BioRxiv
https://doi.org/10.1101/xxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Recycling RNA-Seq Data to Identify Candidate Orphan Genes for Experimental Analysis

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

[1]Genetics and Genomics Graduate Program, Iowa State University, Ames, 541004, US,
[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, US and
[4]Center for Metabolic Biology, Iowa State University, Ames, 50014, US

*Correspondence: evewurtele@gmail.com

Abstract

Motivation: Each organism contains genes with no protein homolog in other species ("orphan genes"). Some of these have arisen *de novo* from non-genic material, while others may be the result of ultra-rapid mutation of existing genes. The challenges of identifying orphan genes and predicting their functions are immense, resulting in under-appreciation of their importance. The yeast genome expresses thousands of transcripts, many that contain ORFs that are translated, that are not annotated as genes. Here, we apply computational approaches to re-cycle and re-evaluate massive raw public RNA-Seq data to identify those ORFs that are the best candidates to represent orphan genes.

Results: We created a pooled, aggregated RNA-Seq dataset from the raw reads and metadata of over 3,400 RNA-seq samples from 172 studies in the NCBI-Sequence Read Archives (SRA) database (Leinonen *et al.*, 2011), and realigned these reads to a transcriptome consisting of the Saccharomyce Genome Database ((Cherry *et al.*, 1998), SGD)-annotated genes and 29,354 unannotated ORFs of the *Saccharomyces cerevisiae* genome. Phylostratigraphy analysis of the predicted proteins from the 29,354 non-annotated open reading frames (ORFs) in the *S. cerevisiae* genome inferred: 15,806 are orphans ("orphan-ORFs"), 11,942 are genus-specific, and 1,606 are more highly conserved. These RNA-Seq data reveal over 150 of transcripts containing orphan encoding-ORFs with mean levels of expression across all samples comparable to half of annotated non-orphan genes. Most orphan-encoding ORFs are highly expressed only under limited conditions. We built a co-expression matrix from the transcription dataset, and optimized partitioning by Markov Chain Clustering. The MCL clustering result is significant different from random clusters based on GO enrichment analysis to show the biological significance. Over 3,000 significant GO terms (p-value<0.05) were found in the clusters, and plenty of unannotated ORFs were found highly correlated (PCC > 0.8) to annotated genes. For example, cluster 112 is composed of seripauperin genes, and smORF247301 is correlated to YPL223C with a 0.95 Pearson correlation. We provide the results of the optimized aggregate-data analysis in a tool that can be used for powerful statistical analysis and visualization of specific transcripts under user-selected conditions. This approach maximizes a user's ability to view potential interactions across experimental perturbations, and provides a rich context for experimental biologists to make novel, experimentally-testable hypotheses as to potential functions of as yet unannotated transcripts.

Contact: evewurtele@gmail.com

Supplementary information: Supplementary data are available at *BioRxiv* online.

1 Introduction

The transcription of unannotated genome sequence is evidenced by transcriptomic studies in which raw reads do not align to known genes (Wu and Knudson, 2018; Pertea *et al.*, 2018b; Delcourt *et al.*, 2018; Struhl, 2007; Lu *et al.*, 2017). Many dismiss this genome expression as "noise" (Pertea *et al.*, 2018b,a; Consortium *et al.*, 2012; Barroso *et al.*, 2018; Lloréns-Rico *et al.*, 2016), However, some of this sequence are translated (Wu and Knudson, 2018; Carvunis *et al.*, 2012; Smith *et al.*, 2014; Olexiouk *et al.*, 2017; Hsu *et al.*, 2016; Ruiz-Orera *et al.*, 2015; Chew *et al.*, 2013), and several functional genes have been identified from it (Ji *et al.*, 2015; Andrews and Rothnagel, 2014). To our knowledge, the levels and patterns of transcription of these unannotated transcripts have not been globally evaluated. Genes that encode species-species proteins (orphan genes) (Arendsee *et al.*, 2019b) are among those least likely to be annotated. Orphan genes may evolve *de novo* from non-coding sequence in regions of the genome lacking genes entirely, or they may evolve

as new reading frames within existing genes (Ruiz-Orera *et al.*, 2015; Arendsee *et al.*, 2019a; Oss and Carvunis, 2019; Tautz and Domazet-Lošo, 2011). Alternately, existing proteins could be highly altered and converted to orphans (Tautz and Domazet-Lošo, 2011; Khalturin *et al.*, 2009), a process that would require selective ultra-rapid mutation. The process by which such ultra-rapid mutation might occur is, to our knowledge, as-yet undefined.

Many annotated orphan genes are up-regulated by an environmental stress; spatially or developmentally localized; or associated with species-specific traits or regulatory patterns (Li *et al.*, 2009; Guo *et al.*, 2007; Colbourne *et al.*, 2011; Arendsee *et al.*, 2014). We propose this is also likely a characteristic of as-yet unnanotated orphans ("orphan-ORFs").

Most annotations of genomes identify genes by combining evidence of homology with genes of other species (Meyer and Durbin, 2004; Proux-Wéra *et al.*, 2012) and *ab initio* gene prediction via canonical motifs (e.g., splice junctions) (Cantarel *et al.*, 2008; Hoff *et al.*, 2016). Homology cannot identify orphan genes, and it is quite possible that *ab initio* gene prediction would not be very effective, as orphans may not yet have evolved canonical motifs.

Here, we introduce an approach and tools to reuse and re-mine aggregated RNA-Seq data from public databases to identify transcribed ORFs and to construct a co-expression network. Integrating these with co-expression networks, GO enrichment analysis and RNA-Seq metadata offers a perspective on which to base experimentally-testable biological hypotheses for unknown genes.

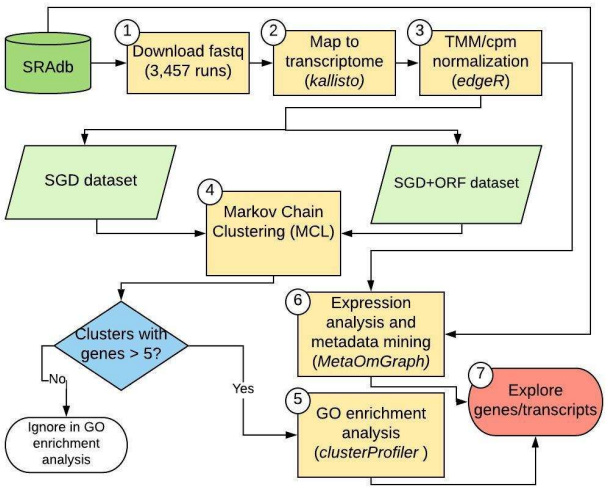


Fig. 1. RNA-Seq data analysis work flow. 1)Raw reads and metadata for 3,457 samples of RNA-Seq were downloaded from SRA. 2)RNA-Seq reads were mapped to transcriptome and quantified abundance by kallisto. 3) TMM normalization was applied to minimize interference of the data with cpm as units edgeR. 4) MCL clustering analysis for SGD dataset and SGD+ORF dataset (all SGD genes and smORFs plus top 25% highly expressed ORFs ranked by mean expression across 3,457 samples). 5) GO enrichment analysis of clusters. 6) Transcript expression statistical analysis, visualization in MetaOmGraph. 7) Combined information to explore transcript function.

2 Methods

2.1 Orphan detection in *S. cerevisiae*

We used the automated R software for phylostratography, *phylostratr* (Arendsee *et al.*, 2019b), to infer the phylostrata of CDSs and ORFs of the *S. cerevisiae* genome. CDSs were obtained from SGD. Unannotated ORFs (⩾150 nt) were extracted by *bedtools2* (Quinlan and Hall, 2010), translated by *emboss* (Rice *et al.*, 2000), yielding 24,912 ORFs; to these we added ORFs less than 150 nucleotides: the set of 1,139 smaller translated sequences (smORFs) identified by ribosome profiling (Carvunis *et al.*, 2012) and 3,303 ORFs identified by TIF-Seq (Lu *et al.*, 2017) (txCDS) that were not included among the ORFs due to small size (⩾150 nt). The analysis compared the proteins predicted from the *S. cerevisiae* CDSs and 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 avoid false positives when identifying orphan ORFs from *S. cerevisiae*, we extracted the translation products from all ORFs (⩾150 nt) from each of the six *Saccharomyces* genomes, and added these to the annotated proteins. (See Figure S1 for workflow, *species.xlsx* for full species list, and *phylostratr_heatmap.pdf* for gene by gene results). Each gene is assigned to the most evolutionarily-distant phylostratum that contains an inferred homolog. A gene is inferred as an orphan if it is assigned a phylostratum level as *S. cerevisiae* (ps=15).

2.2 Raw read processing and data normalization

RNA-Seq raw reads were pseudoaligned to a transcriptome index of SGD-annotated genes plus unannotated ORFs and quantified by *kallisto* (Weijers *et al.*, 2012) (see *S.cerevisiae_RNA-seq.mog* for RNA-Seq metadata). The ORFs over 150 nt with less were filtered out before making the pairwise PCC networks; all SDG genes and smORFs were retained regardless of expression values.

We evaluated the performance of two diverse methods for normalizing raw counts (*Section 5, Supplementary_Material.pdf*) and three Pearson's correlation coefficients (PCC) cutoffs (*Section 3, Supplementary_Material.pdf*) using Markov chain graph clustering (MCL (Dongen and Marinus, 2000)) followed by GO term enrichment analysis. The ORFs with expression values in lowest 75% quartile were removed from the data, and networks were created from the SGD and SGD+ORF datasets by calculating the pairwise Pearson correlation matrix across all 3,457 samples.

Our consideration of which normalization methods to use was twofold. Based on the evaluation of Dillies *et al.* (2013), we used *edgeR* (Robinson *et al.*, 2010). We also normalized the raw data by a single cell RNA-seq normalization *SCnorm* (Bacher *et al.*, 2017). This method examines sequence information from individual cells with the aim to provide a higher resolution of cellular differences. Two features of single cell RNA-Seq analysis are similar to analysis of the orphan-ORF-focused multi-study RNA-Seq data here: 1) raw counts contain an abundance of zero-expression values, 2) there is high technical variability among samples.

Comparison of GO enrichment results led to selection of *edgeR* normalization for further analyses. Networks obtain from PCC cutoffs of 0.6, 0.7 and 0.8 were then evaluated (*Section 3, Supplementary_Material.pdf*).

2.3 Cluster evaluation by Gene Ontology (GO) term enrichment analysis

Clustering results from each of the eight MCL analyses were evaluated by GO enrichment analysis; only clusters with over five genes were considered(*Section 2, Supplementary_Material.pdf*). The GO term representation of each experimental result was compared to that of 100 random sets of clusters that were obtained by permuting gene IDs. For these permutations, the same number of clusters of the same size as the those from the experimental result were assigned to each random set (Mentzen and Wurtele, 2008). The best adjusted p-value p_{min} was recorded for the

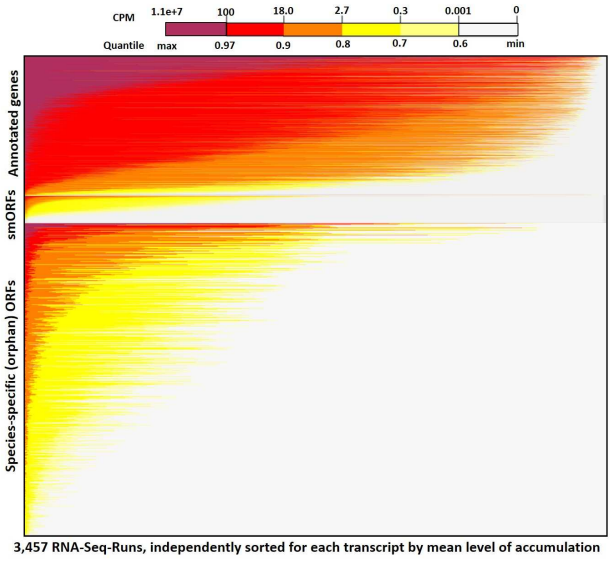


Fig. 2. RNA-Seq expression heatmap across 3,457 samples. Top panel, SGD annotated genes (6,692); middle panel, smORFs (1,139); bottom panel, species-specific ORFs (orphan-ORFs) (15,805). Each group of transcripts is ordered independently by its mean cpm across the 3,457 samples. Each transcript is independently sorted by its cpm. (See Figure S7 for whole expression plot). Many orphan-ORFs have expression comparable to annotated genes

enriched GO terms in each cluster. Each random cluster set was assigned a score S_i, which is the average p_{min} across all clusters in the set.

$$S_i = \frac{\sum_{j=1}^{n} p_{min\ j}}{n} \tag{1}$$

where n indicates the number of clusters (n=49). 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 significantly better than any of the random scores.

3 Results and Discussion

3.1 Identifying potential cryptic orphan genes in *Saccharomyces cerevisiae*

Within the *Saccharomyces sensu stricto* clade, *S. cerevisiae* is the most researched and has the most extensively sequenced and annotated genome. Despite this significant body of research on *S. cerevisiae*, the yeast genome expresses thousands of transcripts that are not annotated as genes (Carvunis *et al.*, 2012; Lu *et al.*, 2017; Oss and Carvunis, 2019; Wu and Knudson, 2018; Smith *et al.*, 2014; Pelechano *et al.*, 2013; Lu *et al.*, 2017), some supported with translational evidence and many of *de novo* origin(Oss and Carvunis, 2019; Lu *et al.*, 2017).

To delineate those ORFs that could encode orphan proteins, we inferred the the oldest phylostratum (PS) to which each *S. cerevisiae* protein (or candidate protein) can be traced using *phylostratr* (Arendsee *et al.*, 2019b), with similarity to proteins of cellular organisms (including prokaryotes) designated as PS1, and no similarity to any protein outside of *S. cerevisiae* being PS15. (See MOG file *S.cerevisiae_RNA-seq.mog* for complete PS assignments by transcript). By this analysis, less than 4% of SGD-annotated genes are inferred as orphans, and 54% of unannotated ORFs are orphans (orphan-ORFs). More than 80% of the conserved ORFs are genus-specific.

The mean CDS length of genes generally increases during evolution, with orphan CDSs being the shortest on average (Arendsee *et al.*, 2014; Palmieri *et al.*, 2014; Arendsee *et al.*, 2019b; Oss and Carvunis, 2019; Toll-Riera *et al.*, 2009) (Figure S4.A). The unannotated orphan-ORFs follow a similar trend, average length is also shorter than that of SGD-annotated conserved genes (Figure S4.A). Similar to the finding of (Basile *et al.*, 2017), the mean GC content for annotated orphans in *S. cerevisiae* is slightly lower than that of more conserved genes; interestingly, Vakirlis *et al.* (2018) report a higher GC content among those orphan genes of confirmed *de novo* origin. The unannotated orphan-ORFs also have a slightly lower mean GC content than those of other phylostratum levels (Figure S4.B).

3.2 Functional inference

To infer potential functions of the yeast orphan genes and orphan-ORFs through co-expression with genes of known function, we downloaded from SRA raw sequence reads of 3,457 RNA-Seq samples from 177 studies. The experimental variables across these samples include a wide variety of mutants, chemical treatments, stresses, nutrition deprivations, and growth stages. We mapped the RNA-Seq reads to a yeast transcriptome composed of SGD-annotated genes and ORFs, and quantified abundances. Then, we analyzed the expression patterns of 29,354 transcripts in the context of the 6,692 annotated genes of *S. cerevisiae* across an aggregated dataset of 3,457 RNA-Seq samples, consisting of the re-processed, *edgeR*-normalized raw reads from SRA and the corresponding metadata.

Figure 2 compares the expression for the matrix of 3,457 RNA-Seq samples for SGD genes, smORFs, and orphan-ORFs (See Figure S7 for full expression plot). The mean expression for all annotated genes across all data is 38 cpm, whereas the mean expression for the 25% most highly expressed ORFs is 18 cpm. Many SGD-annotated genes are expressed in most of the samples. In contrast, most ORFs have very low expression in most samples but accumulate more highly in a few samples. This contributes to the low mean expression of many of the ORFs.

Across 3,457 RNA-Seq runs, conserved SGD annotated genes have significantly higher mean expression than orphan SGD genes and unannotated ORFs (Figure S8.A). Over 600 of unannotated orphan-ORFs have higher mean expression than 10% of conserved SGD genes. Moreover, 36 of orphan-ORFs have mean expression higher than 90% conserved SGD genes (Table S4.A). ORFs with transcription evidence from our study and Lu *et al.* (2017) have higher mean expression than

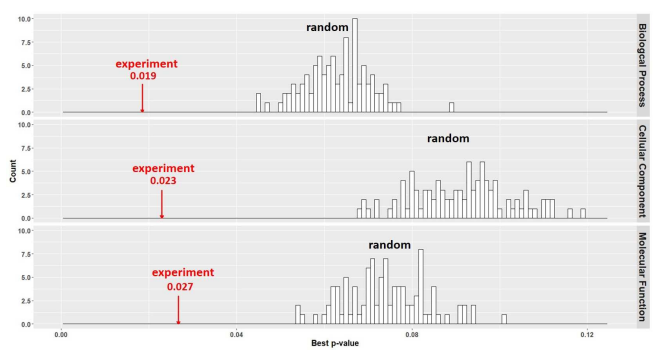


Fig. 3. GO enrichment analysis of experimental data and random test distribution Optimized MCL partitions of the SGD+ORF dataset. Best p-values (mean of the lowest adjusted p-values for GO terms) were determined across all clusters of the experimental data and each random set. (Supplementary_Material.pdf, Sections 2 and 3). Red arrow, experimental data. Black bars, best p-value of 100 randomly-obtained permutations with size and number of clusters identical to experimental data. The clustering result is significantly better for experimental data than any random sets.

smORFs with translation evidence (Carvunis *et al.*, 2012) (Figure S8.B), which indicates these expressed ORFs may not be transcription noise.

3.3 Network inference and MCL clustering of the inferred expression network

Orphan genes have no homologs in other species and are likely to have less canonical structure, hence we cannot use standard gene prediction tools to identify them or infer their functions (Arendsee *et al.*, 2014). Based on the assumption that genes with similar patterns of expression across different conditions are likely to encode proteins with related functions or be involved in the same pathway, we inferred functions for those orphan genes that are co-expressed with genes of known function across many conditions.

We used two datasets for network inference. One network, "SGD genes" is from the 6,692 SGD-annotated genes. The other network, "SGD+ORF", contains 14,885 transcripts in total and includes all SGD-annotated genes, 1139 smORFs, and 7,054 ORFs (comprising the 25% most highly-expressed as ranked by mean expression across 3,475 samples).

To identify sets of co-expressed genes, we used a high-throughput graph-clustering pipeline based on flow simulation (Markov chain graph clustering, MCL (Dongen and Marinus, 2000)) to identify clusters of co-expressed genes. After parameter optimization (see Figure S2, S3, and Table S1) a final network derived from the full SGD+ORF dataset was partitioned by MCL into 544 clusters (Table S2). Since clusters of $<$ five transcripts are problematic for functional analysis, we focused on those clusters with at least 5 genes. Forty-six percent of the 273 SGD genes that have been designated as orphans(Arendsee *et al.*, 2019b) are members of clusters containing more than 5 genes; 59% of the 3,899 expressed orphan-ORFs are in clusters with more than 5 genes.

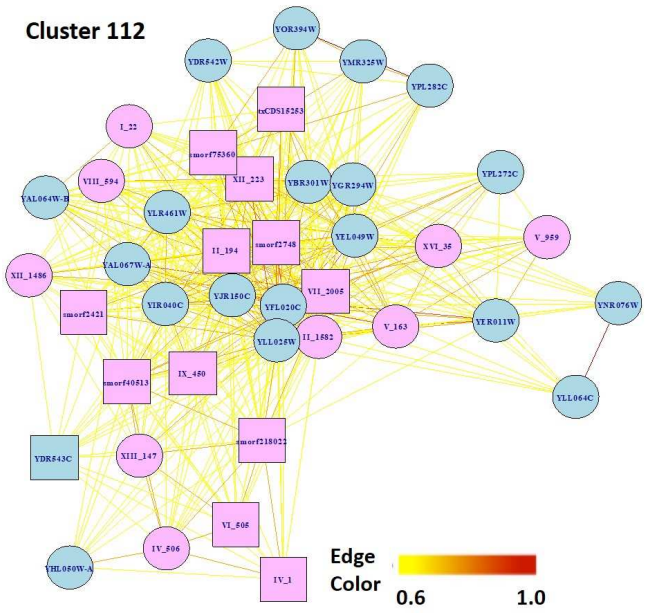
3.4 GO enrichment analysis for clustering results

In order to evaluate the significance of the clustering results, we compared the extent of significant enrichment of Gene Ontology (GO) terms in the set of the clusters of more than 5 genes derived from MCL clustering of experimental data with that of 100 randomly-obtained sets of clusters. For each random set, the number of clusters and the number of genes in each cluster are held the same as 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 (Mentzen and Wurtele, 2008) (Figure 3). 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 3). For each GO ontology category (Biological Process (BO), 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 randomized datasets. This result indicates that the MCL clustering of genes derived from transcriptomic expression is not random, and thus the genes in a given cluster might function in the same or related biological process.

3.5 Exploring Gene Function

As proof-of-concept for the utility of these data, we provide three examples of functional inference combining MCL clustering analysis, GO enrichment analysis, and RNA-Seq expression across multiple conditions. Metadata about each transcript includes: functional annotations (from SGD), CDS/ORF GC content, CDS/ORF length, genomic positional coordinates, orientation, phylostratal assignment, MCL-cluster membership. Metadata about each sample includes: 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. (See *S.cerevisiae_RNA-seq.mog* for the complete expression matrix and gene/ORF and RNA-Seq sample metadata).

In many cases, co-expression clusters are composed of genes and ORFs located across widely diverse regions of the genome. MCL Cluster 112 contains 20 SGD-annotated genes and 21 unannotated ORFs (Figure 4, top panel). Twelve of the genes are in the seripauperin (PAU) family. PAU-rich gene clusters have been identified in independent microarray studies using distinct data sets(Orellana *et al.*, 2014; Magwene and Kim, 2004). Many PAU family members are induced by low temperature and anaerobic conditions, and repressed by heme (Rachidi *et al.*, 2000) and individual PAU proteins confer resistances to biotic and abiotic stresses (Rivero *et al.*, 2015); their molecular function is not known. PAU proteins are localized to the cell wall, plasma membrane, endoplasmic reticulum and vacuole.



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	0.001985
CC	extracellular region	0.006176
CC	anchored component of membrane	0.01868

Fig. 4. Case study: Genes and ORFs grouped into Cluster 112 by MCL partitioning. Cluster 112 contains 12 seripauperin genes. Top, network view of cluster 112. Edge colors, Pearson correlations of 0.6 to 1.0; light blue nodes, SGD-annotated genes; pink nodes, unannotated transcripts; square nodes, orphans; circular nodes, conserved. Bottom, significantly-enriched GO terms. Fourteen of the 19 conserved annotated genes in Cluster 112 are classified by one or more GO terms in the table. (See cluster_example.xlsx and S.cerevisiae_RNA-seq.mog for complete clustering and ontology results.

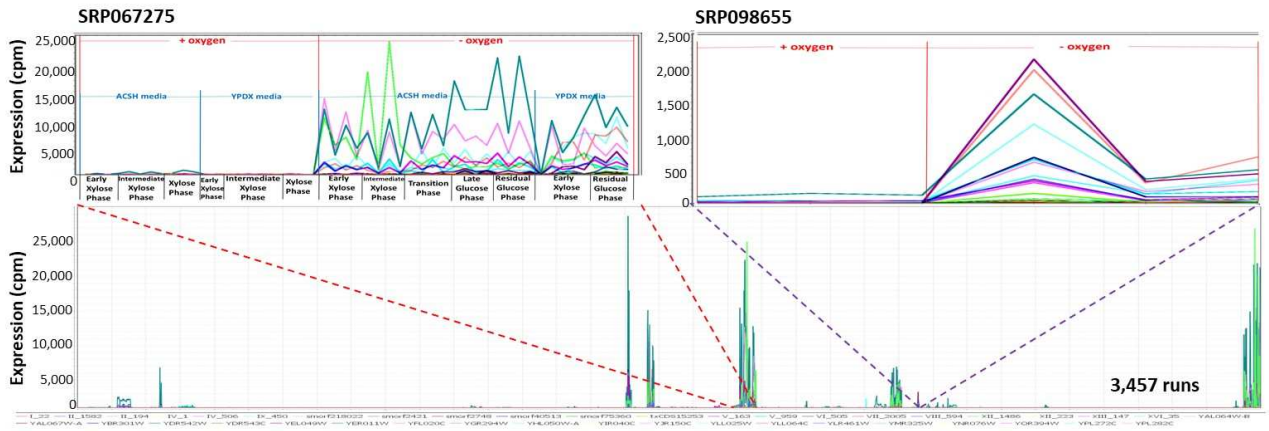


Fig. 5. The 41 genes and ORFs in cluster 112 respond to anoxia. US should we mention the visualizations are produced with MOG Expression patterns of the 41 genes and ORFs in cluster 112 are similar across multiple conditions. (See geneInfo_cluster112and317.xlsx for gene/ORF-list.) X-axis, 3,457-samples, 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 Study SRP067275: "Sequence, annotate, and analyze the genome of GLBRCY22-3, a stress tolerant strain of S. cerevisiae with biofuels potential". 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 containing 60 g/L dextrose and 30 g/L xylose. Top right inset, zoom-in to Study SRP098655: "Lipid engineering reveals regulatory roles for membrane fluidity in yeast fluctuation and oxygen-limited growth". The 41 genes and ORFs of cluster 112 are up-regulated in response to anoxia, regardless of changes in growth media.

YER011W and YJR150C, also in Cluster 112, are localized to the same cellular compartments as seripauperin and have been shown to be induced under anaerobic conditions (Sertil *et al.*, 1997; Kowalski *et al.*, 1995; Kitagaki *et al.*, 1997; Cohen *et al.*, 2001). The other genes in this cluster have no functional annotation. GO enrichment analysis identified eight GO terms to be significantly-over-represented in this cluster (Figure 4, bottom panel). Exploration and visualization of in Cluster 112 in MetaOmGraph (MOG) (Singh *et al.*, 2019). Figure 5 shows the expression of the genes and ORFs across all 3545 samples of the RNA-Seq dataset, and highlights the two studies in this dataset 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 (McIlwain *et al.*, 2016) (Figure 5, 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 (Degreif *et al.*, 2017) (Figure 5, 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 play a role in cellular response to this stress.

In several cases, ORFs are located near an existing gene and also share a similar transcription pattern. This is the case for smORF247301, which is 77nt upstream of YPL223C (Figure 6). MOG analysis indicates smORF247301 and SGD gene YPL223C have a Pearsons correlation 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 (Garay-Arroyo *et al.*, 2000), and induced by osmotic, ionic, oxidative, heat shock and heavy metals stresses. Three independent stress studies (Figure 6 B-D) show increased expression of smORF247301 and YPL223C in response to osmotic, heat and desiccation stresses. smORF247301 has translation evidence through ribosomal profiling(Carvunis *et al.*, 2012).

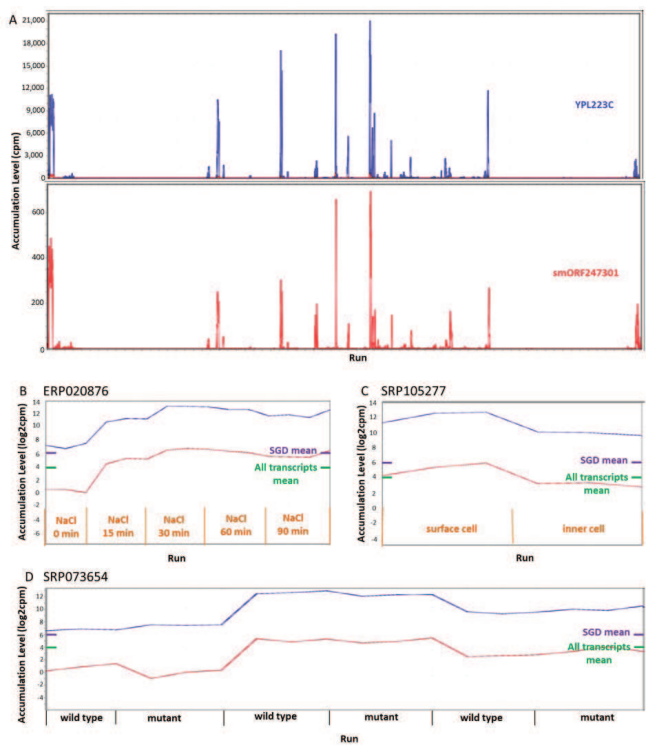


Fig. 6. Expression patterns of smORF247301 and YPL223C smORF247301 and YPL223C are located on nearby regions of chromosome 16 in opposite orientations. A) Expression patterns are similar (Pearson correlation, 0.95) across 3,457 samples. B-D, Expression patterns for smORF247301, YPL223C, and mean of SGD genes and all genes/ORFs in three studies: (B) ERP020876, salt stress; (C) SRP073654: wild type and mutant strains in heat and osmotic stress; (D)SRP105277, surface and invasive cells of biofilms.

One possibility is that the transcription and translation of smORF247301 is "noise" (Eling *et al.*, 2019) associated with the expression of the nearby

YPL223C. Another possibility is that smORF247301 is a young, not-yet-annotated gene that is "piggybacking" on the expression apparatus of YPL223C. In that case, the co-expression of these two transcripts could reflect a mechanism whereby a new gene can emerge provided with a ready-made exposure to transcription factors associated with open chromatin during transcription of a nearby established gene (YPL223C), evolving to confer a survival advantage, e.g., the stress response pathway of YPL223C.

A variety of methods can be used for network inference and partitioning. Each combination of inference and partitioning methods provides complementary information, revealing different types of relationships. Networks can be inferred by correlation among expression values (Ruan *et al.*, 2010; Luo *et al.*, 2007; Mentzen and Wurtele, 2008), by mutual information (Zhang *et al.*, 2012) or by Relatedness approaches (Netotea *et al.*, 2014). Pearson correlation, as in this paper, is highly sensitive at extracting genes whose expression is linearly correlated across multiple conditions, but it misses many more complex, non-linear relationships. To our knowledge, the strengths and weaknesses provided by the various network partitioning methods for extracting different types of biological information has not been investigated. Thus, the identical RNA-Seq expression data can be reanalyzed and further explored to reveal different biological interactions.

4 Conclusion

In this study we have 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. Fourty-nine ORFs had no detectable expression in any sample. Of the 8,193 expressed ORFs, 48% were inferred to be "orphan-ORFs", encoding *S. cerevisiae*-specific proteins. Despite a tendency to be expressed only under specific conditions, 163 of the orphan-ORFs had mean levels of expression greater than that of the median SGD-annotated gene. Over 2,000 of the orphan-ORFs are members of gene clusters identified by MCL partitioning of the network. We postulate that among these ORFs are functional genes, and suggest an approach to hone in on these for experimental analysis.

The SGD+ORF dataset assembled herein represents the expression of SGD-annotated genes and unannotated ORFs under multiple conditions in a readily explorable, user-friendly format. 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.

Acknowledgements

We thank Drs. 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.

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

Andrews, S. J. and Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, **15**(3), 193.

Arendsee, Z. *et al.* (2019a). fagin: High resolution contextual phylostratigraphy of young genes. *Accepted in BMC Bioinformatics*.

Arendsee, Z. *et al.* (2019b). phylostratr: a framework for phylostratigraphy. *Bioinformatics*.

Arendsee, Z. W. *et al.* (2014). Coming of age: orphan genes in plants. *Trends in Plant Science*, **19**(11), 698–708.

Bacher, R. *et al.* (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, **14**(6), 584–586.

Barroso, G. V. *et al.* (2018). The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics*, **208**(1), 173–189.

Basile, W. *et al.* (2017). High GC content causes orphan proteins to be intrinsically disordered. *PLOS Computational Biology*, **13**(3), e1005375.

Cantarel, B. L. *et al.* (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**(1), 188–196.

Carvunis, A.-R. *et al.* (2012). Proto-genes and *de novo* gene birth. *Nature*, **487**(7407), 370–374.

Cherry, J. M. *et al.* (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, **26**(1), 73–79.

Chew, G.-L. *et al.* (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development*, **140**(13), 2828.

Cohen, B. D. *et al.* (2001). 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 Research*, **29**(3), 799–808.

Colbourne, J. K. *et al.* (2011). The Ecoresponsive Genome of Daphnia pulex. *Science*, **331**(6017), 555–561.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Degreif, D. *et al.* (2017). Lipid engineering reveals regulatory roles for membrane fluidity in yeast flocculation and oxygen-limited growth. *Metabolic Engineering*, **41**, 46–56.

Delcourt, V. *et al.* (2018). Small proteins encoded by unannotated orfs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mrna. *Proteomics*, **18**(10), 1700058.

Dillies, M.-A. *et al.* (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, **14**(6), 671–683.

Dongen, V. and Marinus, S. (2000). Graph clustering by flow simulation. *Doctoral dissertation*, page 1.

Eling, N. *et al.* (2019). Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, page 1.

Garay-Arroyo, A. *et al.* (2000). Highly Hydrophilic Proteins in Prokaryotes and Eukaryotes Are Common during Conditions of Water Deficit. *Journal of Biological Chemistry*, **275**(8), 5668–5674.

Guo, W.-J. *et al.* (2007). Significant Comparative Characteristics between Orphan and Nonorphan Genes in the Rice (Oryza sativa L.) Genome.

Hoff, K. J. *et al.* (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**(5), 767–769.

Hsu, P. Y. *et al.* (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proceedings of the National Academy of Sciences*, **113**(45), E7126–E7135.

Ji, Z. *et al.* (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, **4**, e08890.

Khalturin, K. *et al.* (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, **25**(9), 404–413.

Kitagaki, H. *et al.* (1997). Identification and Analysis of a Static Culture-Specific Cell Wall Protein, Tir1p/Srp1p in Saccharomyces Cerevisiae. *European Journal of Biochemistry*, **249**(1), 343–349.

Kowalski, L. R. *et al.* (1995). Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae. *Molecular Microbiology*, **15**(2), 341–353.

Leinonen, R. *et al.* (2011). The Sequence Read Archive. *Nucleic Acids Research*, **39**(suppl_1), D19–D21.

Li, L. *et al.* (2009). Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal*, **58**(3), 485–498.

Lloréns-Rico, V. *et al.* (2016). Bacterial antisense rnas are mainly the product of transcriptional noise. *Science advances*, **2**(3), e1501363.

Lu, T.-C. *et al.* (2017). A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Molecular Biology and Evolution*, **34**(11), 2823–2838.

Luo, F. *et al.* (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, **8**(1), 299.

Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, **5**(12), R100.

McIlwain, S. J. *et al.* (2016). Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research. *G3 (Bethesda, Md.)*, **6**(6), 1757–1766.

Mentzen, W. I. and Wurtele, E. S. (2008). Regulon organization of Arabidopsis. *BMC Plant Biology*, **8**(1), 99.

Meyer, I. M. and Durbin, R. (2004). Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Research*, **32**(2), 776–783.

Netotea, S. *et al.* (2014). ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**(1), 106.

Olexiouk, V. *et al.* (2017). An update on sorfs.org: a repository of small orfs identified by ribosome profiling. *nar*, **46**(D1), D497–D502.

Orellana, M. *et al.* (2014). 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 Research*, **14**(3), 412–424.

Oss, S. B. V. and Carvunis, A.-R. (2019). De novo gene birth. *PLOS Genetics*, **15**(5), e1008160.

Palmieri, N. *et al.* (2014). The life cycle of Drosophila orphan genes. *eLife*, **3**, e01311.

Pelechano, V. *et al.* (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**(7447), 127–131.

Pertea, M. *et al.* (2018a). Chess: a new human gene catalog curated from thousands of large-scale rna sequencing experiments reveals extensive transcriptional noise. *Genome biology*, **19**(1), 208.

Pertea, M. *et al.* (2018b). Thousands of large-scale rna sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv*, page 332825.

Proux-Wéra, E. *et al.* (2012). A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics*, **13**(1), 237.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Rachidi, N. *et al.* (2000). Saccharomyces cerevisiae PAU genes are induced by anaerobiosis. *Molecular Microbiology*, **35**(6), 1421–1430.

Rice, P. *et al.* (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**(6), 276–277.

Rivero, D. *et al.* (2015). Hsp12p and pau genes are involved in ecological interactions between natural yeast strains. *Environmental microbiology*, **17**(8), 3069–3081.

Robinson, M. D. *et al.* (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Ruan, J. *et al.* (2010). A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biology*, **4**(1), 8.

Ruiz-Orera, J. *et al.* (2015). Origins of de novo genes in human and chimpanzee. *PLoS genetics*, **11**(12), e1005721.

Sertil, O. *et al.* (1997). The DAN1 gene of S. cerevisiae is regulated in parallel with the hypoxic genes, but by a different mechanism. *Gene*, **192**(2), 199–205.

Singh, U. *et al.* (2019). Metaomgraph: a workbench for interactive exploratory data analysis of large expression datasets. *bioRxiv*, page 1.

Smith, J. E. *et al.* (2014). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell reports*, **7**(6), 1858–1866.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by rna polymerase ii. *Nature structural & molecular biology*, **14**(2), 103.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, **12**(10), 692–702.

Toll-Riera, M. *et al.* (2009). Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Molecular Biology and Evolution*, **26**(3), 603–612.

Vakirlis, N. *et al.* (2018). A Molecular Portrait of De Novo Genes in Yeasts. *Molecular Biology and Evolution*, **35**(3), 631–645.

Weijers, S. R. *et al.* (2012). KALLISTO: cost effective and integrated optimization of the urban wastewater system Eindhoven. *Water Practice and Technology*, **7**(2).

Wu, B. and Knudson, A. (2018). Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio*, **9**(4), e01024–18.

Zhang, X. *et al.* (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, **28**(1), 98–104.