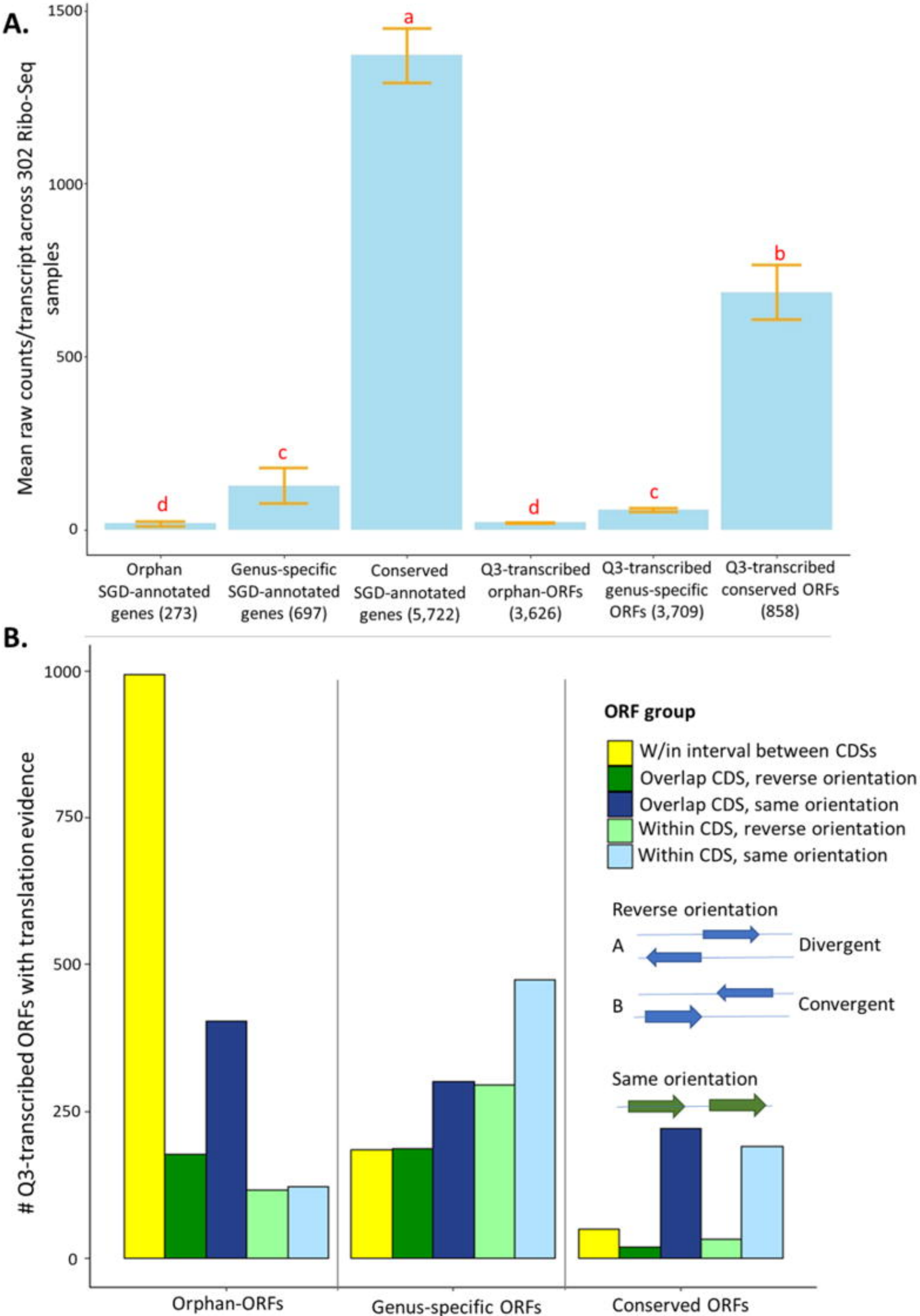
Data processing code: https://github.com/lijing28101/yeast_supplementary

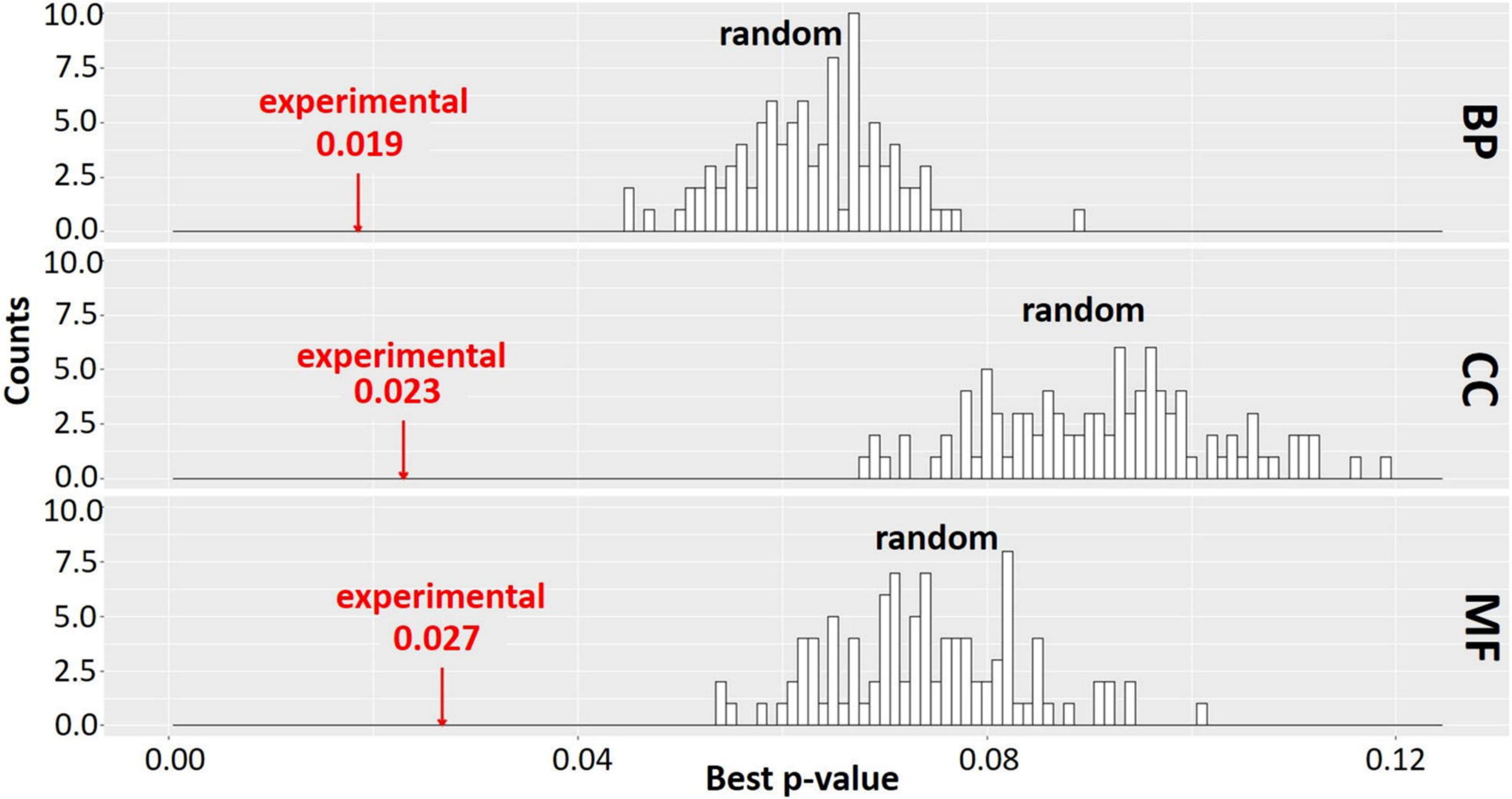

A.
"noise"
dark transcriptome
pseudogenization
de novo
fodder for evolution
pseudogenization
orphans
Rapid divergence
Continued selection
ancient
PROTEIN-CODING GENES
de novo
lncRNAs
smRNAs
regulatory
structural
nuclear scaffolding
rRNAs, tRNAs
fodder for evolution
NON-CODING GENES
B.
SGD-annotated genes
6,692 (5,967)
Orphan SGD-annotated genes
273 (105)
Genus-specific SGD-annotated genes
697 (313)
Conserved SGD-annotated genes
5,722 (5,549)
Unannotated ORFs
29,354 (14,105)
Q3-transcribed ORFs
8,193 (3,857)
Low transcribed ORFs
21,112 (10,236)
Q3-transcribed orphan-ORFs
3,626 (1,850)
Q3-transcribed genus-specific ORFs
3,709 (1,477)
Q3- transcribed conserved ORFs
858 (530)
Low transcribed orphan-ORFs
12,183 (7,367)
Low transcribed genus-specific ORFs
8,179 (2,659)
Low transcribed conserved-ORFs
750 (222)
Non-transcribed orphan-ORFs
49 (12)

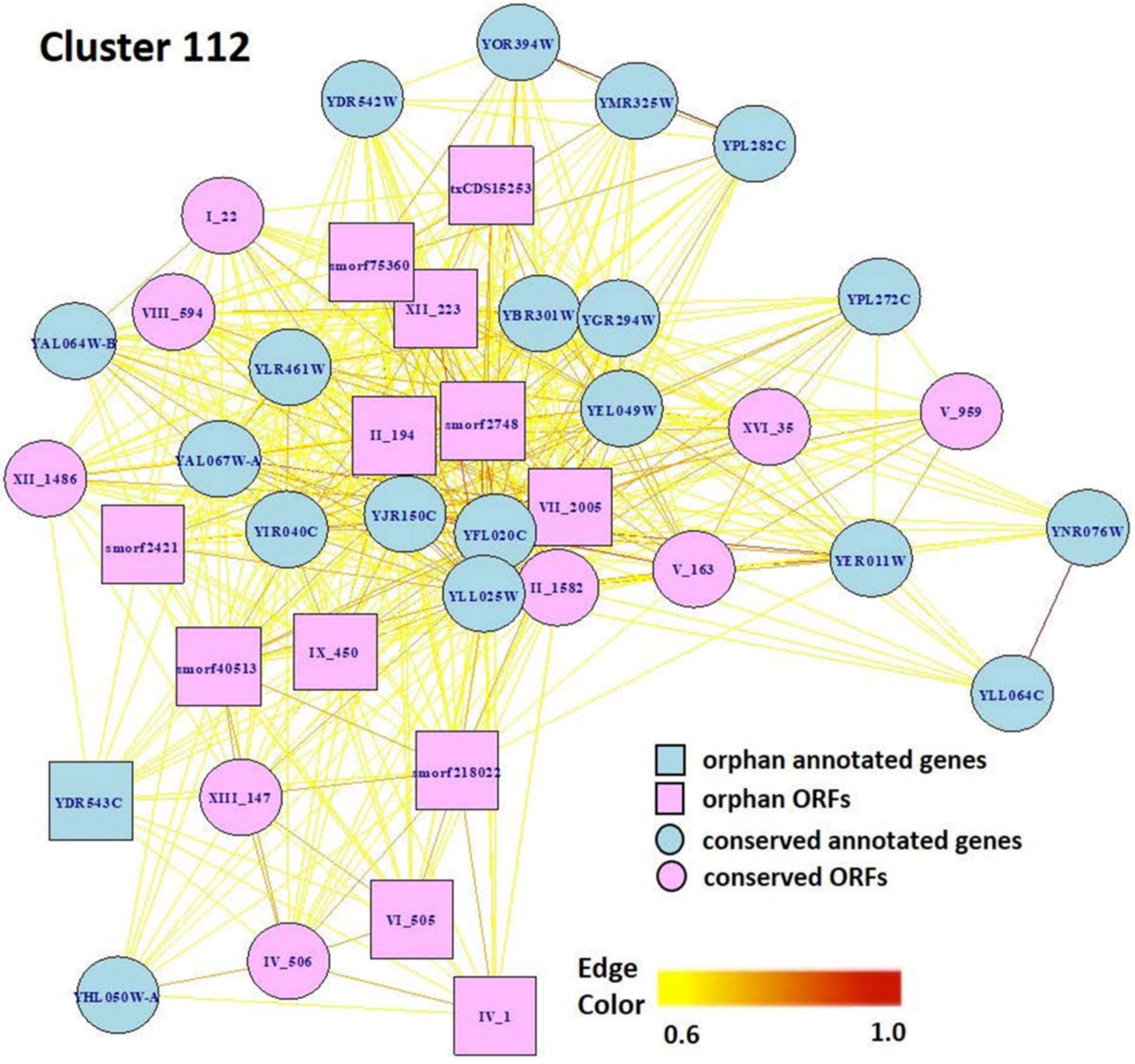

CPM
1.1e+7
100
18.0
2.7
0.3
0.001
0
Quantile
max
0.97
0.9
0.8
0.7
0.6
min
Annotated genes
smORFs
Species-specific (orphan) ORFs
3,457 RNA-Seq-Runs, independently sorted for each transcript by mean level of accumulation

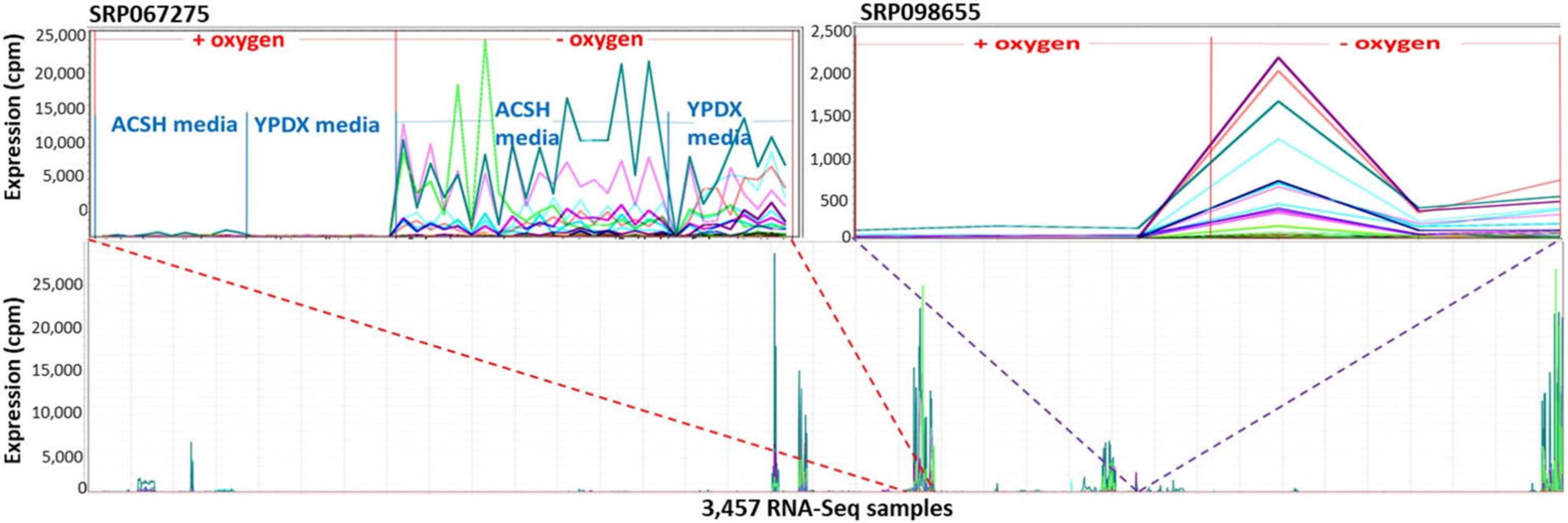

Counts of ORFs
0
500
1000
1500
3,457 RNA-Seq samples
0
500
1000
1500
Top 50 samples

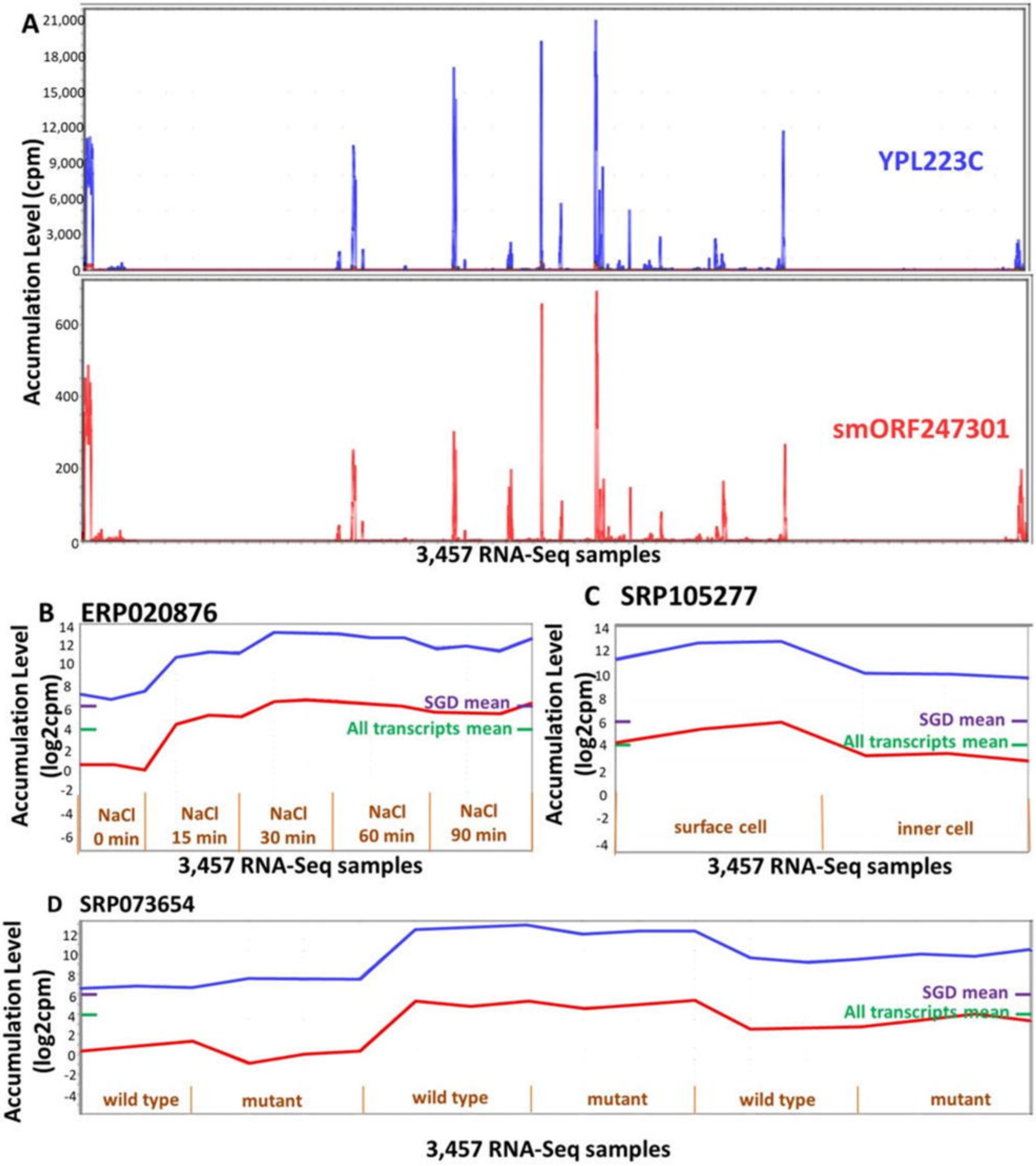

0.8
0.6
0.4
0.2
0.0
density
Type
Q3-transcribed genus-specific & conserved-ORFs (4,567)
Q3-transcribed orphan-ORFs (3,626)
genus-specific & conserved SGDs (6,419)
orphan SGDs (273)
0
12
148
1,808
22,206
mean cpm


A.
Mean raw counts/transcript across 302 Ribo-Seq samples
1500
1000
500
0
a
b
c
c
d
d
Orphan SGD-annotated genes (273)
Genus-specific SGD-annotated genes (697)
Conserved SGD-annotated genes (5,722)
Q3-transcribed orphan-ORFs (3,626)
Q3-transcribed genus-specific ORFs (3,709)
Q3-transcribed conserved ORFs (858)
B.
# Q3-transcribed ORFs with translation evidence
1000
750
500
250
0
ORF group
W/in interval between CDSs
Overlap CDS, reverse orientation
Overlap CDS, same orientation
Within CDS, reverse orientation
Within CDS, same orientation
Reverse orientation
A
Divergent
B
Convergent
Same orientation
Orphan-ORFs
Genus-specific ORFs
Conserved ORFs


random
experimental
0.019
BP
random
experimental
0.023
CC
random
experimental
0.027
MF
Counts
10.0
7.5
5.0
2.5
0.0
0.00
0.04
0.08
0.12
Best p-value


Cluster 112
YOR394W
YDR542W
YMR325W
YPL282C
txCDS15253
I_22
smorf75360
XII_223
YBR301W
YGR294W
YPL272C
VIII_594
YAL064W-B
YLR461W
YEL049W
V_959
smorf2748
II_194
XVI_35
XII_1486
YAL067W-A
VII_2005
YJR150C
YFL020C
YIR040C
smorf2421
YNR076W
V_163
YER011W
II_1582
YLL025W
IX_450
smorf40513
YLL064C
smorf218022
YDR543C
XIII_147
VI_505
IV_506
YHL050W-A
IV_1
orphan annotated genes
orphan ORFs
conserved annotated genes
conserved ORFs
Edge Color
0.6
1.0


SRP067275
Expression (cpm)
25,000
20,000
15,000
10,000
5,000
0
+ oxygen
- oxygen
ACSH media
YPDX media
ACSH media
YPDX media
SRP098655
2,500
2,000
1,500
1,000
500
0
+ oxygen
- oxygen
Expression (cpm)
25,000
20,000
15,000
10,000
5,000
0
3,457 RNA-Seq samples


A
21,000
18,000
15,000
12,000
9,000
6,000
3,000
0
YPL223C
Accumulation Level (cpm)
600
400
200
0
smORF247301
3,457 RNA-Seq samples
B
ERP020876
14
12
10
8
6
4
2
0
-2
-4
-6
Accumulation Level (log2cpm)
SGD mean
All transcripts mean
NaCl 0 min
NaCl 15 min
NaCl 30 min
NaCl 60 min
NaCl 90 min
3,457 RNA-Seq samples
C
SRP105277
14
12
10
8
6
4
2
0
-2
-4
Accumulation Level (log2cpm)
SGD mean
All transcripts mean
surface cell
inner cell
3,457 RNA-Seq samples
D
SRP073654
12
10
8
6
4
2
0
-2
-4
Accumulation Level (log2cpm)
SGD mean
All transcripts mean
wild type
mutant
wild type
mutant
wild type
mutant
3,457 RNA-Seq samples