BCB570 Final Project

Urminder Singh

May 3, 2018

Data

I have RNA-seq and protein datasets from mouse. The RNA-seq dataset is tissue specific and extracted from 18 samples consisting of expression values of 4902 transcripts. The protein dataset consists of expression values of 5565 proteins over 89 samples where each sample was taken at a particular time. To make the datasets directly comparable I removed the etra proteins from the protein dataset and I am left with 4902 proteins which correspond to a transcript in the RNAseq dataset.

I assumed the data given to us was normalized and preprocessed and all the quantified values in the data reflects the true variation in the sample.

```
library(readr)
library("igraph", lib.loc = "~/R/win-library/3.4")
library("Hmisc", lib.loc = "~/R/win-library/3.4")
library("gplots", lib.loc = "~/R/win-library/3.4")
library(GENIE3)
library("doParallel", lib.loc = "~/R/win-library/3.4")
library("doRNG", lib.loc = "~/R/win-library/3.4")
library("MCL", lib.loc = "~/R/win-library/3.4")
library("mixOmics", lib.loc = "~/R/win-library/3.4")
library("yaml", lib.loc = "~/R/win-library/3.4")
library("stringi", lib.loc = "~/R/win-library/3.4")
Urminder <- read_csv("Urminder.txt", col_names = FALSE)
Protein <- read_csv("Protein.txt")
# remove extra prots
Protein <- subset(Protein, (rownames(Protein) %in% rownames(Urminder)))
```

Building individual networks

To build individual networks I used GENIE3. GENIE3 tries to predict protein/transcript in the using all others using regression with random forests. Using these regression models, association links are made between each pair of protein/transcripts.

GENIE3 initially gave a very dense network. I wanted to make a sparse network an keep only meaningful weights while getting rid of any insignificant associations. To do this I chose 10 thresholds and made network using these. The I looked at the density plot and chose a threshold which is near the elbow point in the graph i.e. the point where density suddenly decreases and gets stable. Fig1. shows these plots for the gene network.

Based on the results I chose the following threshold:

- For gene network 0.0086
- For protein network 0.0012

Network properties

```
library(readr)
netprop <- read_csv("netprop.csv")
netprop %>% knitr::kable(caption = "Network Features")
```

Table 1: Network Features

Feature	gene_network_GENIE	prot_network_GENIE
#Edges	9516.0000000	2.544700e+04
#Vertices	3749.0000000	4.519000e+03
Density	0.0013545	2.492700e-03
Diameter	12.0000000	1.600000e+01
Radius	1.0000000	1.000000e+00
Clustering_coeff(global)	0.0549661	2.550987e-01
Clustering_coeff(avg)	0.0659363	8.334630e-02
Avg Shortest Path	5.4621712	4.668710e+00

Powerlaw test

First I plotted the degree distribution of the networks. The degree distribution looks like it follows a power law in both cases. The i used power.law.fit to evaluate the a p-value. Results are in table 2.

```
library(readr)
plaw <- read_csv("plaw.csv")
plaw %>% knitr::kable(caption = "PowerLawTest")
```

Table 2: PowerLawTest

Graph	Alpha	X_min	KS_stat	P_Value	logLik
Gene network	5.102327	16	0.0606798	0.6303486	-391.7622
Protein network	3.151388	18	0.0181309	0.9407989	-3052.8238

From Table 2 we can see that p-value is much geater that significance level and alpha is low less than 6. Thus we can say that these networks follow power and a scale free as many other biological networks.

Finding clusters in the networks

After building my networks, I used Markov clustering to find interesting cluster in my networks. I wanted to compare the clusters and see if the clusters from the two networks are simillar or not.

Running MCL using R

First i ran mcl on my networks using R and default parameters. I found really big clusters in both networks (1475 in gene network, 1007 in protein) while other clusters being very small less that size 10. I chose these two big clusters for comparision and found that only 292 genes/proteins were common (see fig 7).

GO enrichment in the largest clusters

I used the web server GOrilla to find enriched go terms in these largest clusters. I found the gene cluster was enriched in GOterms GO:0001078, GO:0001047, and GO:0000982 which are all related to promoter binding. The protein cluster was enriched in GO:0005488, GO:0008276, and GO:0005515 which are related to protein binding and protein methyltransferase activity. None of the GO terms ovelapped in gene and protein cluster.

Intergrating protein-protein interaction data

To further investigate I downloaded protein-protein interaction (PPI) data from STRINGdb. Then, I mapped my set of genes to PPI network and kept only the edges with high string score i.e. string score above 900. This was done to remove any weakly linked nodes. I used MCL on this network and I got similar cluster distribution and found one big cluster of size 1763. I compared this to the other two biggest cluster and not many genes were common (fig 8).

Then I did GO enrichment on this cluster and fond GO terms GO:0005515, GO:0017016, GO:0005488, GO:0051020, GO:0031267, GO:0017048. From this there were two GO terms which overlapped the protein cluster but none overlapped with the gene cluster.

Running MCL using python

The above clustering results produced really big clusters, to get different clusters I ran MCL using diffferent parameters with python. Running MCL with python is much faster than with R. I set the inflation to 1.4 and expansion to 2. Using this I found 307 clusters in protein and 289 clusters in gene (see protein_mcl_clusters.csv and gene_mcl_clusters.csv). Then to compare these clusters I calculated, pairwise, how many elements are common between the clusters in the two networks. I found that cluster138 from gene (size 38) and cluster17 from protein (size 103) had 10 elements in common. All other had 0,1 or 2 in common (see scores.txt). Surprisingly, none of the clusters were significantly rich in any GO term.

Integrating data

First I did PCA on my data sets. My, dataset looked separable in first two PC (see figs 9, 10,11,12). Please see dataintegration.pdf for full details.

Predicting tissue type

To predict the tissue type with some confidence we should download known tissue specific expression patterns of the given genes and then see which pattern correlates the most with our expression. This approach need additional data to work with. Therefor, I used another approach. First, I split my genes into two sets higly expressed and lowly expressed genes (see highexp.txt and lowexp.txt). Then, I used these genes to as query at http://www.informatics.jax.org/gxd.

I found that my list of least/unexpressed genes are least expressed in heart and liver. Further, the high expressed genes were expressed mostly in brain,testis,heart, liver and lung tissues where expression levels in heart and liver were less than other. This is because certain genes can have high levels of basal expression throughout the body e.g. genes involved in ribosome pathways like Rpl32 etc. Further, I filtered all the samples to be wild type. Then, the low expressed genes were distinctively lower in heart sample than in any other tissue type (see fig 13). Then I checked which genes are not expressed in heart but expressed in liver. I found that 40 of such genes were in my low expressed genes. Based on these observations I can guess that my sample is from heart.

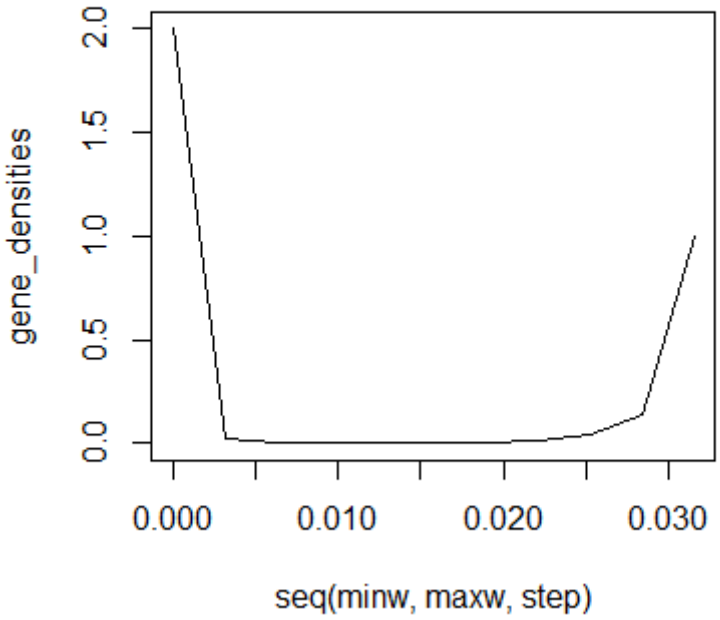

Figure 1: Density vs Threshold in gene network

References

1. Wu XM, Ma X, Tang C, Xie KN, Liu J, Guo W, Yan YL, Shen GH, Luo EP. Protein-protein interaction network and significant gene analysis of osteoporosis. Genet Mol Res. 2013 Oct 18;12(4):4751-9. doi: 10.4238/2013.October.18.12.
2. Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. Bioinformatics. 2015 Jun 15;31(12):i197-205. doi: 10.1093/bioinformatics/btv268.
3. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010 Sep 28;5(9). pii: e12776. doi: 10.1371/journal.pone.0012776.
4. Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.
5. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009 Feb 3;10:48. doi: 10.1186/1471-2105-10-48.
6. Jacqueline H. Finger et.al. The mouse Gene Expression Database (GXD): 2017 update. Nucleic Acids Res. 2017 Jan 4; 45(Database issue): D730-D736.

Figures

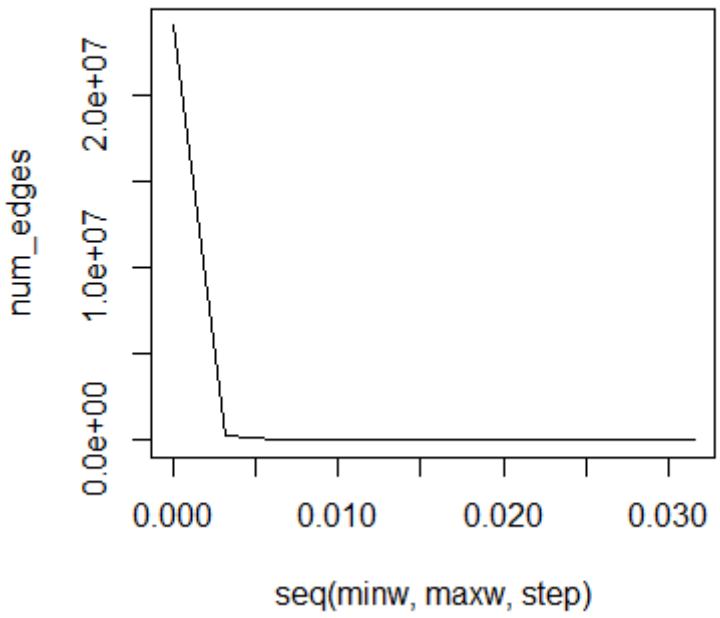

Figure 2: Density vs Edges in gene network

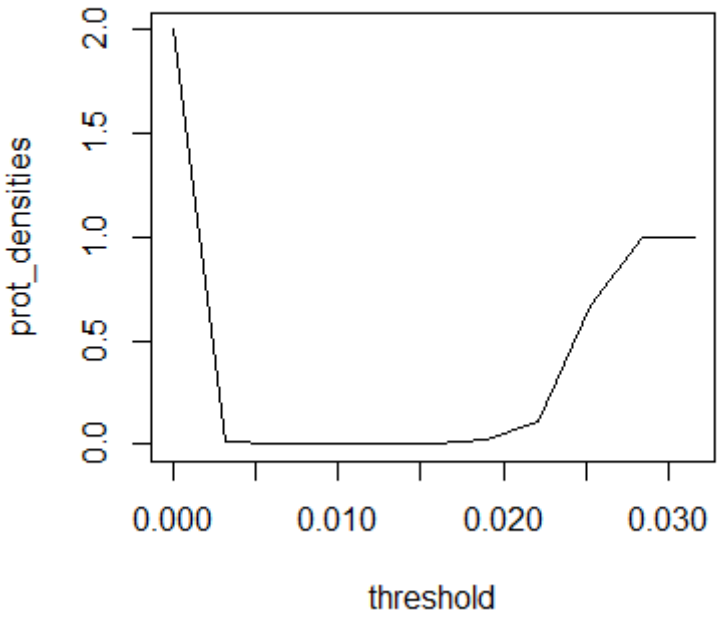

Figure 3: Density vs Threshold in protein network

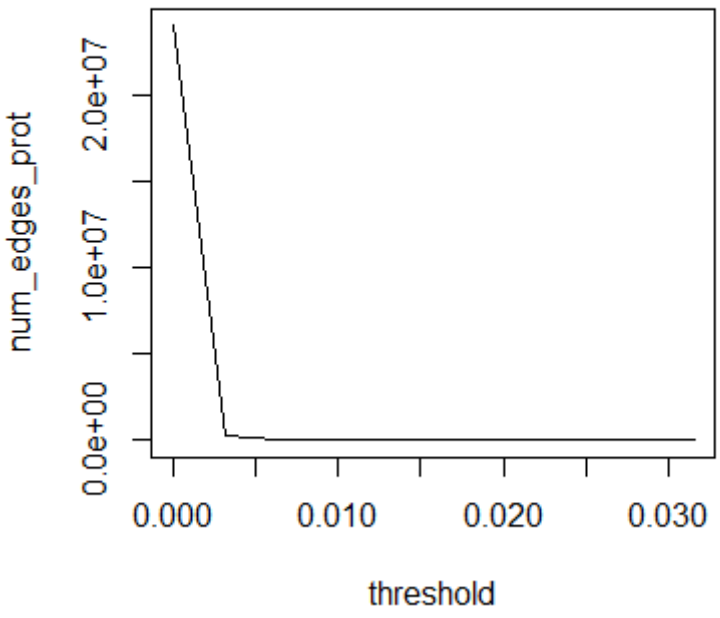

Figure 4: Density vs Edges in protein network

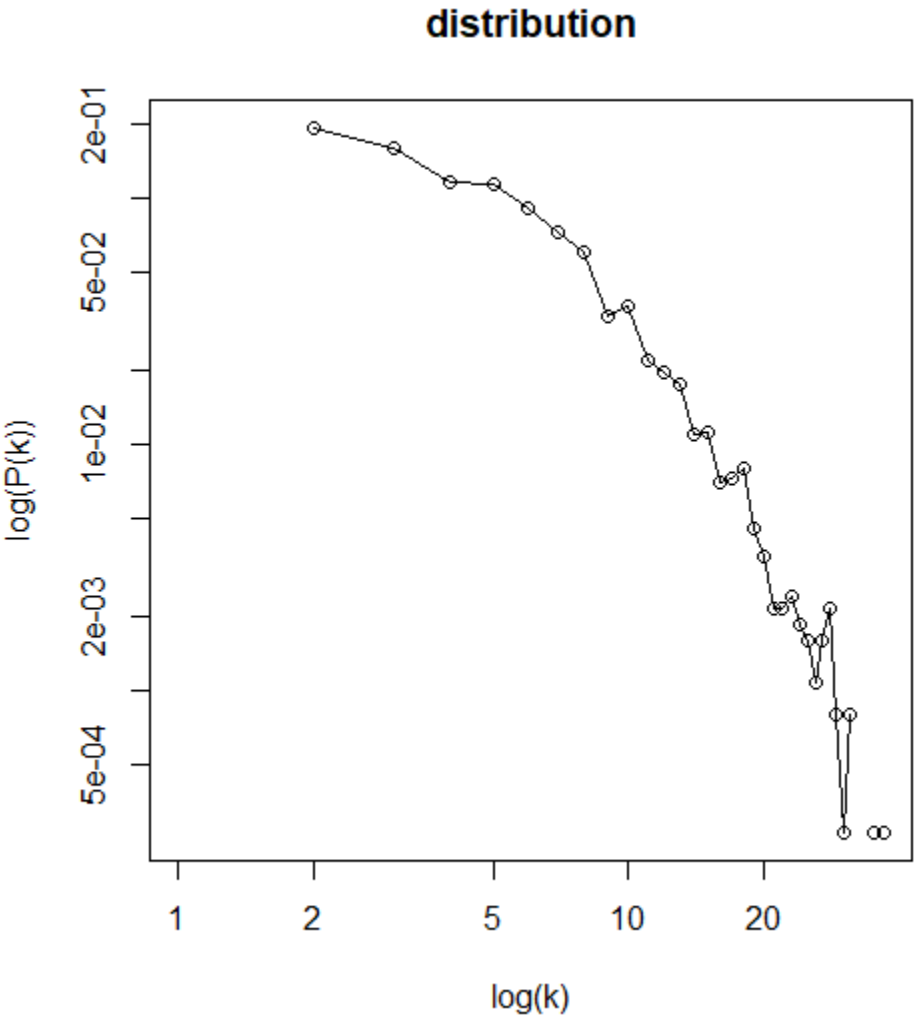

Figure 5: Degree distribution of gene network

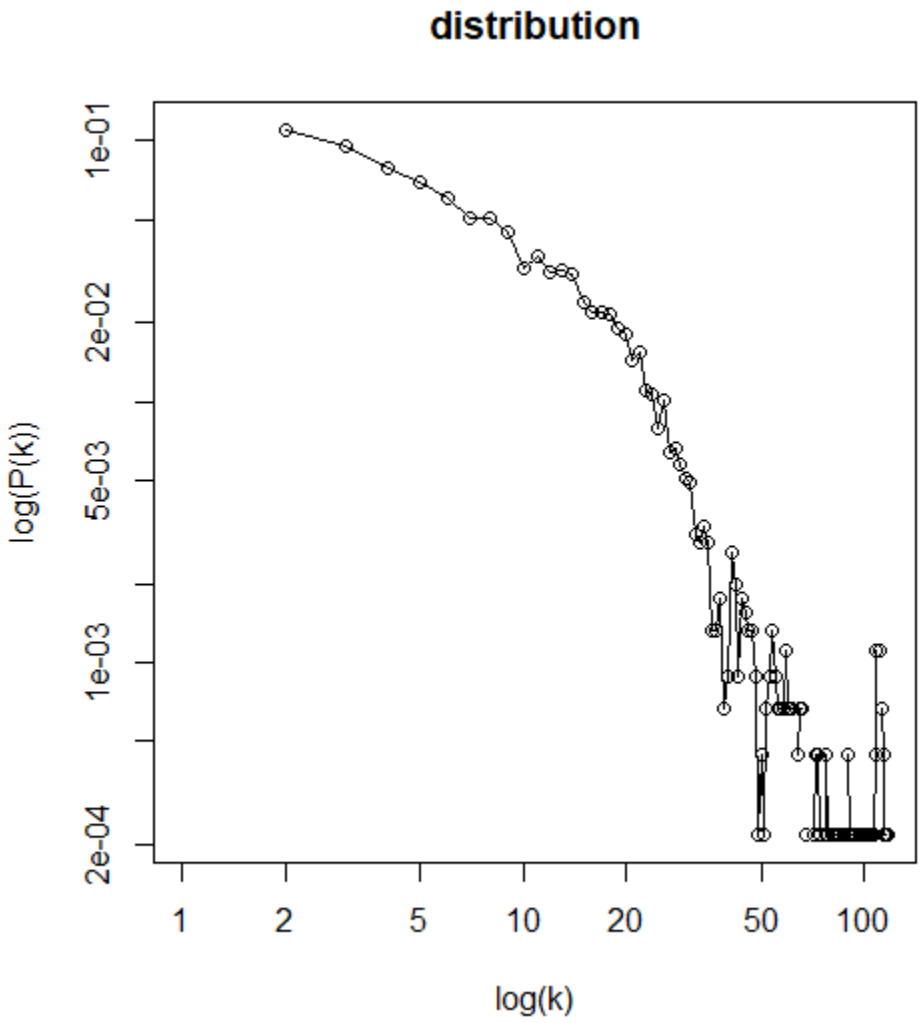

Figure 6: Degree distribution of protein network

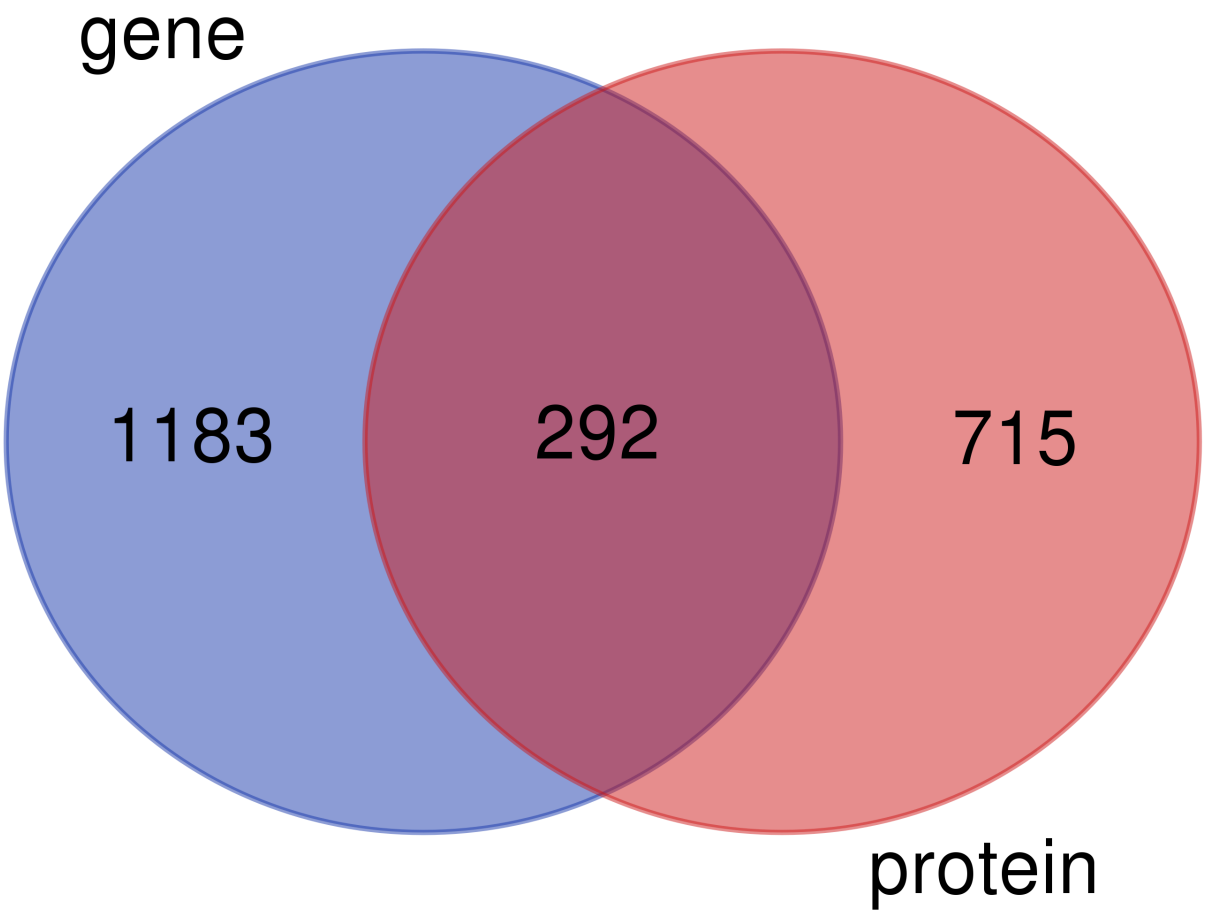

Figure 7: Largest cluster comparision

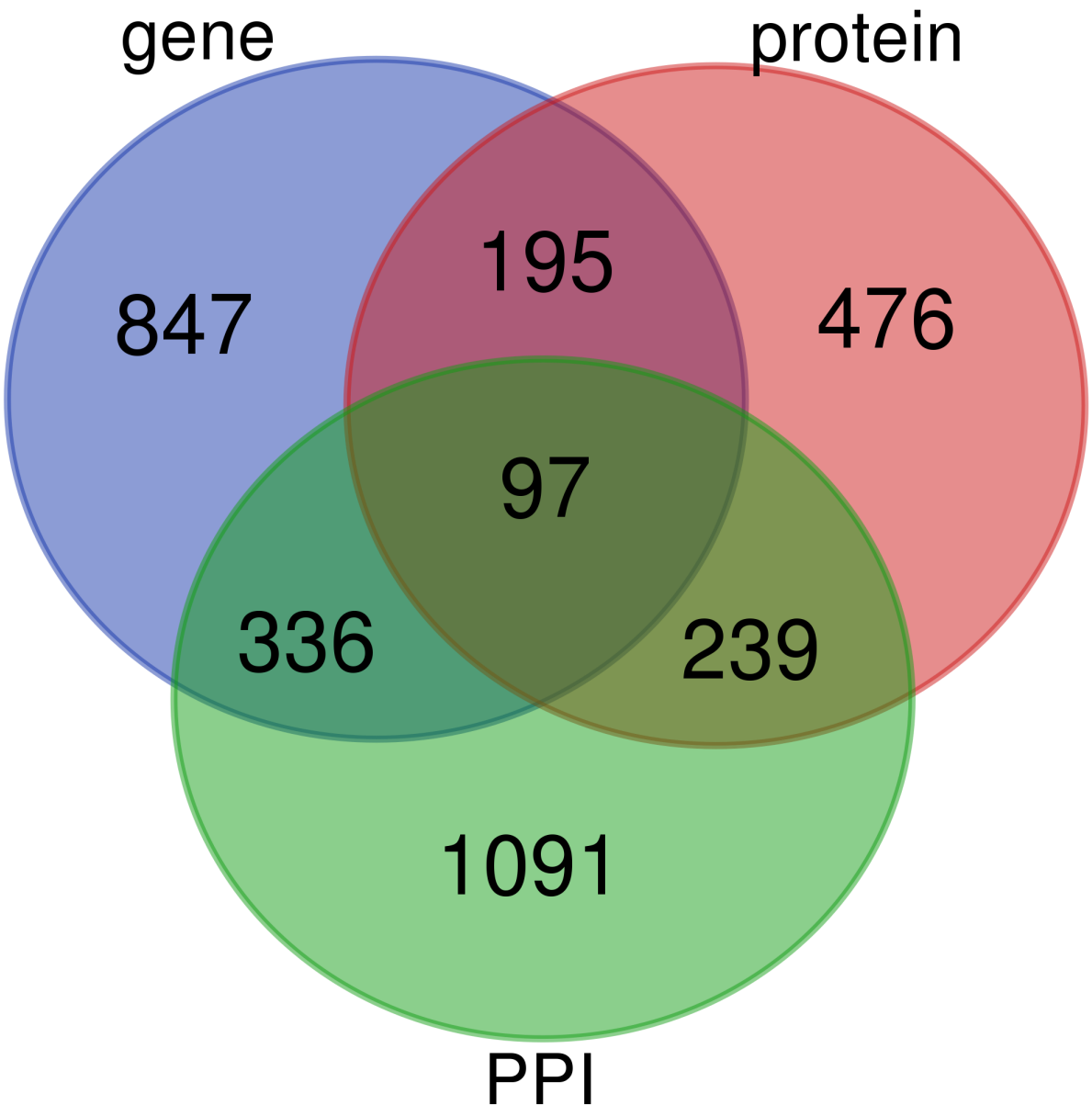

Figure 8: Largest cluster comparision

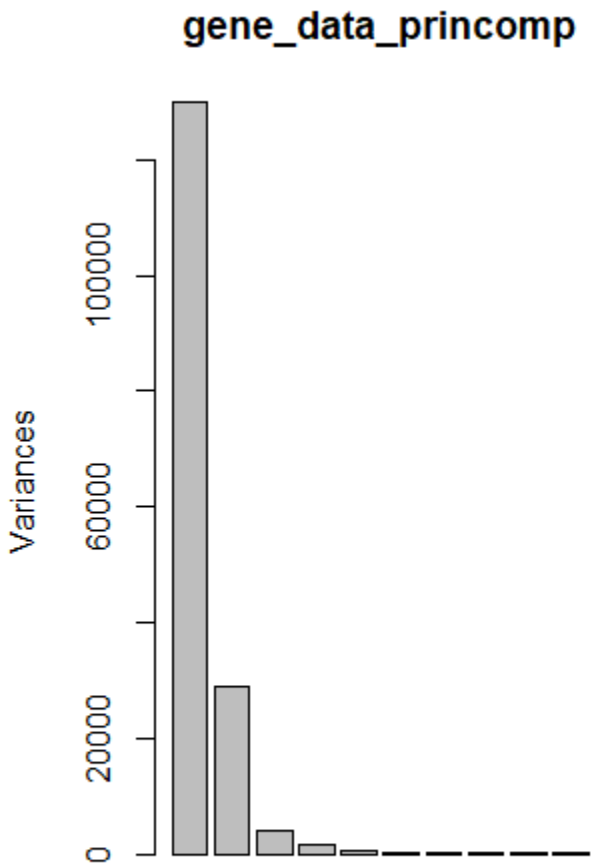

Figure 9: PCA explained variance for gene data

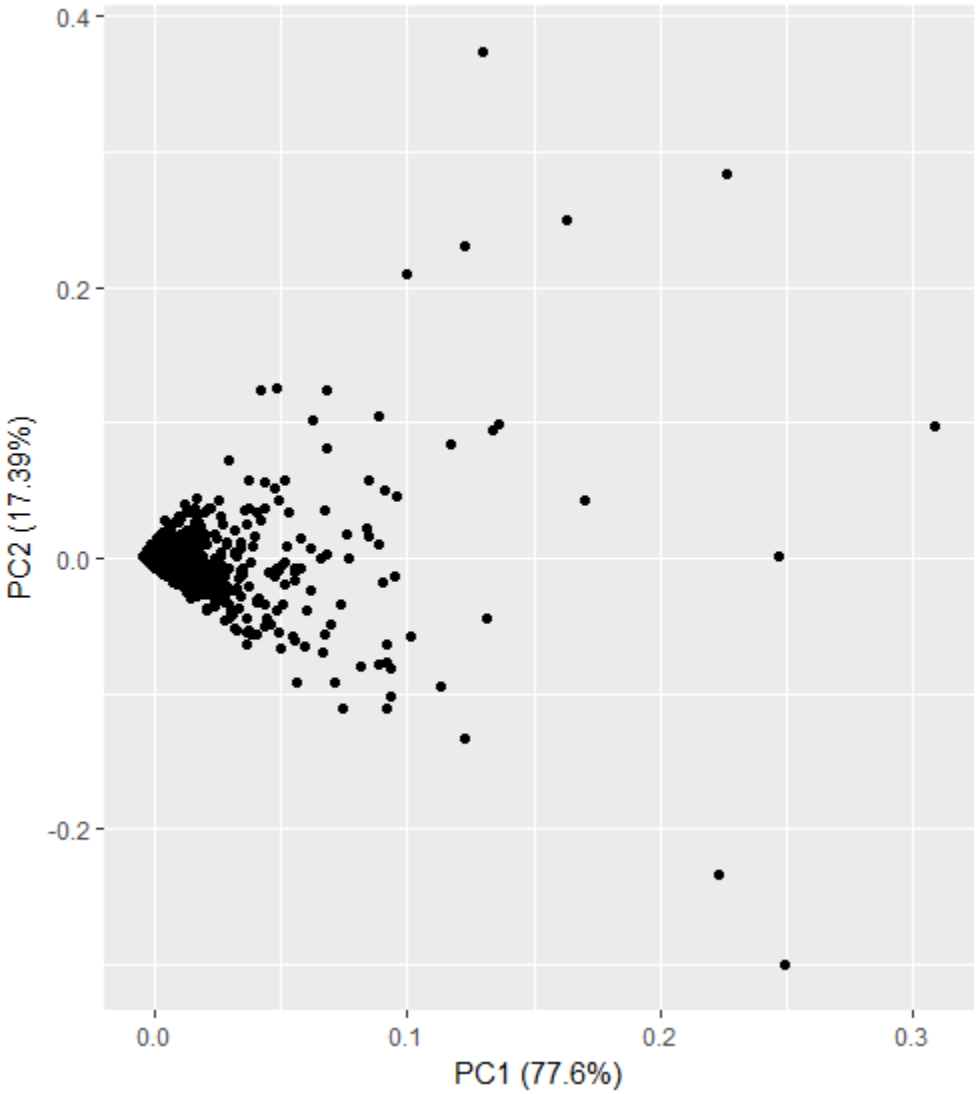

Figure 10: PCA plot for gene dataset

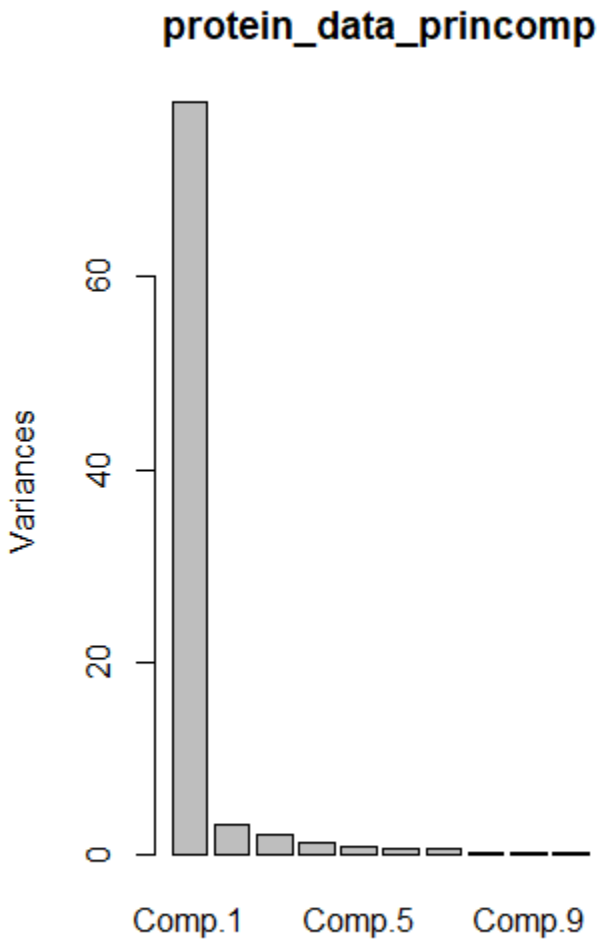

Figure 11: PCA explained variance for protein data

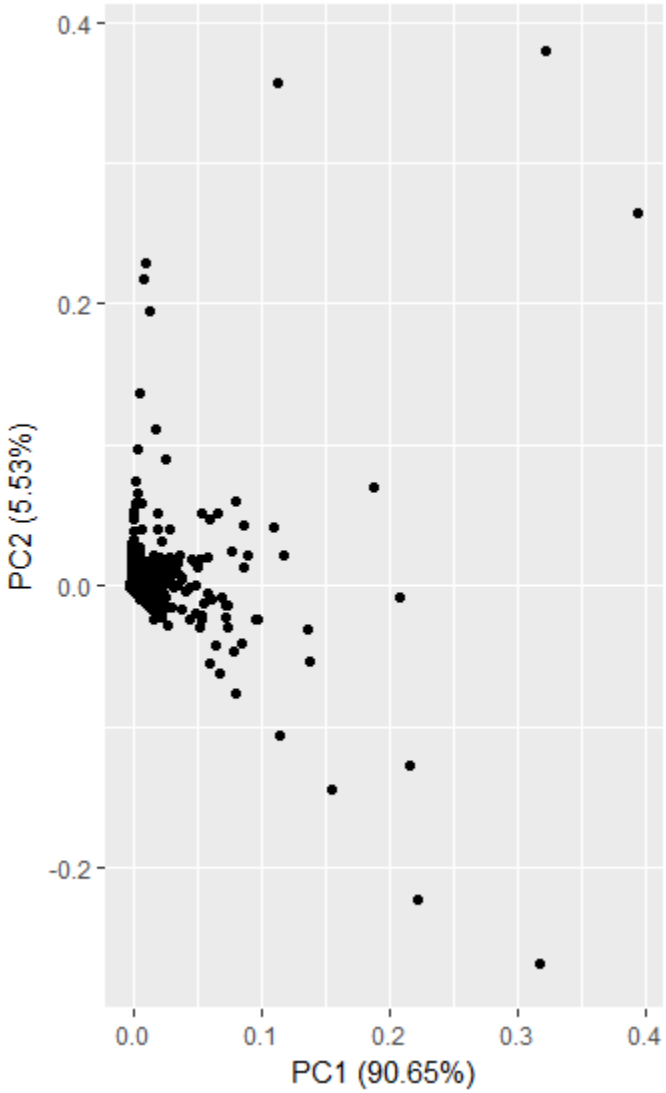

Figure 12: PCA plot for gene dataset

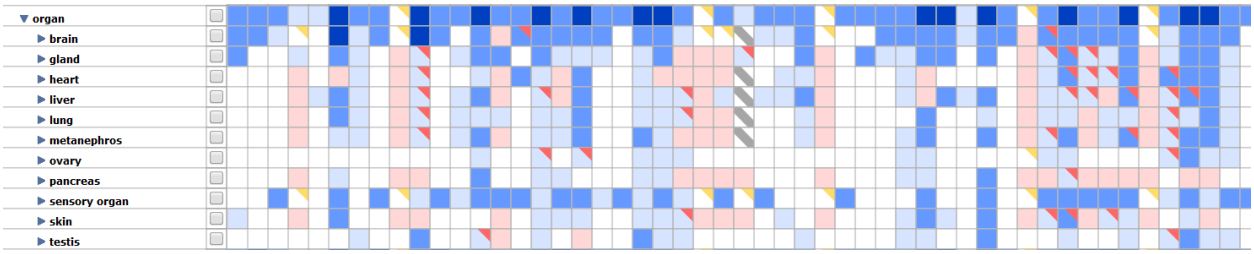

Figure 13: Gene expression of low expressed gene from MGI. Blue: high expression. Red Low/No expression