Final project: Evolution of gene regulation between two yeast species

Gene regulation is important in determining phenotype, and thus evolution often works by modulating gene regulation, rather than protein sequence. We will analyze data from two yeast species that diverged ~5 million years ago, and subsequently adapted to unique challenges (interestingly, though, they are now often found in sympatry). One of the strains we will work with is the familiar lab yeast, *Saccharomyces cerevisiae*, and the other is a wild species, *S. paradoxus.* In an important paper, Scannell let al (2011) produced new genome builds and calls of orthologous genes. Using this data, several groups have studied the sequence and regulatory evolution of many yeast species. We will make use of (part of) the dataset from McManus et al (2014), which consists of both mRNA-seq and ribosome profiling data.

The central question we will explore is how evolution occurred at the transcriptional and translational level between these two species. Recall that mRNAseq measures steady-state mRNA levels; thus, comparing mRNAseq measurements between two samples corresponds to comparing the transcription of the genes of interest. On the other hand, ribosome profiling measures the presence of ribosomes on mRNA; thus, differences in ribosome profiling between two samples reflect differences in translation between the samples. Given that protein expression will ultimately be responsible for participating in molecular and cellular processes, this means that evolution at the level of transcription can be modulated by evolution at the level of translation.

Using the dataset, you will look for differences in mRNA expression and translation efficiency between these two yeast species, to determine the role that evolution at each level of gene regulation played in the divergence of these two species. Specifically, we will look for evidence of two different “modes” of regulatory evolution:

1. Compensatory evolution, in which changes at one level are off-set by changes at another level. For instance, mRNA levels go up, but ribosome occupancy per mRNA goes down.
2. Coordinated evolution, in which both levels act in concert. For instance, mRNA levels increase, as does ribosome occupancy (and vice versa)

You can find all the data that you need for this project in the following tarball: <https://www.dropbox.com/s/e2vhceus3xflysy/final_project.tar.gz>.

The files are

1. S\_cerevisiae.fa: *S. cerevisiae* reference genome
2. S\_paradoxus.fa: *S. paradoxus* reference genome
3. S\_cerevisiae\_genes.bed: A bed file with the gene locations in *S. cerevisiae.* First column is chromosome, second column start, third column stop, fourth column is gene name, sixth column is strand (+ or -). All the other columns are irrelevant. NB: There’s inconsistency between what the chromsomes are called in this file and what they’re called in the reference genome!
4. S\_paradocus\_genes.bed. A bed file of *S. paradoxus* orthologs to the *S. cerevisiae* genes. All columns are the same as for the *S. cerevisiae* file. NB: The gene names in this file has trailing integers (e.g. the first entry is YHR055C.1). You DO NOT want the trailing integers in downstream analysis.
5. A bunch of files with the form <species>\_<data>\_<replicate>.fastq.gz. For example Scer\_RNA\_seq\_1.fastq.gz is the first biological replicate of *S. cerevisiae* RNA seq data.

You will have to write a short paper, with Intro, Methods, Results, Discussion, describing the analysis you undertake. Specifically, you will need to produce the following display items (all items should go in the main text unless otherwise specified):

1. A table of differentially expressed genes between the two species at an FDR of 10% (do not put this in the main text, leave it as a supplemental table).
2. A table showing genes with differential ribosome occupancy between the two species at an FDR of 10% (also not a main text table).
3. A table showing genes with differential translation efficiency between the two species at an FDR of 10% (also not main text).
4. Visualize the results of the previous analyses by making an MA plot for each, and presenting a 3-panel figure.
5. A scatterplot showing log fold change in mRNA abundance between the species vs. log fold change in translational efficiency.
6. A table showing the number of genes where there is compensatory evolution (break it down into mRNA up, TE down and vice versa) as well as the number of genes undergoing coordinated evolution (greak it down into both up, both down).

You will need to interpret these results in the context of what you think about regulatory evolution in the discussion. I’m not expecting a deep treatise, just try to contextualize things. Are you surprised by how many/how few differences you saw at the two modes of regulatory evolution? Are you surprised by the amount of compensatory or coordinated evolution that you see? Just speculate widly!

I’m leaving this somewhat open ended, because I want you to really try to put together a lot of the stuff you’ve learned in the class. Take your time with the project and be careful!

Some hints:

1. You should map all reads to the transcriptomes of the two species, not the reference genomes. That means you will have to use the files with the gene coordinates to extract the sequences of all the genes from the reference genomes, and put them into one big fasta file, where each entry in the file is one gene. BE CONSISTENT WITH THE NAMES OF THE GENES BETWEEN THE TWO SPECIES, YOU WILL REGRET IT IF YOU AREN’T.
2. Use Kallisto to quantify both the mRNAseq and the Ribosome Profiling data
3. Use DESeq2 for differential mRNA expression, just like you have in a problem set.
4. Use DESeq2 for differential ribosome occupancy, essentially treating the ribosome profiling data as if it were mRNA seq data
5. Remember that translational efficiency is the ribosome occupancy (counts from ribosome profiling experiment) normalized by the mRNA abundance (counts from mRNA experiment). This is a slightly more complicated beast than just the mRNA abundance and the ribosome occupancy on its own. Fortunately, DESeq2 can handle this. You should see this post by one of the authors of DESeq2 and see if you can figure out how to use this knowledge to test for differential translation efficiency: <https://support.bioconductor.org/p/61509/>
6. Subhint: in the post, you can think of “assay” as whether its’ the ribosome profiling data or the mRNA data, and “condition” as which species you are looking at.
7. Say a gene evolves under compensatory evolution if the translation efficiency and the mRNA abundance are significantly different and have opposite directions of effect.
8. When in doubt, look things up on google scholar! I’ve given you the reference to the paper we’re using data from, so that should be helpful. You also might be interested in Ingolia *et al* (2009), which is the first ribosome profiling paper.

References:

|  |
| --- |
| 1) Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M. and Hittinger, C.T., 2011. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the Saccharomyces sensu stricto genus. *G3: Genes, Genomes, Genetics*, *1*(1), pp.11-25.  2) McManus, C.J., May, G.E., Spealman, P. and Shteyman, A., 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome research*, *24*(3), pp.422-430.  3) Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S., 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, *324*(5924), pp.218-223. |
|  |