Detecting regulatory evolution occurring in *S. cerevisiae* and *S. paradoxus* species using kallisto and DESeq2

Ravi Patel

BIOL 8250, Spring 2017
Final Project

Introduction

S. cerevisiae is a commonly yeast used not only as a model organism in various areas of biological research, but as an instrumental cooking tool. *S. paradoxus* is the closest known yeast related to *S. cerevisiae*, and is also used as a wild comparator in various research. We use these two species of yeast to study the regulatory differences that may be driving species divergence between two closely related taxa.

In order to determine changes at the protein expression level in an organism, one can use RNA-sequencing to determine changes in gene transcription, and Ribosomal profiling, to determine changes in translation. Together, these metrics can tell a researcher how protein levels are modulated in the species of interest.

Two main forms of regulatory evolution that modulate protein levels in a cell/organism are coordinated and compensatory evolution. Coordinated evolution involves a unidirectional change in both transcription and translation (i.e., both translation and transcription change in the same direction) to increase or decrease protein levels. Compensatory evolution works to maintain protein levels by compensating a change in one (translation or transcription), with an opposite change in the other.

Quantification of mRNA and ribosomal occupancy together can be used to determine translational efficiency, which determines the amount of protein produced in a cell. Thus, we analyze RNA sequencing and Ribosomal profiling data to determine which genes are undergoing regulatory evolution and what type between the two yeast species, *S. paradoxus* and *S. cerevisiae*. Further, Gene Ontologies are used to attempt to determine what phenotypic differences are possibly being driven by this regulatory evolution.

Methods

Acquiring yeast transcriptomes

A custom python script was used to extract the coding DNA sequence (CDS) for each of the 5,474 *S. cerevisiae* protein coding genes. A BED file produced by Scannell et al. [1] provided coordinates for the 5,474 orthologous genes in *S. paradoxus*. The CDS for *S. paradoxus* genes was extracted with the same python script.

The script produced a FASTA file for each species, *S. cerevisiae* and *S. paradoxus*, with a sequence for each CDS labelled with the gene name. These FASTA files formed the respective transcriptomes for the yeast species.

Custom python scripts were also used to ensure that the chromosome and gene names were formatted identically for both *S. cerevisiae* and *S. paradoxus* for downstream analysis.

Quantifying transcript abundance

The raw RNA sequencing (RNA-seq) reads (50bp in length) for *S. cerevisiae* and *S. paradoxus* were mapped against their respective transcriptomes using kallisto [2].The resulting output provides metrics required for differential expression analysis (gene length, abundance estimates, and transcripts per kilobase-millions).

Quantifying translation abundance (Ribosomal profiling)

The raw Ribosomal profiling (Ribo-seq) reads (50bp in length) for *S. cerevisiae* and *S. paradoxus* were mapped against their respective transcriptomes using kallisto [2]. The resulting output provides metrics required for differential ribosome occupancy analysis (gene length, abundance estimates, and transcripts per kilobase-millions).

Analyzing differential expression, translation, and translational efficiency

DESeq2 [3] was used to test for differential expression, ribosomal occupancy, and translational efficiency between the two species. For expression and occupancy analyses, the model tested was simply based on the sample (i.e., *S. paradoxus* vs *S. cerevisiae*). Translational efficiency was tested using a likelihood-ratio-test (LRT) by adding a level for the assay type (RNA-seq or Ribo-seq), as well as an interaction term (RNA-seq:Ribo-seq).

Gene ontology analysis

Gene lists were made for each category of evolution (i.e., coordinated evolution - expression/translational efficiency both up/down, compensatory evolution – expression up/down, translational efficiency down/up). These lists were run through the Gene Ontology (GO) Slim Mapper on the Saccharomyces Genome Database [CITE], using the Process GO set with both manually curated and high-throughput annotation methods.

All scripts, results, and supplementary tables are available on GitHub:
https://github.com/patelravip/BIOL250-final

Results

The transcriptomes for both species contained 5,474 protein coding genes. Not all genes had mapped reads. 5,436 (99%) of protein coding genes were transcribed (as determined by mapping in at least one of the two yeast species. 5,205 (95%) of genes were being translated (as determined by mapping Ribo-seq reads) in at least one of the yeast species.

Differential expression analysis showed that 2,459 of the transcribed genes were significantly differentially expressed (FDR-corrected p<10%). There was an approximately even split of over-expressed and under-expressed genes between *S. paradoxus* and *S. cerevisiae* (52% were over-expressed in *S. paradoxus*) (Fig. 1A).

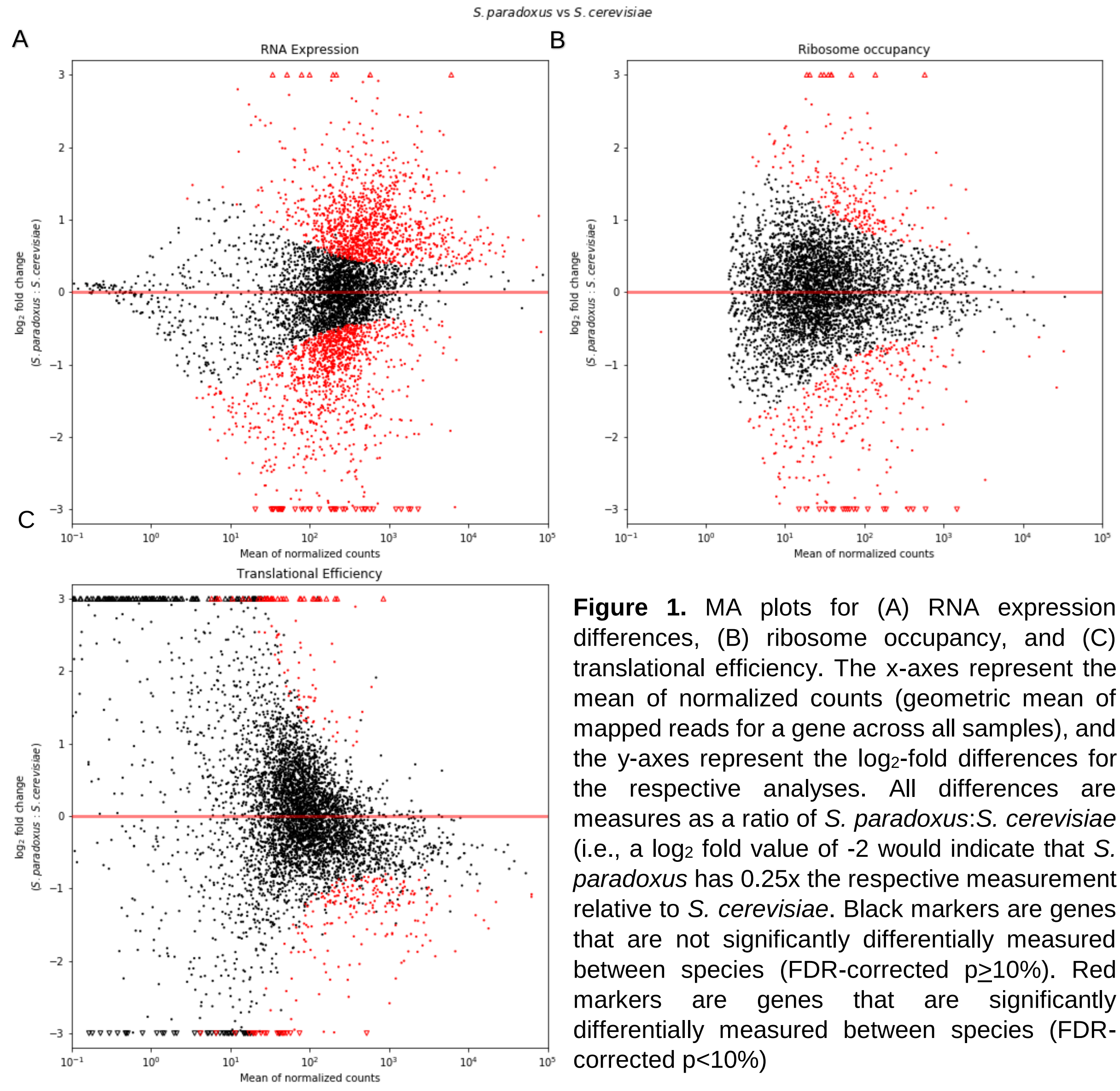

Figure 1. MA plots for (A) RNA expression differences, (B) ribosome occupancy, and (C) translational efficiency. The x-axes represent the mean of normalized counts (geometric mean of mapped reads for a gene across all samples), and the y-axes represent the $\log_2$-fold differences for the respective analyses. All differences are measures as a ratio of *S. paradoxus*:*S. cerevisiae* (i.e., a $\log_2$ fold value of -2 would indicate that *S. paradoxus* has 0.25x the respective measurement relative to *S. cerevisiae*. Black markers are genes that are not significantly differentially measured between species (FDR-corrected p≥10%). Red markers are genes that are significantly differentially measured between species (FDR-corrected p<10%)

568 genes showed significant differential ribosomal occupancy (FDR-correct p<10%) between the two species of yeast (Fig. 1B). 241 (42%) had an increase in ribosomal occupancy in *S. paradoxus* compared to *S. cerevisiae*.

The translational efficiency (TE) measures the usage of each RNA sequence for translation to protein. Only 329 of the yeast protein coding genes showed a significant difference in TE between *S. paradoxus* and *S. cerevisiae* (Fig. 1C). Of these 329 genes, 97 showed an increased in TE in *S. paradoxus*. This shows that there are many genes that are being translated at a slower rate in *S. paradoxus*.

All Genes

Table 1.

	Expression	Translational Efficiency	# of genes
Coordinated n=1917	Increase	Increase	888
	Decrease	Decrease	1029
Compensatory n=3519	Increase	Decrease	1813
	Decrease	Increase	1706

Significant (FDR < 10%)
Differential Expression AND Translational Efficiency

Table 2.

	Expression	Translational Efficiency	# of genes
Coordinated n=42	Increase	Increase	15
	Decrease	Decrease	27
Compensatory n=215	Increase	Decrease	168
	Decrease	Increase	47

Tables 1 and 2. Counts of genes in each evolution category. Table 1 contains gene classifications for all 5,436 protein coding genes. Table 2 contains classifications for genes that are both significantly differentially expressed and have different translational efficiency (FDR-corrected p<10%).

Tables 1 and 2 show the breakdown of type of evolution occurring in the genes between *S. paradoxus* and *S. cerevisiae*. In general it seems that more genes are undergoing compensatory evolution in at least one of the species to ensure protein levels (e.g., concentration) remain stable. Almost 2x more genes are undergoing compensatory than coordinated evolution when considering the entire transcriptomes, and as high as 5x more genes when considering genes with double-significance (both expression and TE are significantly different between species at FDR-corrected p<10%) (Fig. 2). Generally, there seems to be more compensatory evolution that occurs by increasing expression and decreasing TE than vice versa.

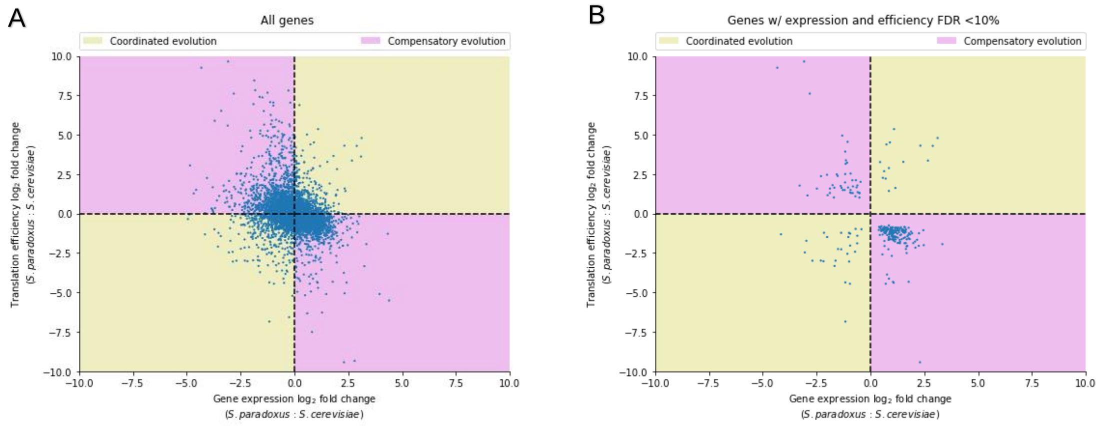

Figure 2. Plots of gene expression change between *S. paradoxus* and *S. cerevisiae* (x-axis), compared to translational efficiency change for (A) all 5,436 protein coding genes and (B) genes with significant differential expression and translational efficiency (FDR-corrected p<10%).

For genes undergoing coordinated evolution, there seem to be more genes that are decreasing overall protein product, than increasing in *S. paradoxus*.

The GO results for genes undergoing significant coordinated or compensatory evolution show that the top processes affected are generally unknown. However, carbohydrate metabolic, response to chemical, and cellular amino acid metabolic process genes consistently show up at higher proportions than found throughout the yeast genome.

Discussion

The finding that almost half of the protein coding genes were differentially expressed between *S. paradoxus* and *S. cerevisiae* was not surprising. It has been previously posited that many species level phenotypic differences can be attributed to regulatory changes, and not necessarily protein coding changes (e.g., humans and chimpanzees). The same reasoning applies to the ribosomal occupancy differences. ~10% of protein coding genes showed differential ribosomal occupancy between *S. paradoxus* and *S. cerevisiae*. This combined with the large amount of differentially expressed genes would support that the species divergence was largely impacted by regulatory evolution.

The results become surprising when the translational efficiency is put into the picture. Generally (for both significantly and non-significantly different genes) more genes showed signatures of compensatory evolution. This would imply that the amount of protein produced overall remained the same. If regulatory evolution was driving divergence, then protein levels should be affected to change the phenotype, and more genes should be undergoing coordinated evolution. This unexpected result suggests that changes at the non-regulatory level (e.g., protein-coding changes) may be playing a larger role in species differences. However, it is important to note, that protein levels still may be changing if the magnitude of expression and TE differences are not the same, such that they do not completely compensate for one another.

It is also interesting to see that of the genes undergoing compensatory evolution, more are compensating by increased expression, and decreased TE in *S. paradoxus*. This result cannot be interpreted any further however, as it cannot be known which species is actually increasing/decreasing expression or TE, only the relative differences between the two species (i.e., one cannot differentiate between an increased expression in *S. paradoxus*, or decreased expression in *S. cerevisiae*; only that *S. paradoxus* has higher expression than *S. cerevisiae*).

The genes that do show regulatory evolution can be clustered into a few main phenotypic process categories: those that affect response to the cells environment (refer to results for GO terms) and growth. Since *S. paradoxus* are often considered the wild counterparts to *S. cerevisiae*, these process terms can possibly explain some of the characteristic differences between the two species.

1. Scannell, D.R., et al., *The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus.* G3 (Bethesda), 2011. **1**(1): p. 11-25.
2. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.
3. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.