Github link

https://github.com/Intro-Sci-Comp-UIowa/biol-4386-course-project-tatagozli/tree/main Github link

Title

Erives AJ, Fassler JS (2015) Metabolic and Chaperone Gene Loss Marks the Origin of Animals: Evidence for Hsp104 and Hsp78 Chaperones Sharing Mitochondrial Enzymes as Clients. PLoS ONE 10(2): e0117192. https://doi.org/10.1371/journal.pone.0117192

Reference

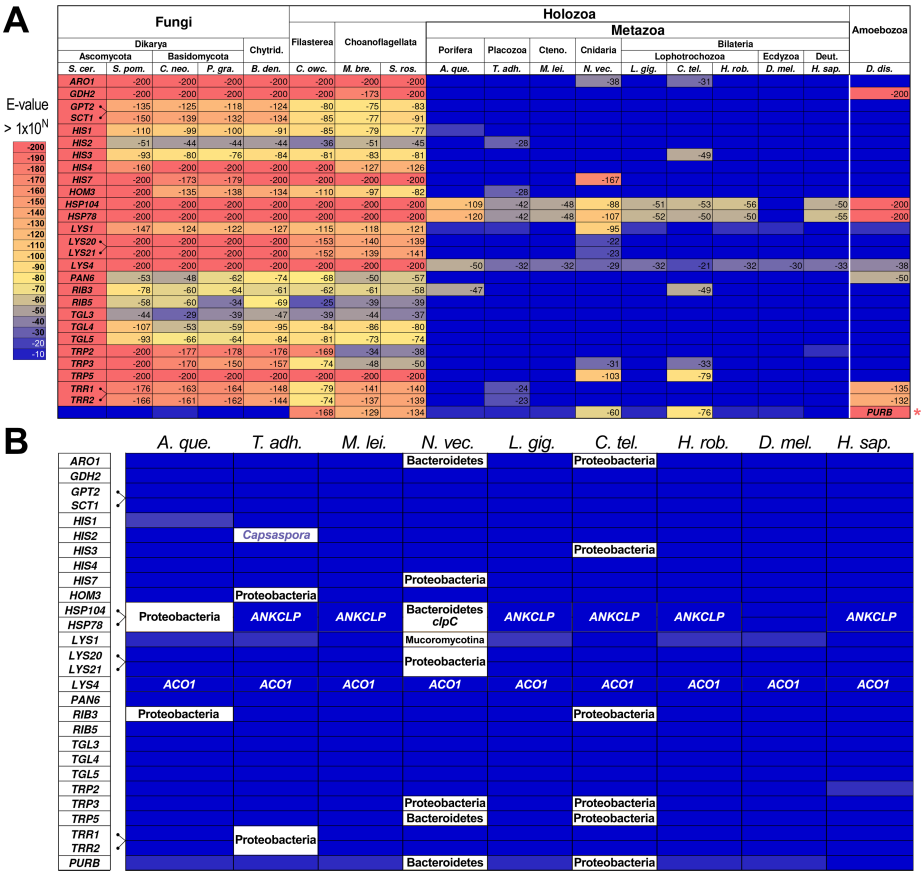
Link Link 2

Introduction and Abstract

The authors of this paper were interested in seeing which genes were lost in the evolution of animals. While the expansion of genes, particularly transcription factors, in animals has been well-studied, the loss of genes from single-celled eukaryotes to animals has been poorly studied. Understanding which genes were lost is important because it may lend insight into the context, origin, and constraints associated with new transcription programs in animals.

They predicted 27 genes were maintained in non-animal eukaryotes but lost in animals. To test this hypothesis, they used BLASTP to compare the fungal (S. cerevisiae) homolog of those genes with 17 other organisms from fungi and holozoa. Their results confirmed their hypothesis. However, they only looked at 7 non-animal genomes and 10 animal genomes. Since that time, many more genomes have been sequenced. This project expands on their finding by using BLASTP to compare the S. cerevisiae genes to the 64 genomes in Paps et al., 2018.

Original Figure



~Figure 1

Figure 1 Legend

(A) Heat map of twenty-seven genes (first twenty-seven rows with gene names) from the budding yeast S. cerevisiae (S. cer.) and their orthologs in related organisms with whole genome sequence assemblies and their closest matches in animals. Six of the twenty-seven yeast genes are the result of lineage-specific duplications in yeast (indicated by dark gray brackets, left of table), leaving twenty-four genetic functions either lost or not measurably conserved in metazoans. The heat map is based on the BLASTP expect values of the best match to the indicated yeast gene (see key, lower right of table). Species abbreviations: S. cer., Saccharomyces cerevisiae; S. pom., Schizosaccharomyces pombe; C. neo., Cryptococcus neoformans; P. gra., Puccinia graminis; B. den., Batrachochytrium dendrobatidis; C. owc., Capsaspora owczarzaki; M. bre., Monosiga brevicollis; S. ros., Salpingoeca rosetta; A. que., Amphimedon queenslandica; T. adh., Trichoplax

adhaerens; M. lei., Mnemiopsis leidyi; N. vec., Nematostella vectensis; L. gig., Lottia gigantea; C. tel., Capitella teleta; H. rob., Helobdella robusta; D. mel., Drosophila melanogaster; and H. sap., Homo sapiens. Other abbreviations: Chytridio., Chytridiomycota; Placo., Placozoa; Ctenoph., Ctenophora; Ecdyso., Ecdysozoa; and Chor., Chordata. The last row is from an alternative screen for lost animal genes (see text) in which Amoebozoa PFAM domains that are absent in animals were identified (five genes already identified plus one new gene, PURB). (B) Close-up of the lost animal gene columns from panel A. This table shows either the taxonomic origin of a weak animal match, or the name of the gene if it is not directly related (i.e., not orthologous) to the named yeast gene (first column) for all matches with BLASTP expectation values < 1e-20. Most of the animal matches to the candidate lost genes correspond to lineage-specific horizontal gene transfers and/or environmental contaminants from unrelated (non-opisthokont) clades. Some matches correspond to non-orthologous genes such as ANKCLP (S2 Fig.) in the case of the eukaryotic clpB genes HSP78 and HSP104, or ACO1 in the case of LYS4 (Fig. 3). Only HIS2 might have been lost soon after early animal radiation based on the presence of a weak match in Trichoplax adhaerens. This gene is most similar to the version from Capsaspora and not to either of the two choanoflagellates, so it is of uncertain origin.

My description of the figure

Figure 1 of this paper uses computational methods to identify genes and genetic functions retained in fungi and choanoflagellates but lost in animals. The left column is the list of genes predicted by the Ensembl pipeline to be conserved in fungi and choanoflagellates, but absent in animals. The boxes are the log of the e-values. The red boxes indicate a close BLASTP match with the yeast gene, and the blue boxes represent a weak to no match. This figure clearly shows that these genes are conserved in fungi and choanoflagellates and absent in animals (metazoans). There is a weak match in some metazoans with Hsp104 and Hsp78, but this is a nonfunctional variant of the gene.

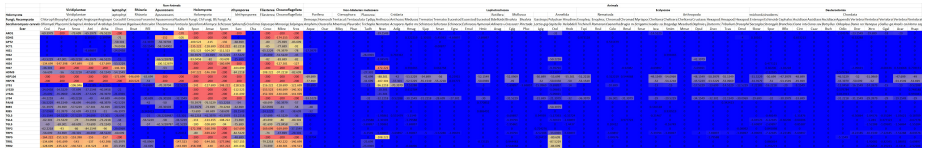
My Project

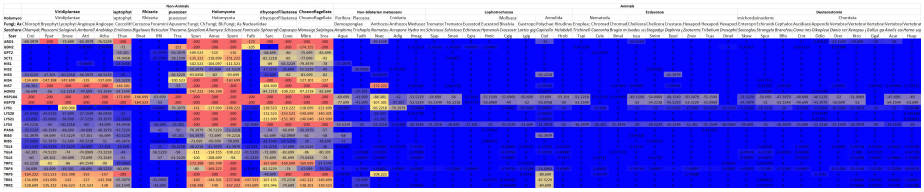
This is a powerful figure, but lots of other genomes have been sequenced since this figure was made by Erives and Fassler. These genomes can be found in Paps et al., 2018. My objective is to redo their BLASTP analysis with more genomes, including more non-animal eukaryotes (besides fungi and choanoflagellates), and more animal genomes.

Methods

1. Obtain the protein sequences for the 27 yeast genes in FASTA format from the supplemental figures of Erives and Fassler, 2015.
2. Obtain the list of sequenced genomes from Paps et al., 2018
3. Use the RefSeq_protein database in NCBI BLASTP to check for alignments between the 27 candidate genes and the sequenced genomes. For genomes not available in the RefSeq database, use the non-redundant database
4. Put the E-values in Microsoft excel, and obtain the log of the e-values. Enter an e-value of 1 for no matches (log of 1 is 0), and a manually enter a value of -200 for matches with an e-value of 0
5. To make the heat map use the conditional formatting feature in Excel. Set -200 as red, -100 as yellow, and the highest value as blue.
6. For the trimmed figure, remove genomes (columns) from the heat map with no BLASTP matches.
7. Copy and paste the heat map cells into Microsoft paint
8. Save the file as a JPEG

My Figure

~My Figure 

~Trimmed Version 

Results and Conclusions

My figure agrees with the original conclusion by the authors that these 27 yeast genes are mostly lost in animals but retained in fungi and choanoflagellates. However, there are a few differences between my figure and the original one: 1.) While the 27 genes are found in fungi and choanoflagellates, they are not present in plants or rhizaria. 2.) Hsp104 and Hsp78 is conserved in animals as Clpb, a nonfunctional variant. However, some genomes, including the widely used model organism, *C. elegans*, do not even have this nonfunctional variant 3.) There are individual animal genomes in the figure which seem to have somewhat strong matches to some of the genes

Discussion

There's a number of hypotheses I have for the novel conclusions in my new figure which would require further research. 1. These genes are missing from plants and rhizaria because 1.) They are not compatible with these organisms, 2.) These genes emerged after plants and rhizaria had diverged from the eukaryotic tree. To test these hypotheses we could look at the phylogeny of these genes to see when they appeared or do transform these genes into plants and see what happens. 2. We know that Hsp104 is toxic to developing animals, including *C. elegans* and *X.laevis* (Skuodas and Clemons, 2020), presumably because Hsp104 interferes with native aggregates. I would be interested in introducing ClpB into *C. elegans* to see if it is also toxic. 3. Some exceptions could result from horizontal gene transfer.To test this I would want to know if the match resembles a bacterial variant of the protein

Reflection

As someone new to computer science, I had a lot of hurdles during this work, but I was able to overcome them and produce an acceptable figure while learning a lot in the course. My original goal in using this figure was to use the ensembl pipeline like the authors did to see if I would have a different list of genes than the 27 they predicted. However, I could not figure out what they did from what was written in the paper, and more specifically I could not figure out how to use ensembl to compare multiple genomes are once. Thus, I was stuck on this step for a long time. Eventually I decided to start with the 27 genes they had already predicted, and to build on their original figure by using an expanded list of genomes. At this point, my progress on the project was more smooth. With the BLAST analysis, I made sure to read up on E-values and available database (nonredundant vs refseq), but for the most part I didn't have any problems. I figured out how to make the heat map in excel by using google and youtube. If I was repeating this project, and I had more time, I would consider using HMMR in addition to BLASTP and comparing the results. I would also have tried to use R to make the heat map. Lastly, I would have asked for help sooner instead of remaining stuck on the same step for such a long time.

Overall, I am pleased with my final product, and I look forward to discussing the figure with my P.I. to get her thoughts on it. Regarding the course, I got a lot out of it. That being said, I am far from skilled in R, Python, or Unix. However, I have an introductory understanding of all these systems and feel I could learn them more easily if I needed to in the future. I did see a lot of personal improvement in this course. Most notably, I am comfortable with Unix and Github. For example, I can make directories, files, and write stuff in markdown in my fastx terminal and push any changes to github relatively easily. I was not able to do these things before I took the course. I think these basic skills will be helpful if I take future courses or workshops in computer science.