Final Research Project
BMI 565/665 Python Scripting

Eric Leung

December 8th, 2014

Part I: Identify Pathway Focus

We were given three files to analyze.

- `H5N1_VN1203_DE_Probes.txt`
- `H5N1_VN1203_UNIVERSE_Probes.txt`
- `KEGG_Pathway_Genes.txt`

Upon analysis, I found that there were six pathways that have a larger number of differentially expressed genes than would be expected by chance (which was measured by an odds ratio greater than 1.5).

Pathway	Odds Ratio
ECM-receptor interaction	2.2812
RIG-I-like receptor signaling pathway	**2.2828**
Focal adhesion	1.5999
Cytokine-cytokine receptor interaction	1.5165
Small cell lung cancer	2.3641
Toll-like receptor signaling pathway	2.1792

Table 1: Significant pathways with odds ratios greater than 1.5

The pathway I chose is "RIG-I-like receptor pathway." It has an odds ratio of 2.2828, which is greater than 1.5, which we deem significant enough to study. Figure 1 can be found at `http://www.kegg.jp/kegg-bin/show_pathway?hsa04622`.

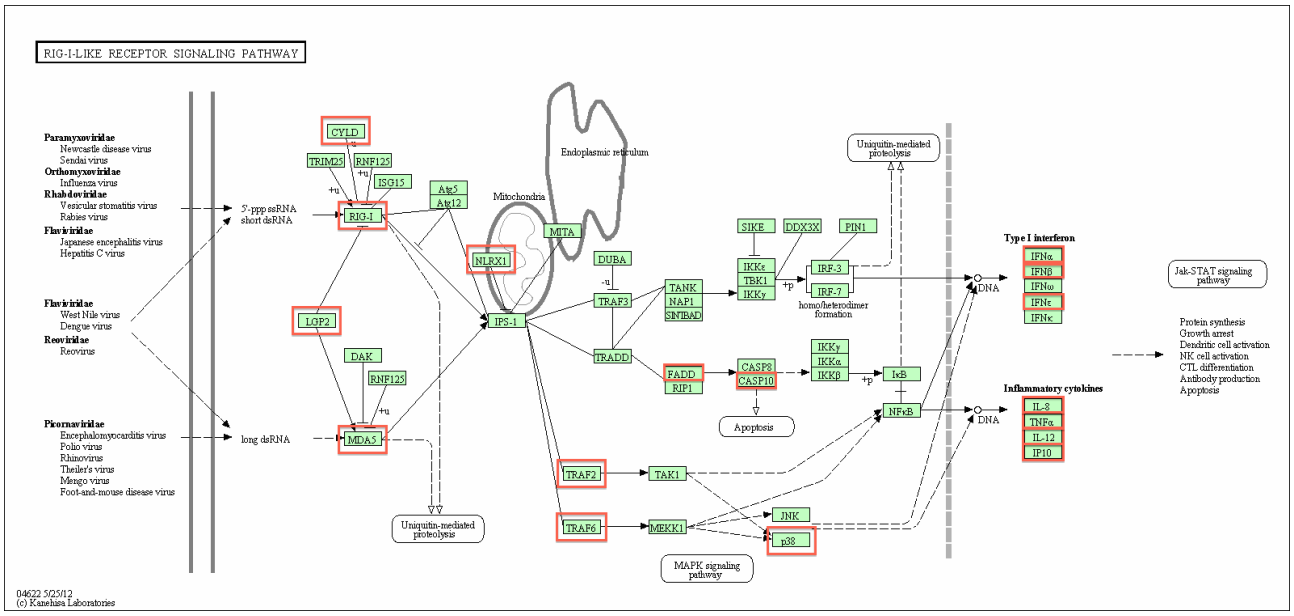

Figure 1: RIG-I-like receptor signaling pathway for Homo sapiens with differentially expressed genes highlighted.

Part II: Examine Cross-Species Conservation of Pathway Genes

I studied the conservation of genes among Humans (*Homo sapiens*), Mice (*Mus musculus*), and Chimpanzees (*Pan troglodytes*).

Instead of taking the "RIG-I-like receptor pathway" genes listed in the "KEGG Pathway Genes" text file, I decided to go back to the KEGG website [1] to find the genes in the pathway for each species that I focused on. I relied on the KEGG API to search for and parse the right page to extract the list of genes.

I decided to disregard the list that was in the file given because I found that some of the "genes" listed for my pathway did not correspond to any human gene. Two were cDNA clones (AW511634 and AI423557), a partial CDS for a transcription factor (U88316), and another was from HeLa cells (U56433). Thus, I decided that this was not a list of human genes and went straight to the source (KEGG) for the list of genes to study.

Additionally, I came across a duplicate gene, but under difference names. There is a gene CR936771 in my list in the KEGG pathways file that when I look up, gives me SIKE1, which is already in my list, but it is listed as SIKE, but SIKE1 is the official gene name, according to HGNC.

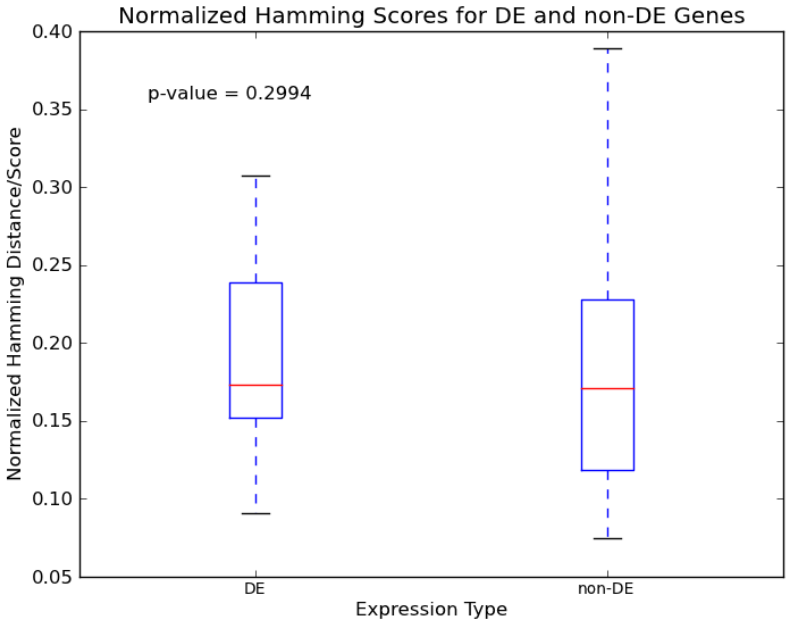

Figure 2: Boxplot of normalized Hamming distances for differentially expressed and non-differentially expressed genes in the RIG-I-like receptor pathway for Humans (*Homo sapiens*), Mice (*Mus musculus*), and Chimpanzees (*Pan troglodytes*).

The decision to go to the KEGG database for the list of genes to study resulted in eighteen genes from the "KEGG Pathway Genes" file not being found for my pathway. As it was mentioned earlier, I found that I wouldn't have been able to find legitimate genes for five "genes". Thus, I did not search for thirteen genes. This may have had an

[1] `http://www.kegg.jp/dbget-bin/www_bget?hsa04622`

effect on the final analysis. However, as you will see in the end, these thirteen missing genes must have had radical differences in them to change the overall statistical test.

I wanted to only compared genes that existed among all three species. This makes analysis easier instead of having to make a separate analysis for a pair of genes and all three genes.

We found there was no significant difference between differentially expressed genes and non-differentially expressed genes in terms of Hamming distance among three different species (p-value $= 0.2994 > 0.05$).

H5N1 Infection