Project Report – Claire Green

Overall aim: to discover a set of differentially expressed pathways that represent a functional signature for diseases expressing TDP-43 pathology.

Current aim: to optimise the quality of my data, explore enrichment within each individual data set, and compare multiple methods of generating differentially expressed functional gene sets.

Data

I have 5 data sets:

1. Motor neurons from 8 patients with C9orf72 mutations (ALS), and 3 controls
2. Motor neurons from 3 patients with CHMP2B mutations (ALS), and 7 controls
3. Motor neurons from 7 patients with sALS, and 3 controls
4. Cortical neurons from 6 patients with progranulin mutations (FTLD), 10 with sporadic FTLD, and 8 controls
5. Muscle cells from 7 patients with VCP myopathy, and 3 controls

Step One: Quality Control (*completed*)

To increase our confidence in the list that we have generated, the first step was to conduct quality control. This process indicates if there are outliers in the data, or samples with a particularly large distribution. The three methods used were:

1. Principle Components Analysis (2D and 3D)
2. Boxplots
3. WGCNA Cluster Analysis

From this I identified two possible outliers: one patient in the C9orf72 data set, and one control in the CHMP2B data set. Overall, quality of data sets was good.

Step Two: Enrichment Analysis

Enrichment can be investigated using multiple methods. By processing my data via multiple methodologies, it will help to show which gene sets/pathways are commonly found to be differentially expressed.

Part One: GSEA (*in progress*)

Gene set enrichment analysis of each data set will provide a list of gene sets that are differentially expressed. I can use this information in three ways;

1. I can run further QC analysis by comparing results with and without suspected outliers to see if they have an effect on the gene set p values.
2. I can look more closely at the three motor neuron data sets and see how similar/different they are from one another in terms of DE gene sets.
3. I can run similar analyses to the Pathprint method above where I find which gene sets are common across all 5 data sets.

Part Two: Other Enrichment Analysis Methods

Once I am confident that I have discovered what I can using GSEA, I would like to then turn to other suggested analyses including Global Ancova and G:Profiler, so as to see the difference between the available methods.

The Next Step…

Jobs for the next week+ include:

* Finish conducting GSEA on all data sets, and establish whether outliers should be removed by looking for changes in p values
* Construct networks in Cytoscape using the GSEA results from each data set to visualise relationships of enriched pathways
* Create presentation showing results so far
* Pathprint each individual patient and look for clusters within each sample