Andrew Judell-Halfpenny
BIOL607: Introduction to Biological Data Analysis

Final Project Proposal

I plan to combine RNA-seq data produced by the Center for Personalized Cancer Therapy (CPCT) at Umass Boston from Benign Prostatic Hyperplasia (BPH)/Lower Urinary Tract Syndrome (LUTS) symptomatic patients with public data-sets from the GEO-profiles repositories to produce a large enough sample size for analysis and identification of gene expression biomarkers of BPH/LUTS. The RNA-Seq data produced by the CPCT consists of 9 symptomatic and 9 asymptomatic males. The CPCT's BPH/LUTS data was initially processed by Tailor, an RNA-seq analysis pipeline that Professor Riley and I built to perform analysis and visualization of differential gene expression between two or more groups. Unfortunately, our current signal to noise ratio is quite low because of the (still) prohibitively high expense of RNA-sequencing and its subsequently small sample sizes. In the CPCT's BPH/LUTS data-set, intra-group variability is biased by several patients with incredibly high expression levels which causes our current hypothesis testing protocol (Cuffdiff) to produce significance calls that fail to elucidate the etiology of the disease. I will seek (have sought) out data-sets with similar experimental protocols to merge into the CPCT data-set to increase the power of my analysis. I will describe the effect size and power of the new data set before identifying the most likely distribution of gene expression for each gene within each group using maximum likelihood. I will then identify BPH/LUTS biomarkers by using a chi square or Kolmogorov-Smirnov test with some type of multiple hypothesis correction. To conclude the analysis, I will compare the output of this analysis to several commonly used alternate methods such as Limma, DESeq2, Rsubread, and EasyRNASeq. I may incorporate some classification techniques which are outside the bounds of this course such as partitioning the data-set into training and test sets to determine the accuracy of this model as a protocol for BPH/LUTS testing as well as a method to identify gene expression biomarkers of under-researched diseases. I hope that this analysis refines the final leg of Tailor and can be combined into the method for inclusion in a scientific paper.