Lab Book 13/11/15

Claire Green

Monday

1. I began by creating pathway lists for each combination of data sets, including each individual tissue type and pairwise combinations. The top 20 differentially expressed pathways of each condition is as follows:

Motor Neurons .
Complement Activation, Classical Pathway (Wikipathways)
Complement and Coagulation Cascades (Wikipathways)
{CALM1,30} (Static Module)
Phototransduction (KEGG)
{RAN,17} (Static Module)
{SREBF1,11} (Static Module)
Glycosphingolipid biosynthesis - lacto and neolacto series (KEGG)
Glyoxylate and dicarboxylate metabolism (KEGG)
Butanoate metabolism (KEGG)
{RARA,17} (Static Module)
Starch and sucrose metabolism (KEGG)
Drug metabolism - other enzymes (KEGG)
Phase I, non P450 (Wikipathways)
{CDK5,14} (Static Module)
Acetylcholine Synthesis (Wikipathways)
{F2,46} (Static Module)
Ubiquinone and other terpenoid-quinone biosynthesis (KEGG)
Tryptophan metabolism (KEGG)
Complement and coagulation cascades (KEGG)
Leukocyte transendothelial migration (KEGG)

Cortex .
{BCLAF1,25} (Static Module)
Benzo(a)pyrene metabolism (Wikipathways)
Muscle contraction (Reactome)
{CNOT2,13} (Static Module)
{HSP90AA1,18} (Static Module)
{C5,29} (Static Module)
ABC transporters (KEGG)
Heart Development (Wikipathways)
Codeine and morphine metabolism (Wikipathways)
Mitochondrial LC-Fatty Acid Beta-Oxidation (Wikipathways)
Monoamine GPCRs (Wikipathways)
{SIX3,11} (Static Module)
Oxidative phosphorylation (KEGG)
Alzheimer's disease (KEGG)
Type II diabetes mellitus (Wikipathways)
Nicotine Activity on Dopaminergic Neurons (Wikipathways)
NLR proteins (Wikipathways)
Osteoblast Signaling (Wikipathways)

Fatty Acid Biosynthesis (Wikipathways)
TCA Cycle (Wikipathways)

Muscle .
{CNOT2,13} (Static Module)
Glycosaminoglycan biosynthesis - heparan sulfate (KEGG)
{ACY1,11} (Static Module)
{CHRNA1,13} (Static Module)
Taurine and hypotaurine metabolism (KEGG)
Long-term potentiation (KEGG)
Nicotine Activity on Dopaminergic Neurons (Wikipathways)
Irinotecan Pathway (Wikipathways)
{VCP,17} (Static Module)
Primary bile acid biosynthesis (KEGG)
Purine metabolism (KEGG)
{EPRS,15} (Static Module)
mRNA surveillance pathway (KEGG)
Bladder cancer (KEGG)
Non-small cell lung cancer (KEGG)
Vitamin B12 Metabolism (Wikipathways)
Estrogen signaling pathway (Wikipathways)
{POLR2H,109} (Static Module)
Neurotrophin signaling pathway (KEGG)
Influenza A (KEGG)

MN + Cortex .
Complement Activation, Classical Pathway (Wikipathways)
Complement and Coagulation Cascades (Wikipathways)
Phototransduction (KEGG)
{RAN,17} (Static Module)
{SREBF1,11} (Static Module)
Glycosphingolipid biosynthesis - lacto and neolacto series (KEGG)
Glyoxylate and dicarboxylate metabolism (KEGG)
Butanoate metabolism (KEGG)
Starch and sucrose metabolism (KEGG)
Phase I, non P450 (Wikipathways)
{F2,46} (Static Module)
Complement and coagulation cascades (KEGG)
Prion diseases (KEGG)
Statin Pathway (Wikipathways)
Signaling by Insulin receptor (Reactome)
Collecting duct acid secretion (KEGG)
{ESR1,24} (Static Module)
Dorso-ventral axis formation (KEGG)
{C5,29} (Static Module)
Nucleotide GPCRs (Wikipathways)

MN + Muscle .
Complement Activation, Classical Pathway (Wikipathways)
Complement and Coagulation Cascades (Wikipathways)
Phototransduction (KEGG)
{RAN,17} (Static Module)

{SREBF1,11} (Static Module)
Starch and sucrose metabolism (KEGG)
Phase I, non P450 (Wikipathways)
{CDK5,14} (Static Module)
{F2,46} (Static Module)
Tryptophan metabolism (KEGG)
Complement and coagulation cascades (KEGG)
Prion diseases (KEGG)
Signaling by Insulin receptor (Reactome)
Collecting duct acid secretion (KEGG)
{ESR1,24} (Static Module)
Pantothenate and CoA biosynthesis (KEGG)
Dorso-ventral axis formation (KEGG)
Tamoxifen metabolism (Wikipathways)
Vitamin B12 Metabolism (Wikipathways)
{VCP,17} (Static Module)

Cortex + Muscle .
Benzo(a)pyrene metabolism (Wikipathways)
{CNOT2,13} (Static Module)
{HSP90AA1,18} (Static Module)
ABC transporters (KEGG)
Monoamine GPCRs (Wikipathways)
Type II diabetes mellitus (Wikipathways)
Nicotine Activity on Dopaminergic Neurons (Wikipathways)
NLR proteins (Wikipathways)
Fatty Acid Biosynthesis (Wikipathways)
Irinotecan Pathway (Wikipathways)
{KCNAB1,18} (Static Module)
Neuroactive ligand-receptor interaction (KEGG)
Complement and coagulation cascades (KEGG)
Phase I, non P450 (Wikipathways)
Complement Activation, Classical Pathway (Wikipathways)
Biogenic Amine Synthesis (Wikipathways)
{RAD21,11} (Static Module)
Nitrogen metabolism (KEGG)
Phototransduction (KEGG)
Collecting duct acid secretion (KEGG)

All .
Complement Activation, Classical Pathway (Wikipathways)
Complement and Coagulation Cascades (Wikipathways)
Phototransduction (KEGG)
{RAN,17} (Static Module)
{SREBF1,11} (Static Module)
Starch and sucrose metabolism (KEGG)
Phase I, non P450 (Wikipathways)
{F2,46} (Static Module)
Complement and coagulation cascades (KEGG)
Prion diseases (KEGG)
Signaling by Insulin receptor (Reactome)
Collecting duct acid secretion (KEGG)
{ESR1,24} (Static Module)

Dorso-ventral axis formation (KEGG)
Type II diabetes mellitus (KEGG)
Urea cycle and metabolism of amino groups (Wikipathways)
{HSPA8,34} (Static Module)
Fructose and mannose metabolism (KEGG)
Ganglio Sphingolipid Metabolism (Wikipathways)
Jak-STAT signaling pathway (KEGG)
Fat digestion and absorption (KEGG)
Biogenic Amine Synthesis (Wikipathways)
ABC transporters (KEGG)
Pentose and glucuronate interconversions (KEGG)
Nitrogen metabolism (KEGG)

2. I began trying to get to grips with GSEA using the JAVA platform but I will do it through R instead because it is becoming difficult to use it with Cytoscape and knowing the R route will be better for me.

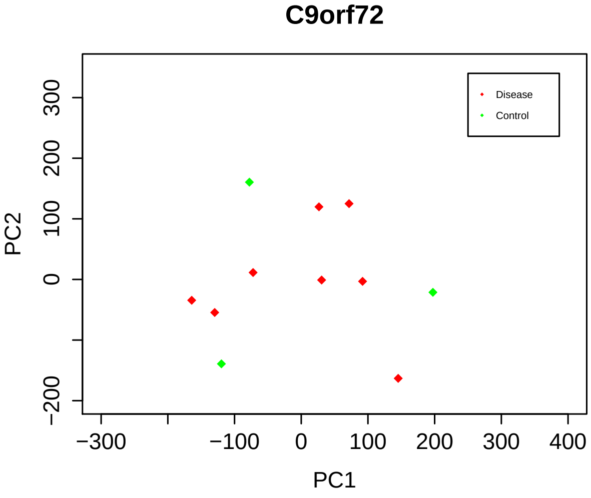
Tuesday

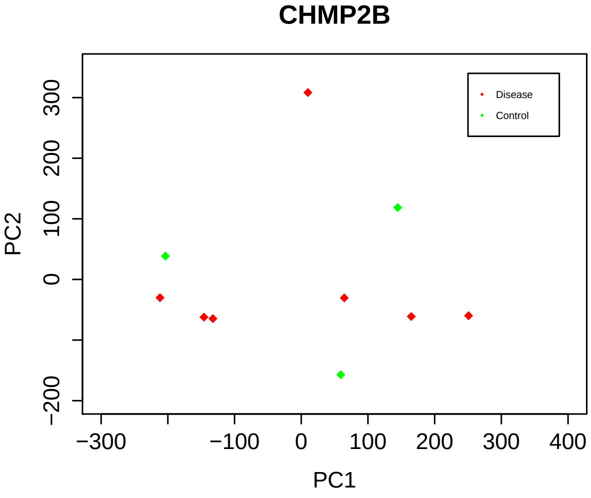
After talking with John, the aim of today was to visualise the data and see if there is variance/outliers etc. First I did a principle component analysis of each data set. This type of analysis takes the data and presents it in a way that shows the most variation.

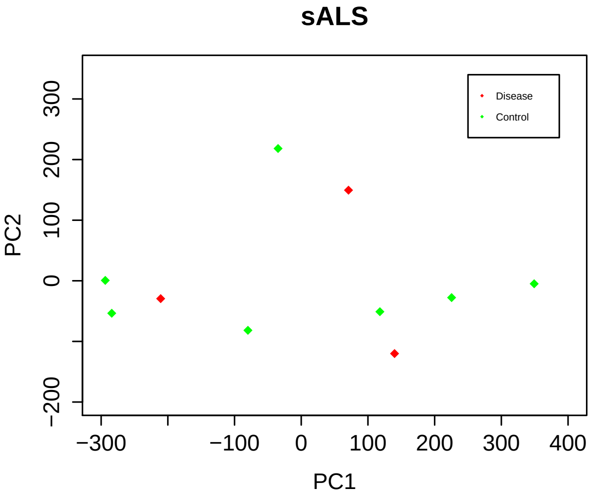
1. I started with visualising the data in 2D. To do this I used the function "prcomp" on each data set, and visualised the controls and disease as separate colours. An example of the code can be seen below

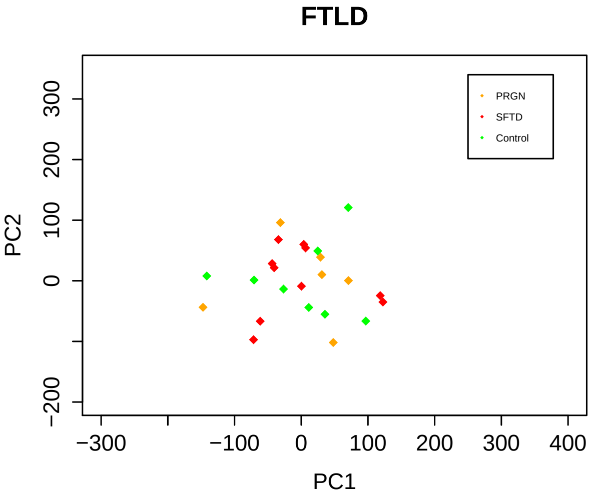
```
### C9orf72 ###
dev.off() #removes previous plot
C9d <- exp_C9.LCM[,c(1:8)] #take expression data from disease columns
C9c <- exp_C9.LCM[,c(9:11)] #take expression data from control columns
pcaC9d <- prcomp(t(C9d)) #run pca
pcaC9c <- prcomp(t(C9c))

plot(pcaC9d$x[,1:2], pch=18, cex=1.25, col="red", xlim=c(-300, 400), ylim=c(-200, 350),main = "C9orf72"
points(pcaC9c$x[,1:2], pch=18, cex=1.25, col="green", xlim=c(-300, 400), ylim=c(-200, 350))
legend(260, 340,pch=18, legend=c("Disease", "Control"), col=c("red","green"), cex=0.7)
```

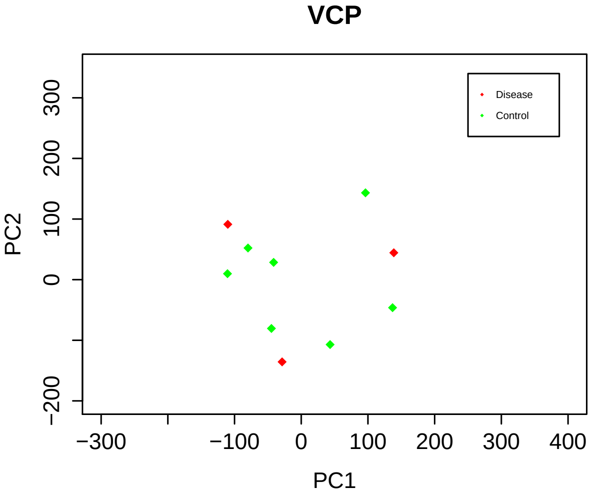
This process was repeated for each data set.











Results

As you can see there appears to be a bit of an outliar in the CHMP2B data, but until I have conducted another analysis I can't be sure that it is a problem. I will just bare it in mind for now. Other than a bit of a spread of data in the sALS set, all other sets look okay.

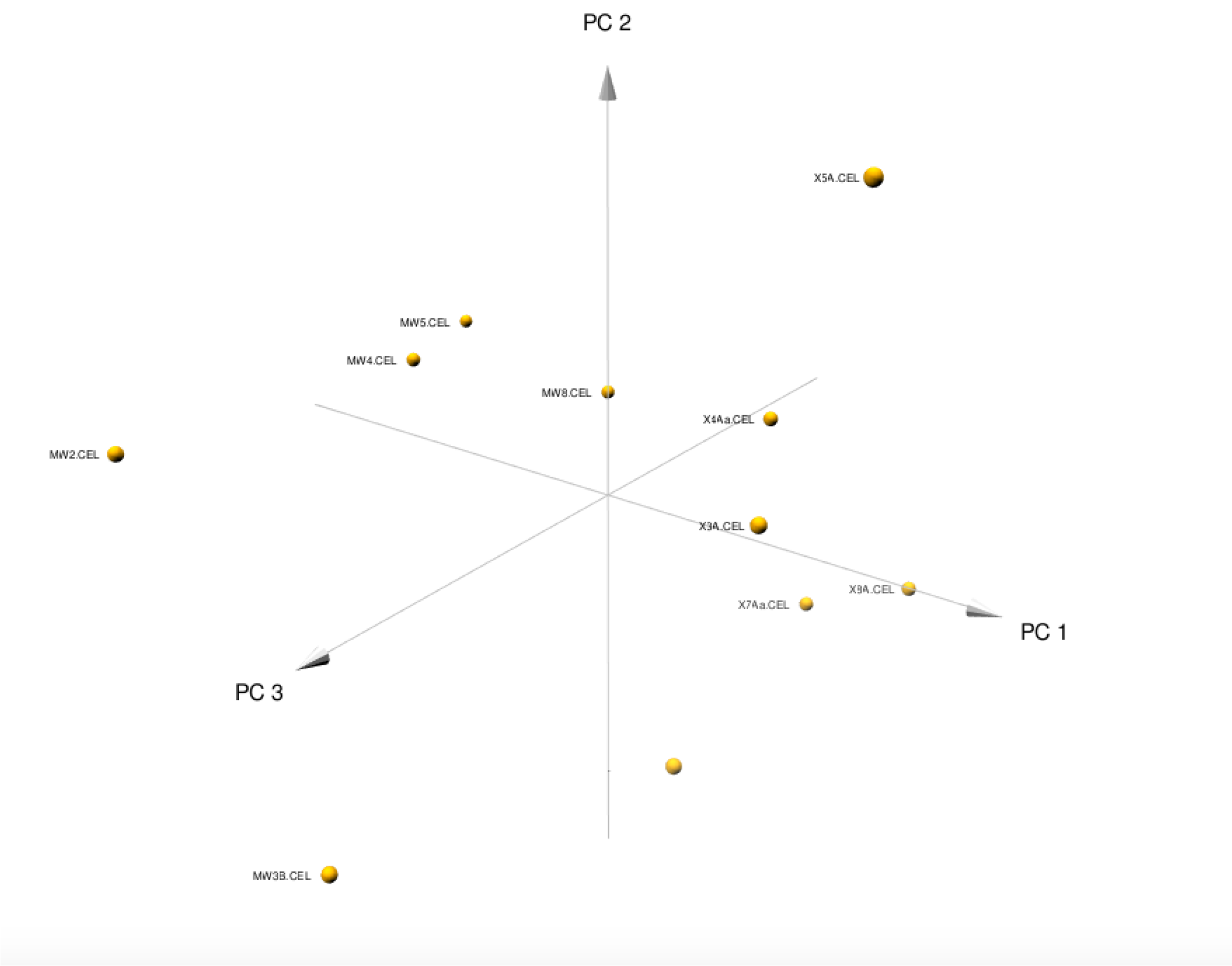
2. I began attempting to create a 3d PCA plot for an added dimension of variance, but the library that was recommended called “rgl” was complicated to install so I left it for another day.

Wednesday

1. I was able to get the rgl package to work, allowing me to visualise 3d PCA plots. The problem I was having was that the X11 graphics system is not installed on Mac so I had to install XQuartz to install X11 and that meant I could load the rgl library. The output doesn’t travel so well to a document such as this, but the code is as follows:

```
## 3D Principle Component Analysis ##

setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets")
load("TDP-43 Analysis Environment.RData")


####DISCLAIMER
#To use rgl you have to have X11 installed. If you get the error that X11 is not found (usually
#mac), then go to XQuartx and download it. After that, load the rgl source("y") again.

library(rgl)
library(pca3d)

###C9orf72###
pcaC9 <- prcomp(t(exp_C9.LCM)) #conduct PCA
pca3d(pcaC9$x[,1:3])
text3d(pcaC9$x[,1:3], text=rownames(pcaC9$x), adj=1.3, color="black", cex = 0.5) #add labels

And repeat for the other 4...
```

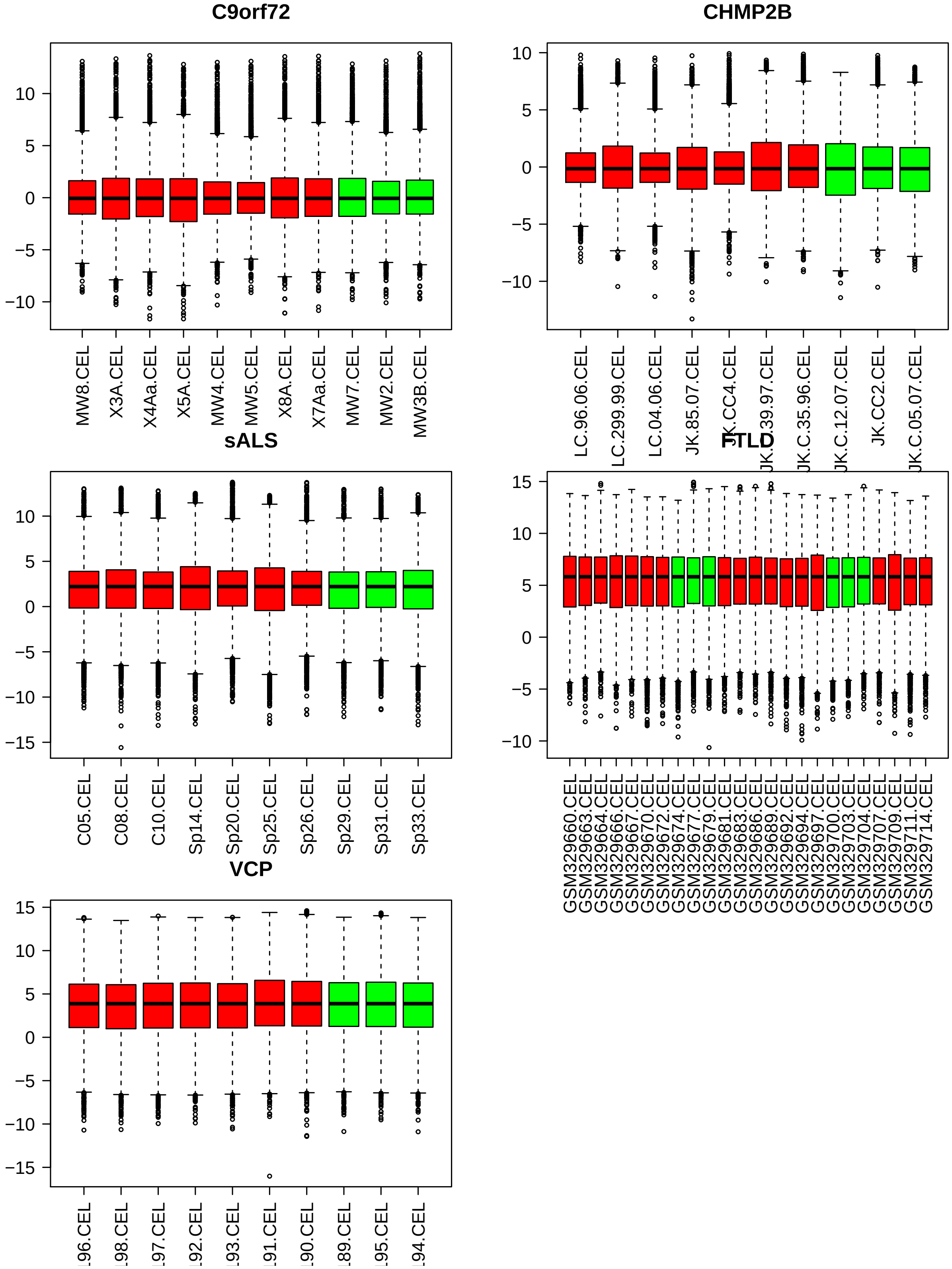
Results An example of the XQuartx output can be seen below. This is interactive and allows for exploration of the 3D space. The outliar of the CHMP2B was very clearly different than all other samples on PC2. It again supports that it could be a problem. The variance in other data sets was more obvious in 3D, however John said not to worry too much, as I guess adding the 3rd dimension can make it look worse than it is.



2. Next, I began further analysis of variance by creating boxplots for each of the data sets. The code is as follows:

```
####BOXPLOT ANALYSIS OF MICROARRAY DATA####

dev.off()
par(mfrow=c(3,2))

#C9orf72
boxplot(exp_C9.LCM, las=2,
        col = c("red","red","red","red","red","red","red",
                "red","green","green","green"), main = "C9orf72")
legend(100, 100, legend=c("Disease", "Control"), col=c("red","green"))

etc...
```

Output below (red = patient, green = control)



Here you can see that the variance in CHMP2B is perhaps not as bad as it looks through PCA. We decided that we will conduct GSEA on the data set with and without that sample (JK.85.07.CEL), just to see if it makes any kind of difference. If it does, we may consider leaving it out, but only if we feel confident that it is the right thing to do. Otherwise, there is very little variance to be seen across all the data sets.

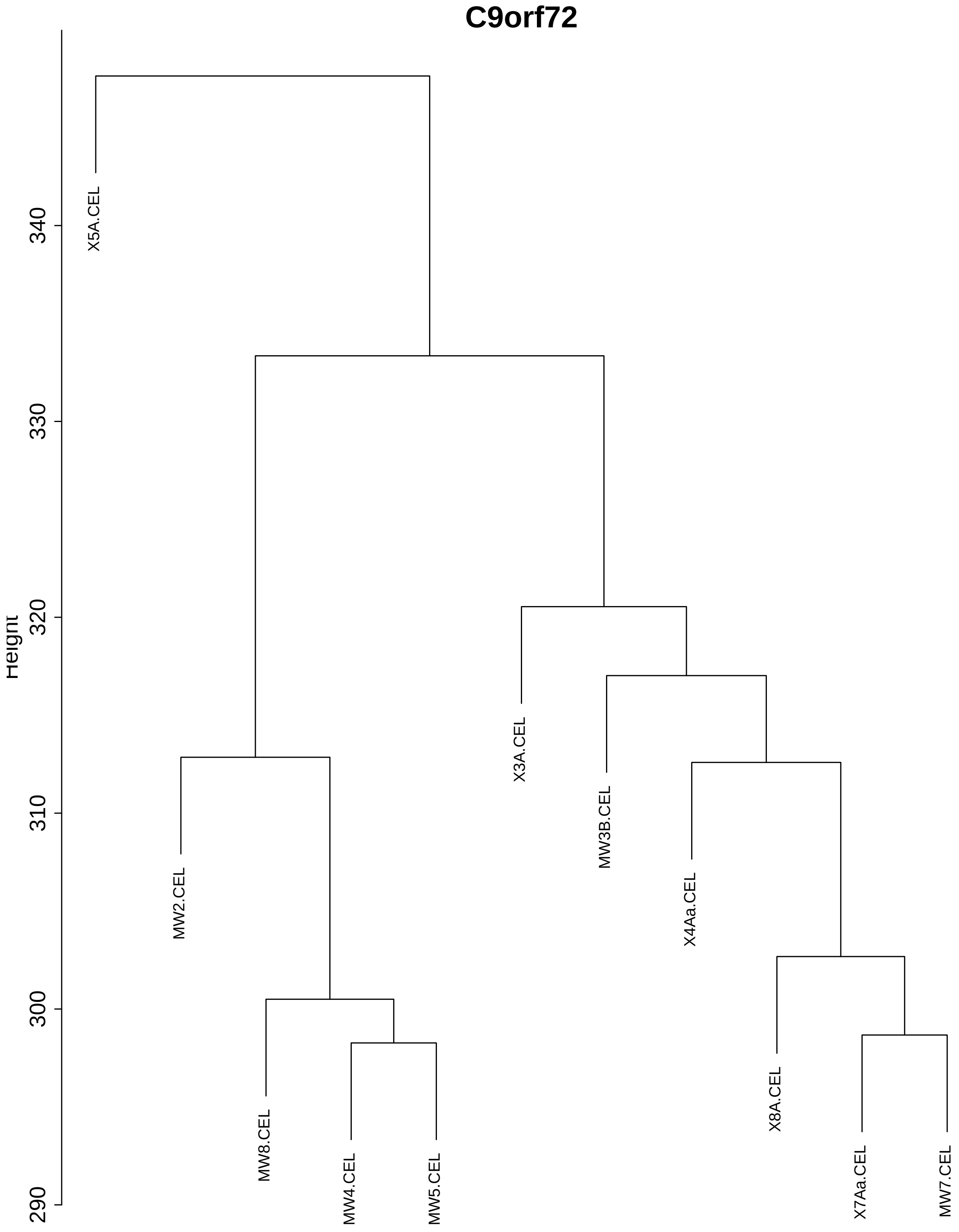
3. I then turned to conducting GSEA on my 5 datasets, but this turned out to be more complicated than I expected. There is the GSEA software package that I mentioned above but it seems a little "one size fits all" and it's clear that I will be better off learning how to write my own analysis using the packages available. The problem I came to was knowing which packages were the best ones to use (and then how to actually use them). The Broad Institute have a 10 year old script that they provide which looks extremely complicated and not particularly well written, and then there are R packages such as TopGO and GSEABase. John sent me a script that seems to use a combination of WGCNA and BiomaRt, so tomorrow I will try and understand this further.

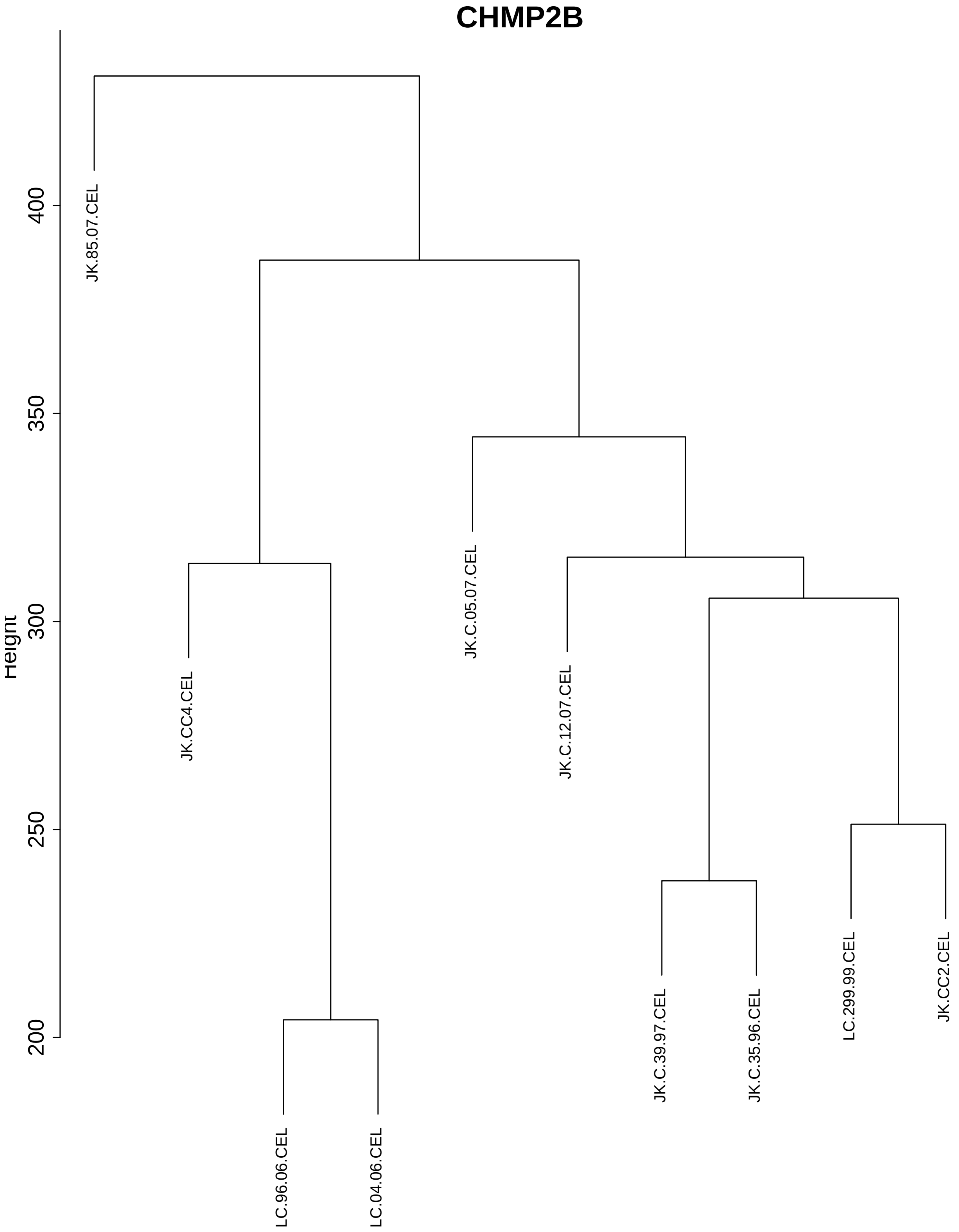
Thursday

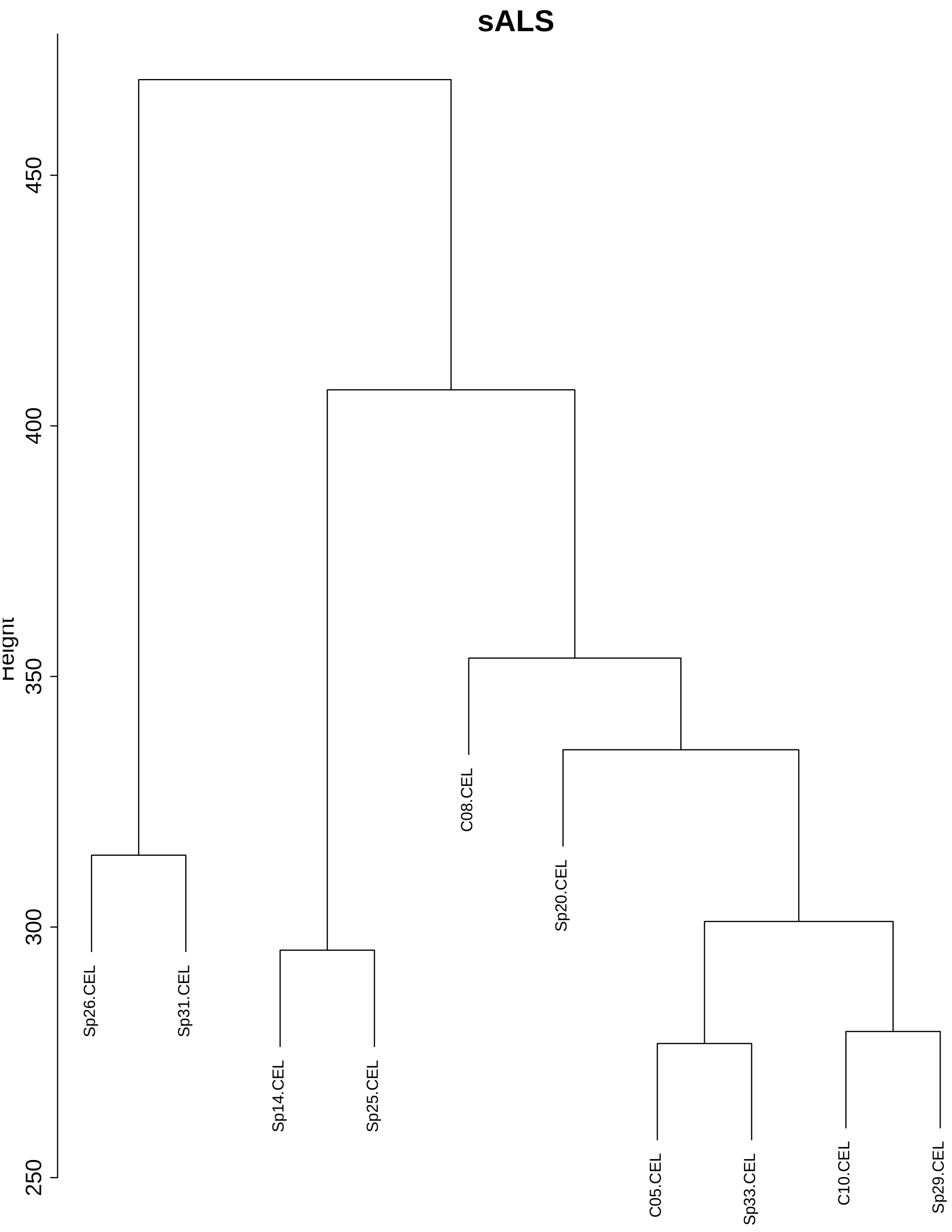
1. Until the GSEA could be addressed, I spent some time looking at WGCNA and playing with sample clustering to detect outliars in the data sets. To do this, I developed the following code:

```r
library(WGCNA)

### C9orf72 ###
# Display the current working directory
setwd ("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets/C9orf72_LCM")
options(stringsAsFactors = FALSE);

#Read in data set
exp_C9.LCM <- read.csv ("eset_NineP_150612_exprs.csv", header=TRUE)

row.names (exp_C9.LCM) <- exp_C9.LCM[,1] #specify that first column contains gene names
exp_C9.LCM<- exp_C9.LCM[,2:12] #specify that all other columns are gene expression data
exp_C9.LCM <- t(exp_C9.LCM) #transpose data set so that samples are rows
                            #and genes are columns (required by sampleTree)

###Check that there are any excessive missing values and identification of outlier
#microarray###
gsg = goodSamplesGenes(exp_C9.LCM, verbose = 3);
gsg$allOK

#If this comes out as False, look up documentation here http://labs.genetics.ucla.edu/horvath
#/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-01-dataInput.pdf

###Cluster the samples###

C9_sampleTree = hclust(dist(exp_C9.LCM), method = "average");

dev.off()
par(cex = 0.6);
par(mar = c(0,4,2,0))
plot(C9_sampleTree, main = "C9orf72", sub="", xlab="", cex.lab = 1.5,
     cex.axis = 1.5, cex.main = 2)
```

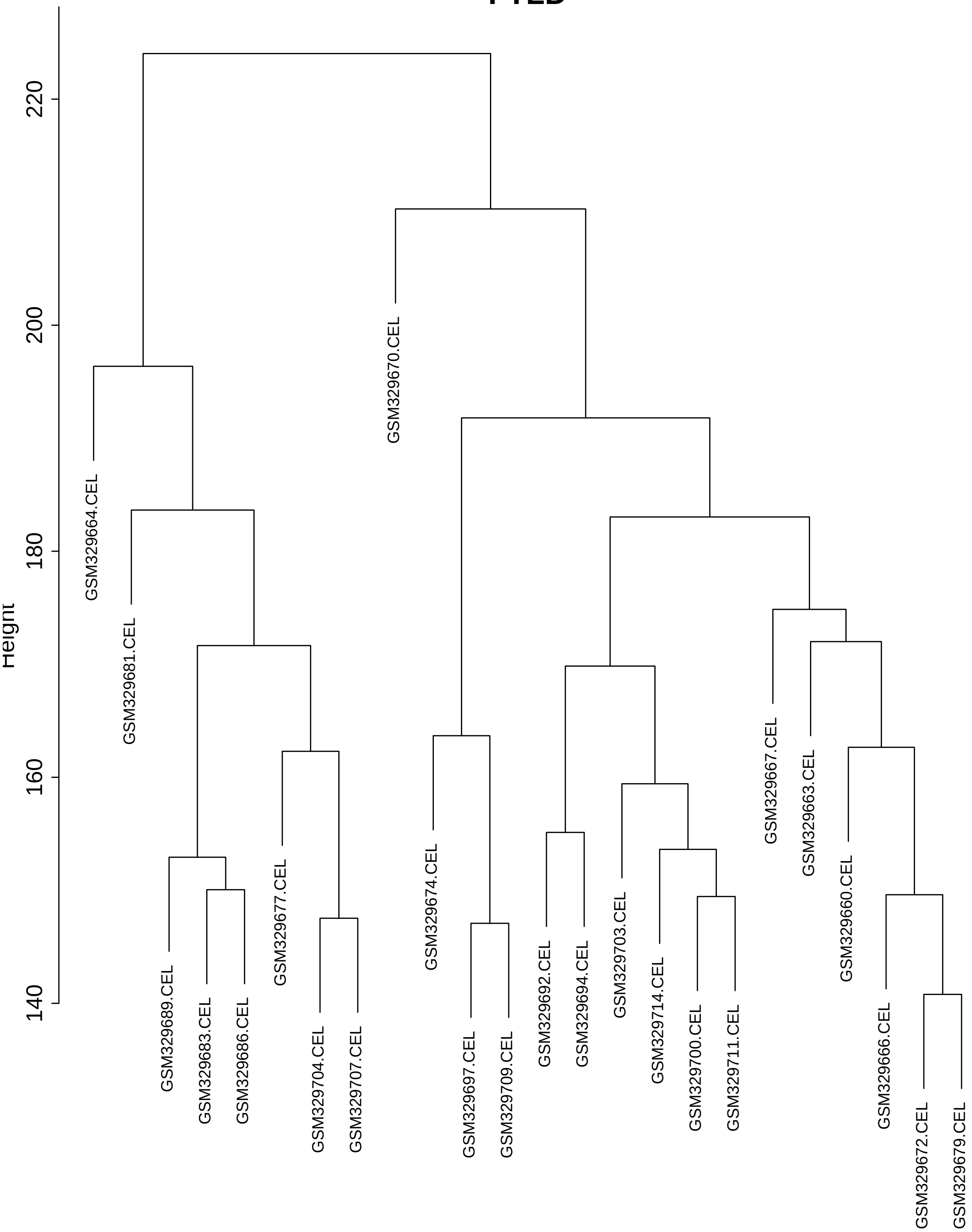
C9orf72 example And so on for the other data sets. . . results are:

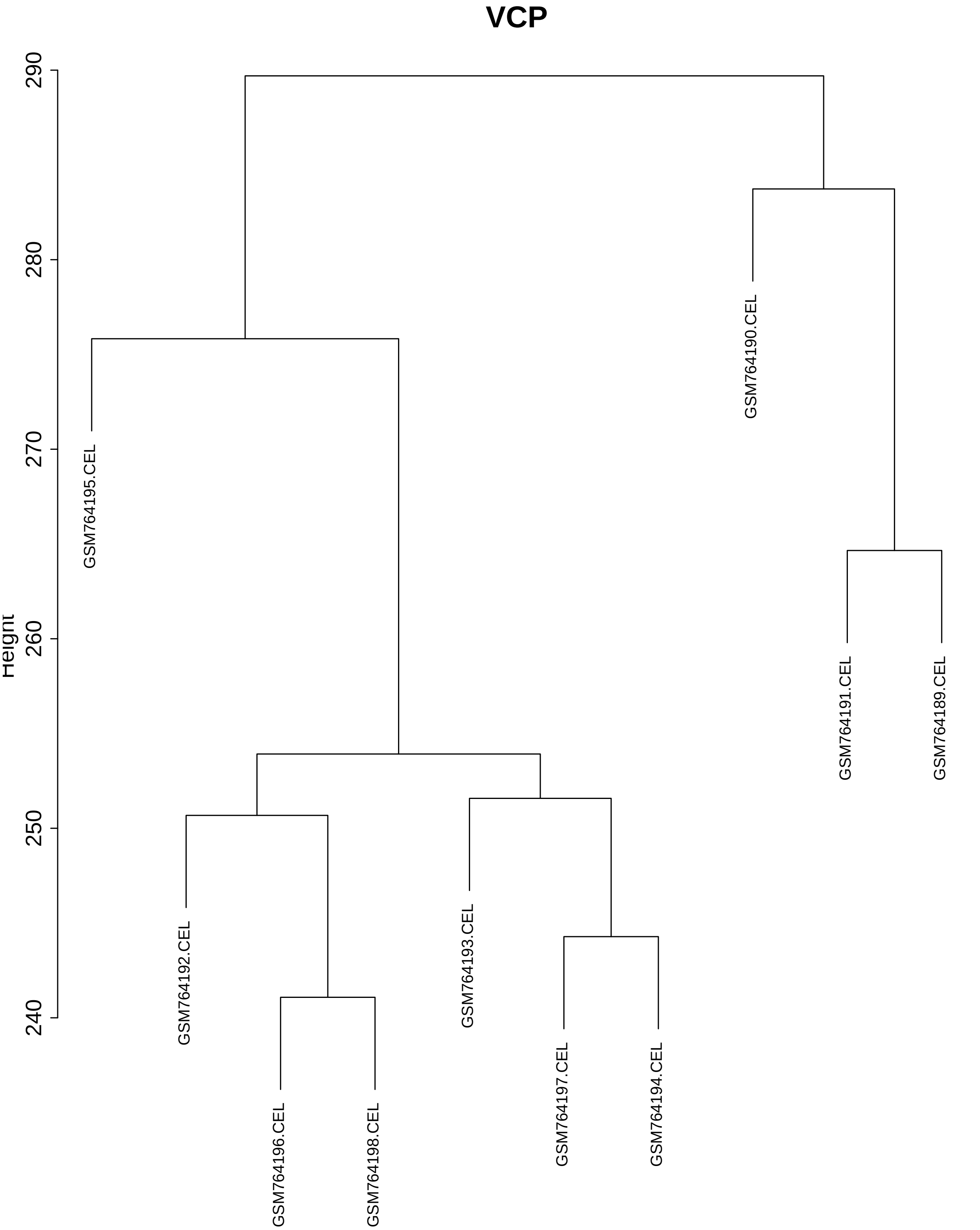

C9orf72
Height
340
330
320
310
300
290
X5A.CEL
MW2.CEL
MW8.CEL
MW4.CEL
MW5.CEL
X3A.CEL
MW3B.CEL
X4Aa.CEL
X8A.CEL
X7Aa.CEL
MW7.CEL


CHMP2B
Height
200
250
300
350
400
JK.85.07.CEL
JK.CC4.CEL
LC.96.06.CEL
LC.04.06.CEL
JK.C.05.07.CEL
JK.C.12.07.CEL
JK.C.39.97.CEL
JK.C.35.96.CEL
LC.299.99.CEL
JK.CC2.CEL


sALS
Height
450
400
350
300
250
Sp26.CEL
Sp31.CEL
Sp14.CEL
Sp25.CEL
C08.CEL
Sp20.CEL
C05.CEL
Sp33.CEL
C10.CEL
Sp29.CEL

FTLD


Height
220
200
180
160
140
GSM329664.CEL
GSM329681.CEL
GSM329689.CEL
GSM329683.CEL
GSM329686.CEL
GSM329677.CEL
GSM329704.CEL
GSM329707.CEL
GSM329670.CEL
GSM329674.CEL
GSM329697.CEL
GSM329709.CEL
GSM329692.CEL
GSM329694.CEL
GSM329703.CEL
GSM329714.CEL
GSM329700.CEL
GSM329711.CEL
GSM329667.CEL
GSM329663.CEL
GSM329660.CEL
GSM329666.CEL
GSM329672.CEL
GSM329679.CEL



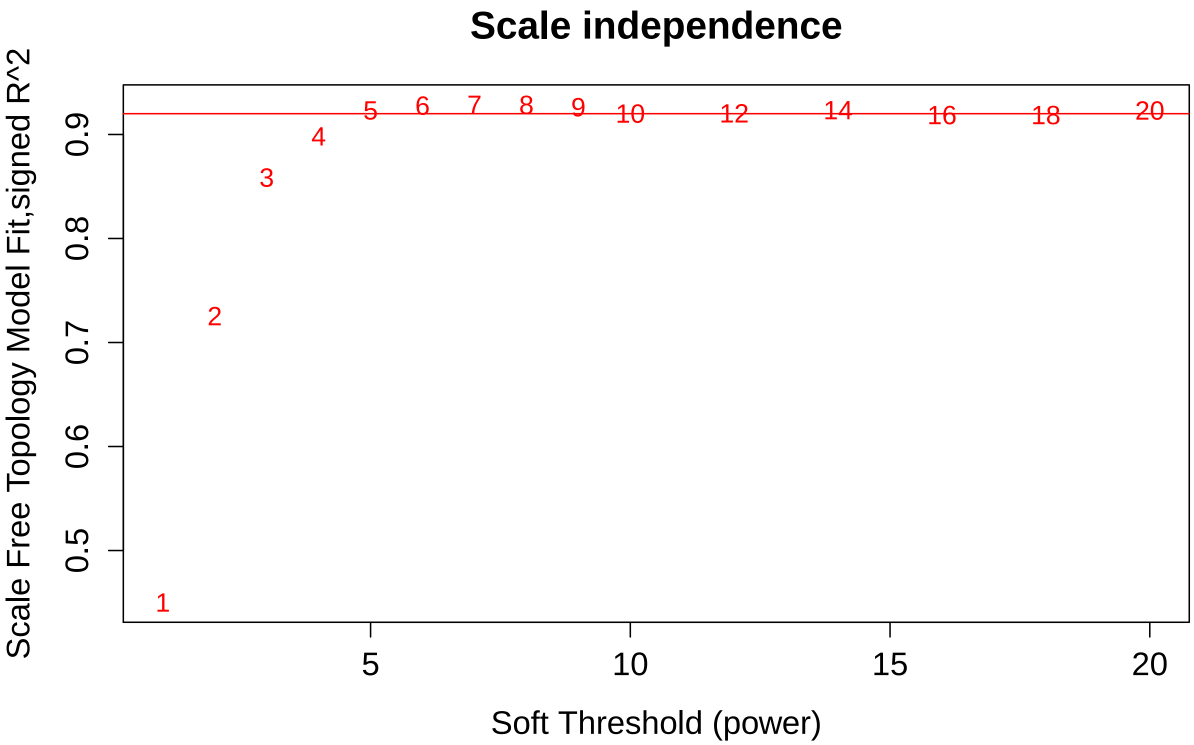
Again, a couple of red flags: the X5A sample from C9orf72 (look at page 7 to see where it lies on the 3D PCA plot) and the JK.85.07.CEL CHMP2B sample that has been suspected multiple times. These samples I may have to be wary of, and potentially test their effect on the GSEA.

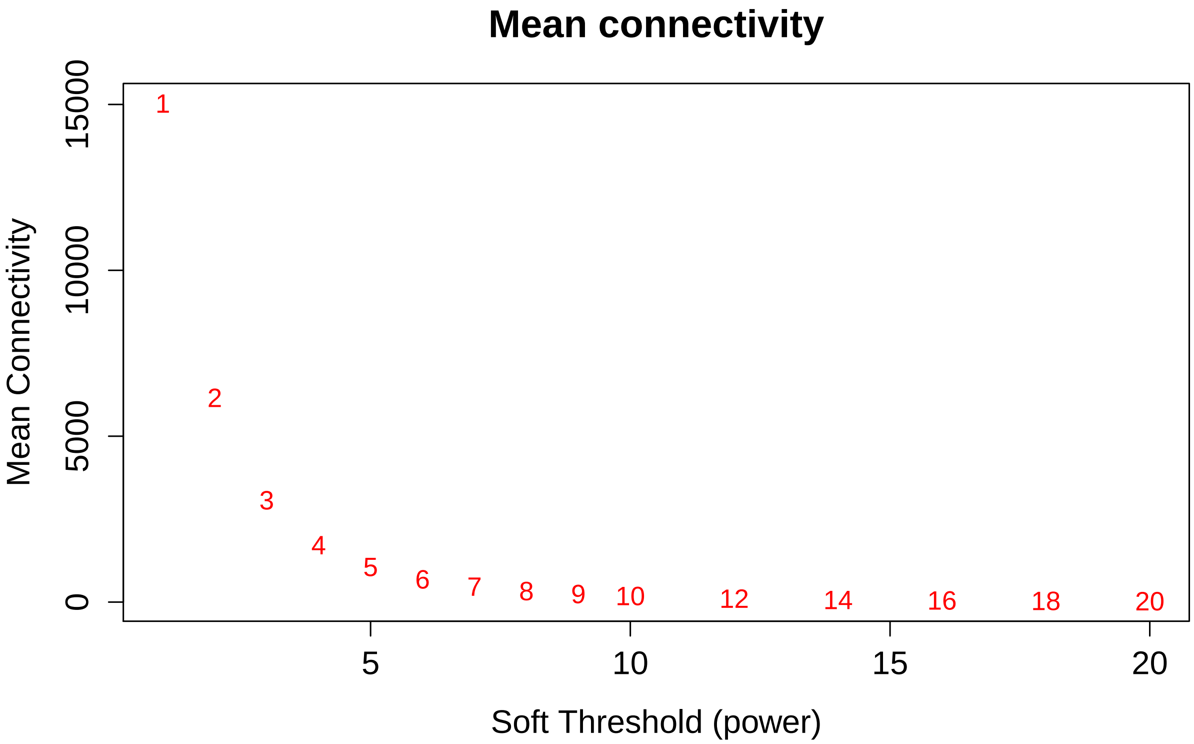
Friday

On Friday I continued investigating the capabilities of WGCNA by following the tutorials. I had a go at automatic network construction and module detection with the C9orf72 LCM Motor Neuron data.

To construct a gene network, I mostly used the code provided by the WGCNA tutorial, and adapted it for my data. The first step is to pick a soft threshold value which is the power to which expression is raised, and this allows the calculation of adjacency. The code is as follows:

```r
### Automatic construction of the gene network and identification of modules ###


##Choosing a threshold Power##

#Display the current working directory
setwd("/Users/clairegreen/Documents/PhD/TDP-43/TDP-43 Data Sets/Quality Control Analysis/WGCNA")


# Load the WGCNA package
library(WGCNA)
# The following setting is important, do not omit.
options(stringsAsFactors = FALSE);


# Load the data saved in the first part
lnames = load(file = "TDP-43_Analysis Environment.RData");
#The variable lnames contains the names of loaded variables.
lnames
exp_C9.LCM <- t(exp_C9.LCM)

# Choose a set of soft-thresholding powers
powers = c(c(1:10), seq(from = 12, to=20, by=2))

# Call the network topology analysis function
sft = pickSoftThreshold(exp_C9.LCM, powerVector = powers, verbose = 5)

# Plot the results:
par(mfrow = c(2,1));
cex1 = 0.8;
# Scale-free topology fit index as a function of the soft-thresholding power
plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
     xlab="Soft Threshold (power)",ylab="Scale Free Topology Model Fit,signed R^2",type="n",
     main = paste("Scale independence"));
text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
     labels=powers,cex=cex1,col="red");

# this line corresponds to using an R^2 cut-off of h
abline(h=0.92,col="red")

# Mean connectivity as a function of the soft-thresholding power
plot(sft$fitIndices[,1], sft$fitIndices[,5],
     xlab="Soft Threshold (power)",ylab="Mean Connectivity", type="n",
     main = paste("Mean connectivity"))
text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1,col="red")
```

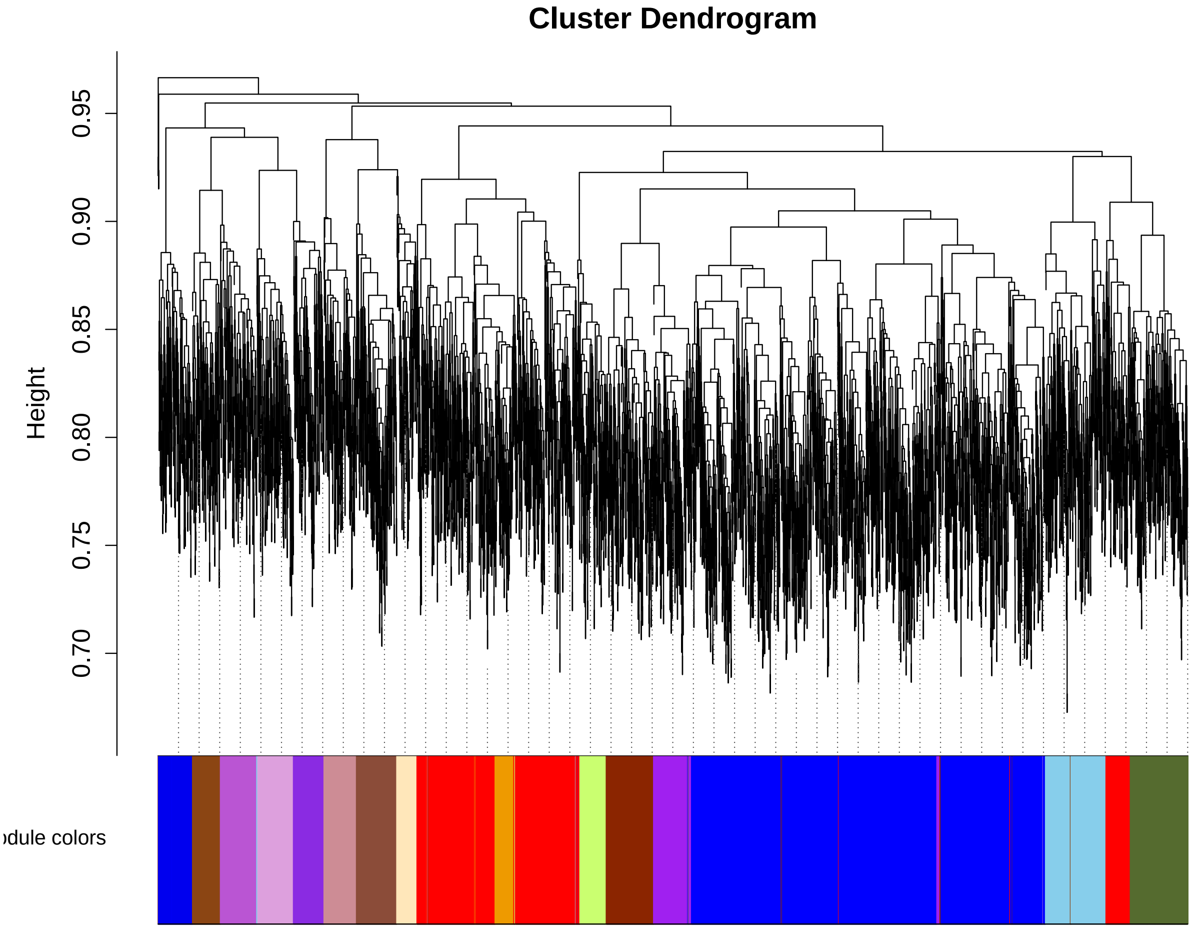




Results

As we can see from the graph, 5 is the lowest power for which the scale-free topology fit index curve flattens out upon reaching a high value (~0.92).

Finally, I used this threshold to perform network construction and module detection. The resulting dendrogram can be seen below:


Cluster Dendrogram
Height
odule colors