Khan Inan
Transcriptomics project proposal

American and European eel microarray

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL19017

This particular experiment stood out to me because it focuses on pollution/contamination and how it has affected the transcriptome of eels. In a world where many water-sites are being polluted, it is important to understand how certain species can survive and what makes them resilient. North American and European eels in particular can survive and even thrive in low oxygen conditions and a variety of salinity levels, so it would be interesting to see how their population has been affected by pollutants in the water.

After analyzing the data I'm hoping to discover the extent to which pollutants have affected gene expression levels for certain parts of the eels transcriptome. This would be valuable information because if we can understand the role that these affected gene's play, then it would help us understand the ways that eels have adapted to these pollution concentration levels.

There are 8 factors in this experiment because they collected data from 8 separate sites. Of these site 4 were Canadian and 4 were french. Additionally, in terms of levels for each factor, there were 6 fish (eels) collected from each site. The data that was collected in addition to the actual specimens themselves, included water contamination levels (metals and organic pollutants) as well as oxygen level, salinity and even temperature.

I would say there are 5 other biological replicates for each samples. If location isn't a concern and the data for each eel only need to be compared with others if its own species, there should be 23 other replicated for each sample (4 locations x 6 fish for each location)

My goal for analyzing this data set is that I essentially want to understand how the pollution levels for the various locations affects the gene expression levels for certain parts of the sequenced transcriptome. Since this experiment is using microarrays, there are certainly some methods we can use to analyze the data. One method is power analysis, which involves a lot of hypothesis testing to determine if there is an effect or correlation between location/pollution and gene expression levels. More specifically i can use the standard t-test method or other t-test methods for calculating the difference between expressed genes. If I want a more visually intuitive method. I can go for a clustering approach, such as hierarchical clustering or k-means clustering. My general approach will be to try and identify how genes are differentially expressed in varying conditions and pollution levels.

Useful resource: https://ologyjournals.com/beij/beij_00001.php