Nate Evans

Mike Mooney

BMI665

12/3/18

Bioinformatics Scripting: Research Project

There are 4 main components (\*) to this project and must be in the following file structure:

evans\_project-part1.py\*

evans\_project-part2.py\*

evans\_lib.py\*

evans\_proj\_run.sh\*

/data/

/BMI565\_ResearchProject\_Data/

H5N1\_VN1203\_DE\_Probes.txt

H5N1\_VN1203\_UNIVERSE\_Probes.txt

KEGG\_Pathway\_Genes.txt

/outputs/

empty

/temp/

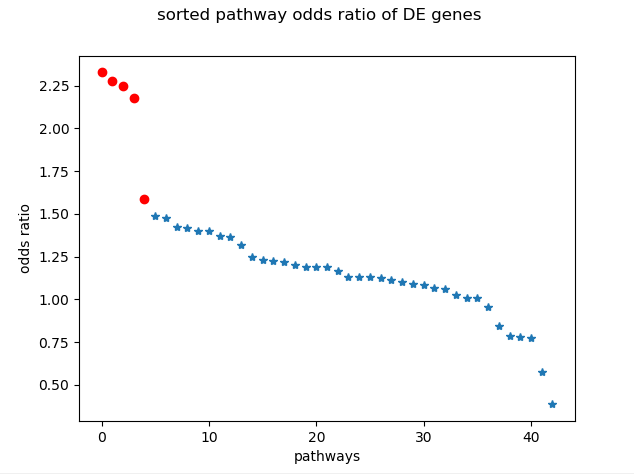
empty

To run this script, navigate to the appropriate base folder and run:

$ ./evans\_proj\_run.sh

# Results

For part 1, it was fairly straight forward to run an odds ratio comparing the number of differentially expressed genes that were in each pathway and after sorting the pathways by highest DE odds ratio to lowest, the following plot was produced. From this, we can see that there were 5 pathways with an odds ratio above 1.5 and of those I chose Small cell lung cancer as it had the highest odds ratio.



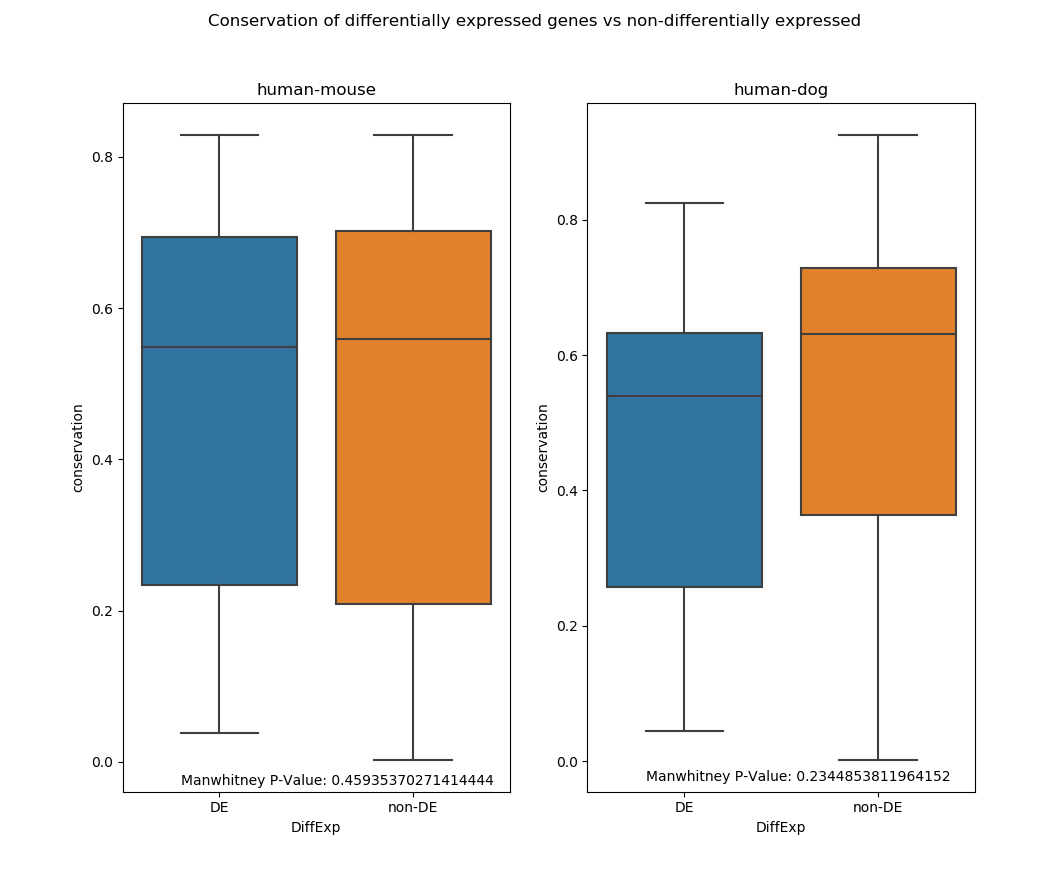
In order of DE odds ratio:

1. Small cell lung cancer
2. RIG-I-like receptor signaling pathway
3. ECM-receptor interaction
4. Toll-like receptor signaling pathway
5. Focal adhesion

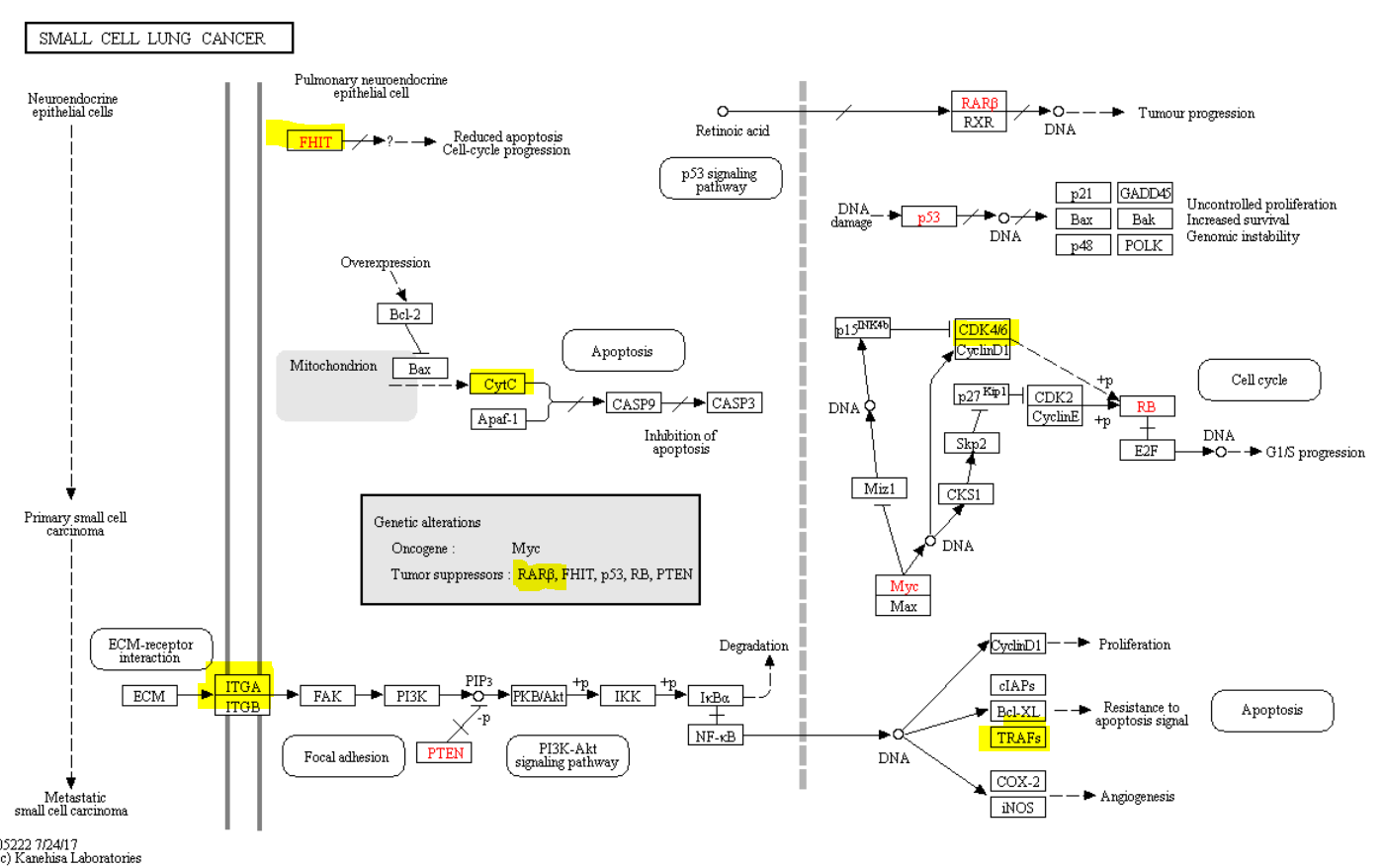
In part 2, I systematically pulled all the gene sequences for each gene in the pathway for three different species from the entrez ‘nuccore’ database:

1. homo sapiens
2. mus musculus
3. canis lupus familiaris

These sequences were then concatenated into a fasta file, written to disk and then aligned by clustalW. The resulting alignment file was read back into memory, and the edit distance was calculating, normalizing by gene length. The gene edit distances for each species were compared between differentially expressed genes and non-differentially expressed genes, resulting in the plot below. From these results, and a manwhitney test, there are no statistically significant differences between groups.



Below is a graph of the DE genes within the pathway map itself. I couldn’t locate all the genes and I didn’t find any overarching trends within it.



# Problems

There were two serious issues encountered during this project, first was the time required to both retrieve gene sequences from entrez and then to align those sequences using clustalW. This bottleneck significantly throttled development and in the future, I would work to optimize this either algorithmically or by upgrading the hardware.

Second, the entrez search didn’t always return a gene for each species, indicating that many genes were not homologous OR my search terms weren’t effective enough. For this reason, its important to not that there were 14/97 genes that failed to align and were therefore not included in the final analysis.