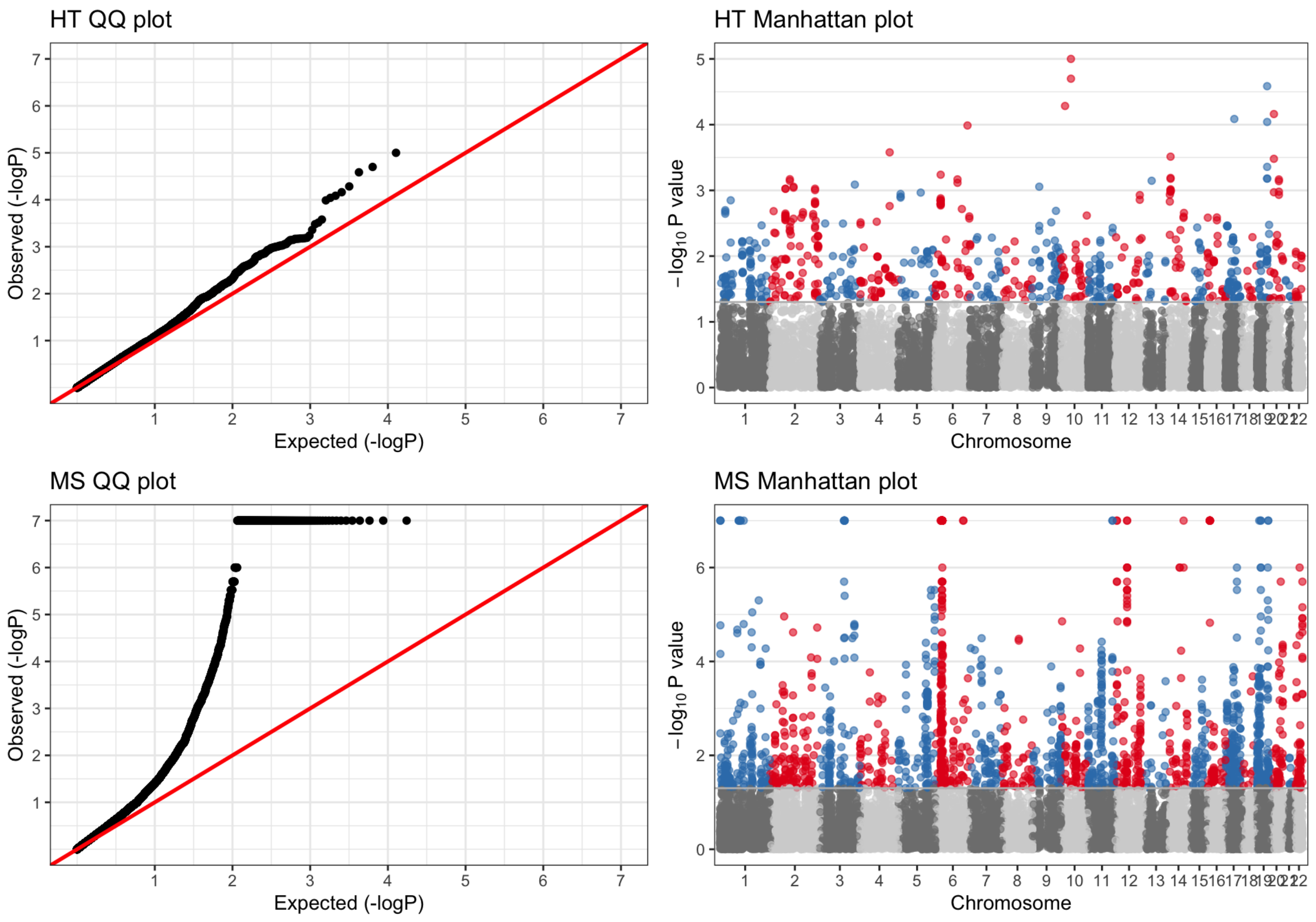
BMI 206: Networks Lab

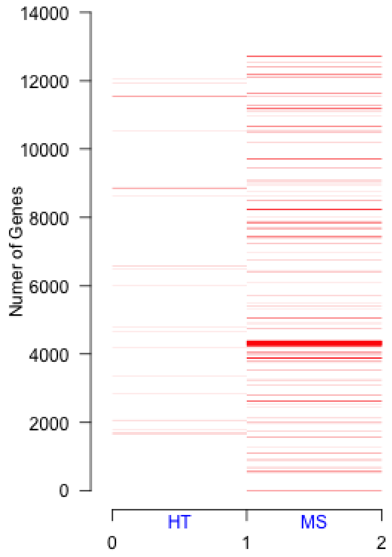
```r
#turned sergio's code into functions for ease.

sergio_qqplot <- function(pvals){
  observed <- sort(pvals$GenePvalue)
  lobs <- -(log10(observed))

  expected <- c(1:length(observed))
  lexp <- -(log10(expected / (length(expected)+1)))

  g1 <- ggplot(data = data.frame(cbind(lexp, lobs)), aes(x=lexp, y=lobs)) +
  geom_point() +
  xlab("Expected (-logP)") + ylab("Observed (-logP)") +
  theme_bw() +
  scale_x_continuous(breaks=c(1:7), limits = c(0,7)) +
  scale_y_continuous(breaks=c(1:7), limits = c(0,7)) +
  geom_abline(intercept = 0, slope = 1, color="red", size=1)
  NULL
  return(g1)
}

sergio_manhattan <- function(wtcc){
  wtcc$Start <- wtcc$Start/1000
  for(j in 2:22){
    wtcc[wtcc$Chr==j,"Start"] <- max(wtcc[wtcc$Chr==(j-1),"Start"])+wtcc[wtcc$Chr==j,"Start"]
    wtcc[wtcc$Chr==j, "Tick"] <- (min(wtcc[wtcc$Chr==j,"Start"]) +
                                  max(wtcc[wtcc$Chr==j,"Start"]))/2}
  wtcc[wtcc$Chr==1,"Tick"] <- (min(wtcc[wtcc$Chr==1,"Start"]) + max(wtcc[wtcc$Chr==1,"Start"]))/2
  wtcc$Discovery_log <- -log10(wtcc$GenePvalue)
  wtcc$Color_Dis <- wtcc$Chr %% 2
  wtcc$Color_Dis <-ifelse(wtcc$GenePvalue< 0.05, (wtcc$Chr %% 2)+2, wtcc$Color_Dis)

  colours <- c("#D3D3D3","#808080",brewer.pal(n = 3, name = "Set1"))

#pdf("Manhattan_plot.pdf",width=12,height=6)
  g1 <- ggplot(wtcc, aes(Start, Discovery_log)) +
  geom_point(size=1.5,alpha=0.6,aes(colour=as.factor(Color_Dis)))+
  scale_colour_manual(values = colours) +
  #scale_color_brewer(palette="Set1")+
  geom_hline(yintercept=-log10(0.05),size=0.5, colour="gray")+
  ylab(expression(paste(-log[10]~'P value')))+
  theme_bw()+
  theme(legend.position = "none",
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank()) +
  scale_x_continuous(name = "Chromosome", breaks = unique(wtcc$Tick), labels = unique(wtcc$Chr),expand=c(0.01,0))

  return(g1)
}
```

Q1: After exploring the Manhattan plots, and qq-plots from each GWAS, what can you tell about the power of each study?

ANS: The MS GWAS has a higher power than the HT study. This can be seen by the flat line in the Q-Q plot of the latter. Advanced comment: The early departure from the diagonal in the MS study, might suggest genomic inflation, but its power is clearly superior.

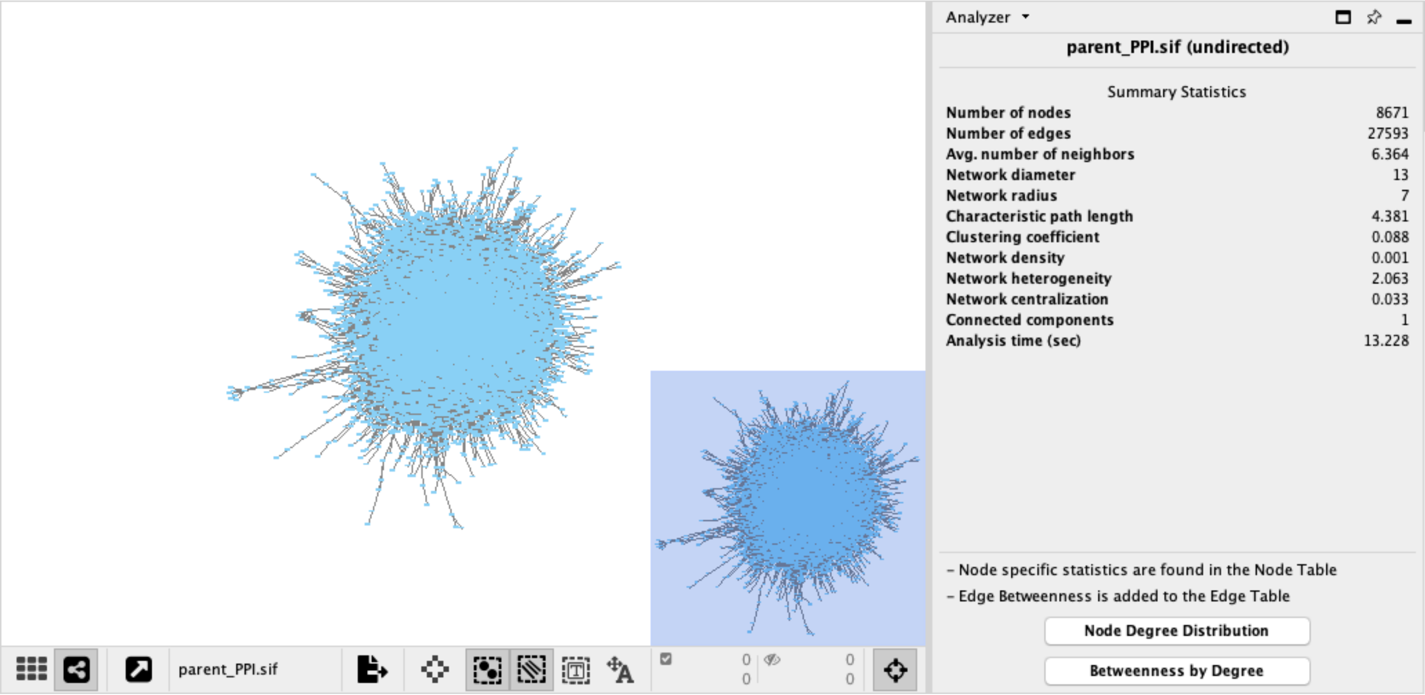
```
g1 <- sergio_qqplot(read.table("MS.pvals.out", header=T))
g2 <- sergio_manhattan(read.table("MS.pvals.out", header=T, as.is=T))

g3 <- sergio_qqplot(read.table("HT.pvals.out", header=T))
g4 <- sergio_manhattan(read.table("HT.pvals.out", header=T, as.is=T))

grid.arrange(g3+ggtitle("HT QQ plot"), g4+ggtitle("HT Manhattan plot"),
             g1+ggtitle("MS QQ plot"), g2+ggtitle("MS Manhattan plot"),
             ncol=2)
```





Q2: Using Cytoscape, analyze the PPI and describe its main network properties (this may take 20-40 min! do at home)

ANS: There are ~8K nodes (proteins) and ~27K edges (protein interaction/binding). The network is scale-free. Answers could also describe the clustering coefficient, and closeness metrics.

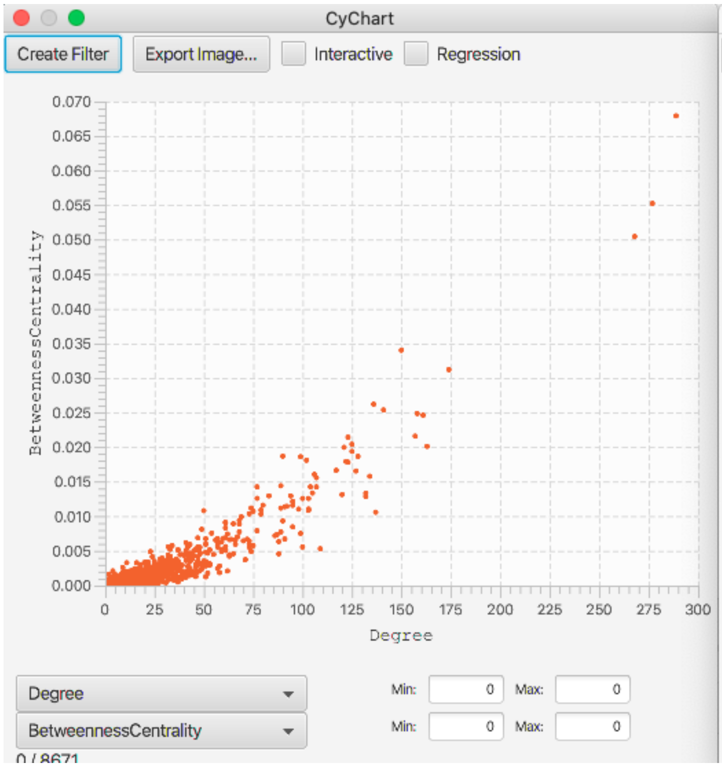


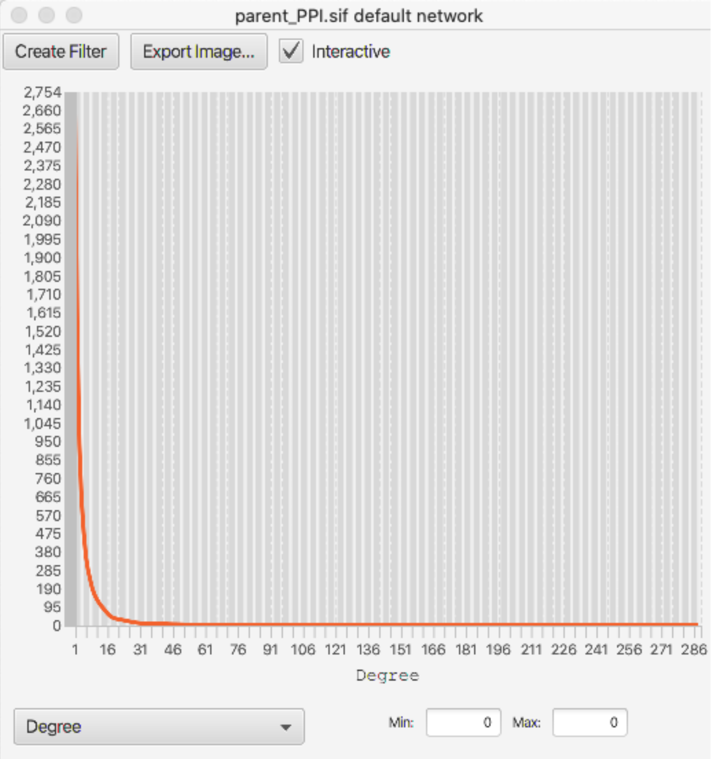
parent_PPI.sif (undirected)

Summary Statistics

Number of nodes	8671
Number of edges	27593
Avg. number of neighbors	6.364
Network diameter	13
Network radius	7
Characteristic path length	4.381
Clustering coefficient	0.088
Network density	0.001
Network heterogeneity	2.063
Network centralization	0.033
Connected components	1
Analysis time (sec)	13.228

- Node specific statistics are found in the Node Table
- Edge Betweenness is added to the Edge Table





Q3: Using Cytoscape, find the first order networks (p<0.05) for each GWAS

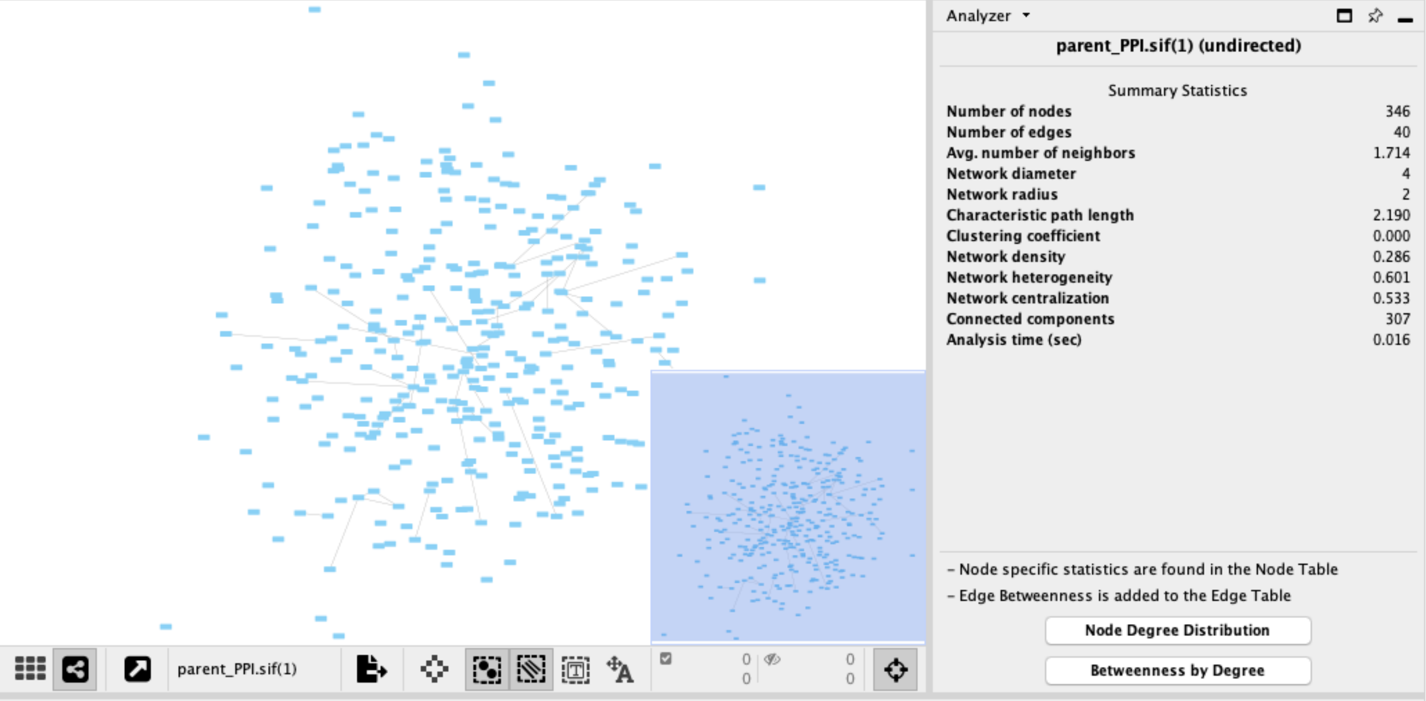
ANS:

Filter > + > Node:Pvalue between 0 and 0.05 inclusive (N=346 for HT, N=667 for MS)

File > New Network > From selected nodes, all edges

Tools > Analyze Network

The HT first order network contains 346 nodes and 40 edges.

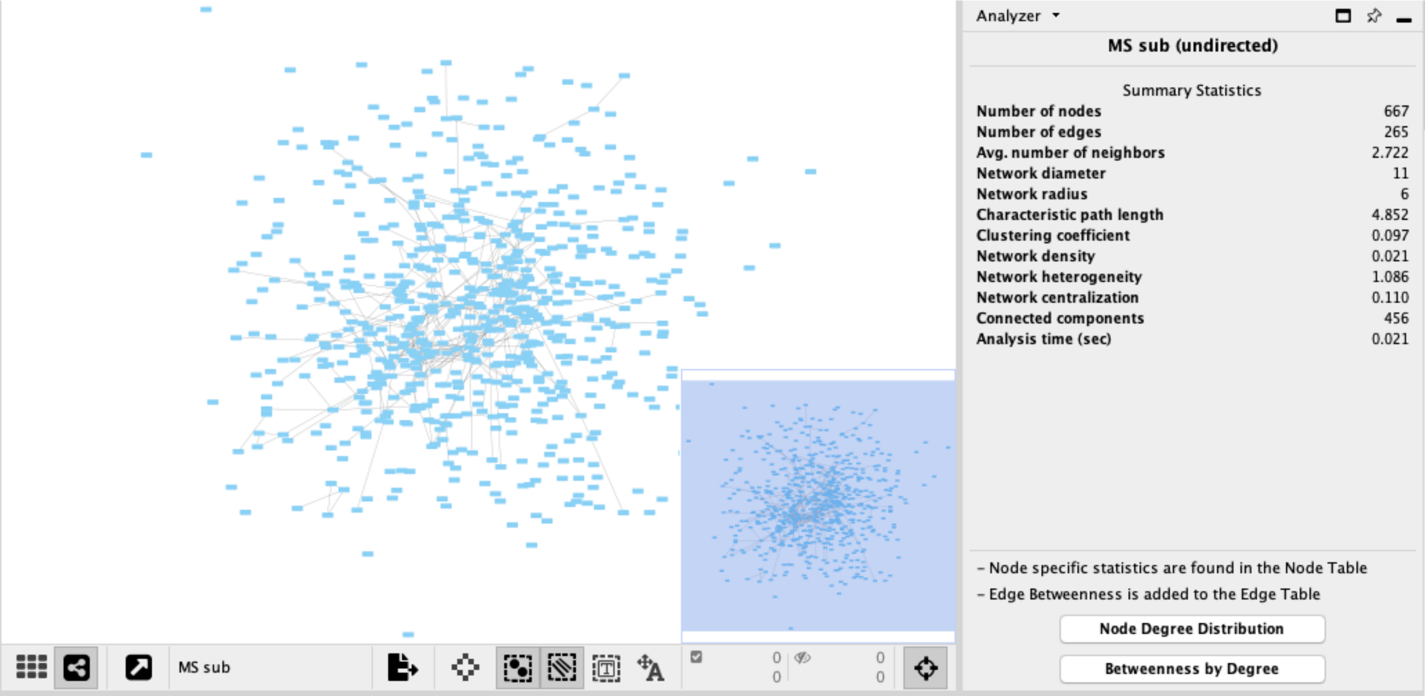


parent_PPI.sif(1) (undirected)

```
Summary Statistics
Number of nodes                     346
Number of edges                      40
Avg. number of neighbors          1.714
Network diameter                      4
Network radius                        2
Characteristic path length        2.190
Clustering coefficient            0.000
Network density                   0.286
Network heterogeneity             0.601
Network centralization            0.533
Connected components                307
Analysis time (sec)               0.016
```

- Node specific statistics are found in the Node Table
- Edge Betweenness is added to the Edge Table

The MS first order network contains 667 nodes and 265 edges. In contrast to the HT subnetwork, this network has metrics reflecting it is fairly connected and may contain more hub genes (via network heterogeneity).



MS sub (undirected)

Summary Statistics

Number of nodes	667
Number of edges	265
Avg. number of neighbors	2.722
Network diameter	11
Network radius	6
Characteristic path length	4.852
Clustering coefficient	0.097
Network density	0.021
Network heterogeneity	1.086
Network centralization	0.110
Connected components	456
Analysis time (sec)	0.021

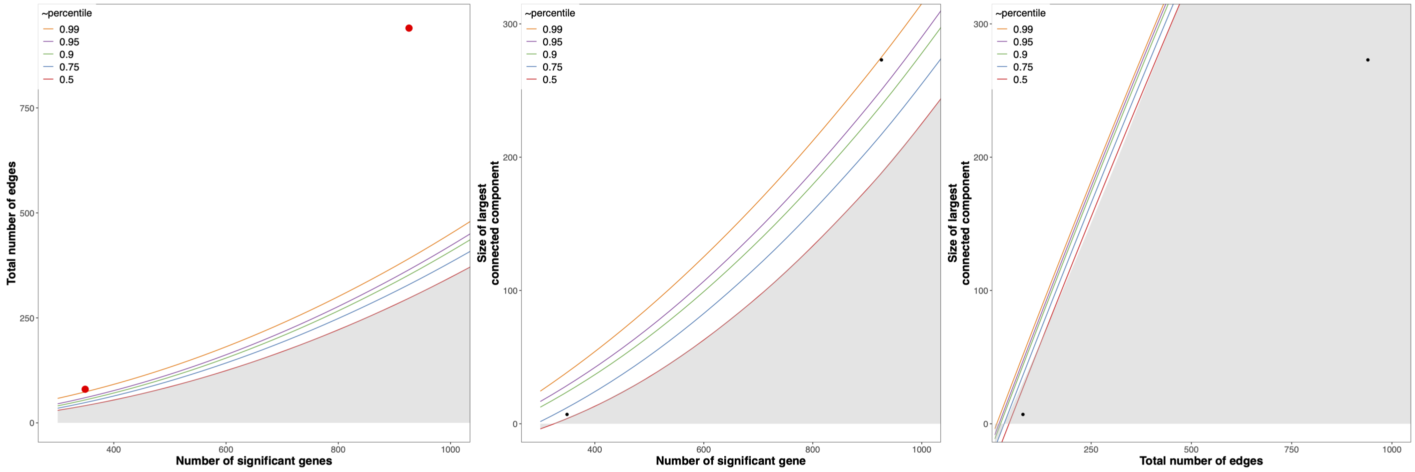
– Node specific statistics are found in the Node Table

– Edge Betweenness is added to the Edge Table

Q4: Source “Pathway_permutation.r”. Are the first order networks from both GWAS more connected than expected? What does this mean?

ANS: The sub-network of MS is more connected than expected (way above the 99th percentile) The sub-network of HT is not.

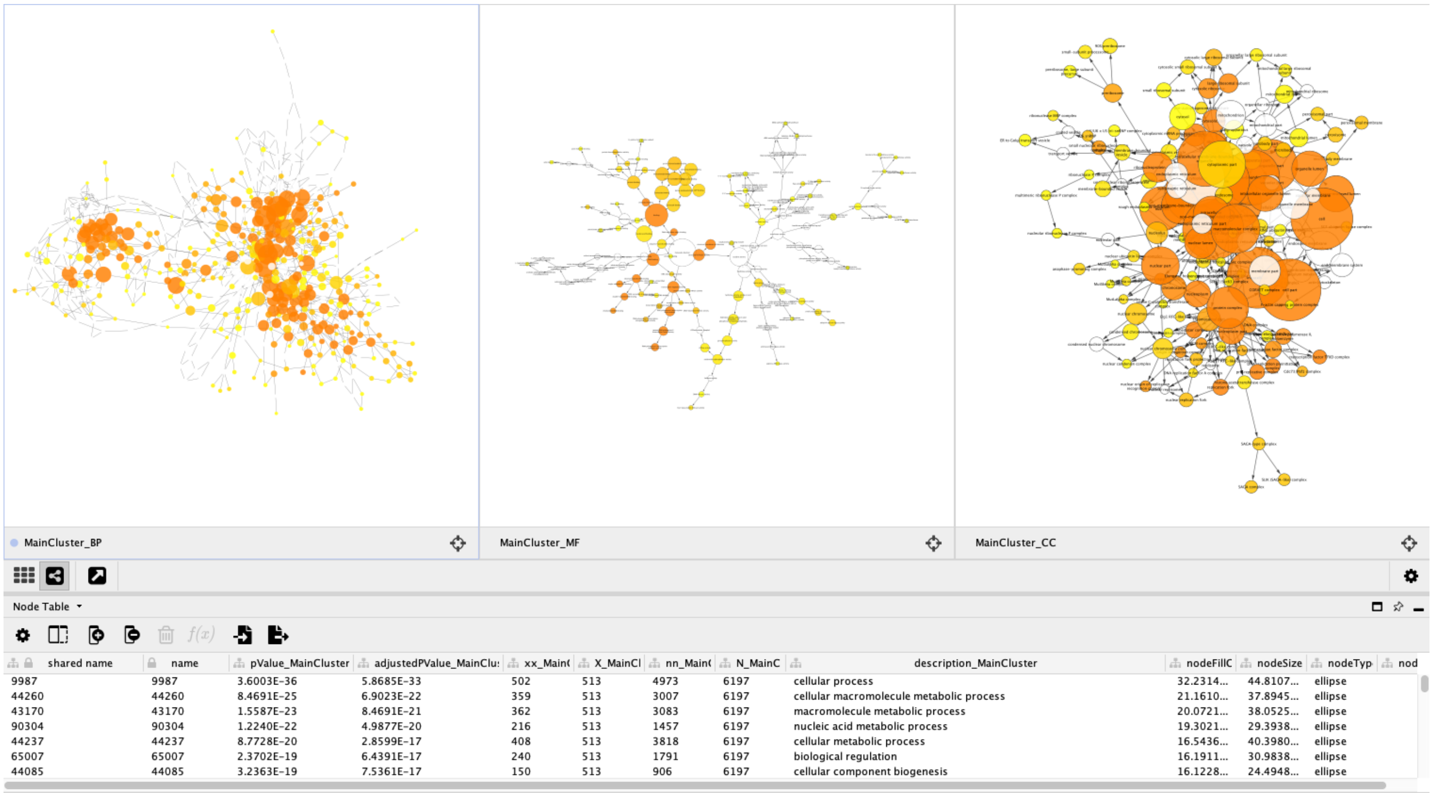
```
source("Pathway_permutation.r")

  Region Extracted_nodes Edges largest_nodes
1     HT             349    80             7
2     MS             926   940           273
```



Q5: Run BINGO App on all nodes from largest connected component. What biological processes emerge from the first order networks?

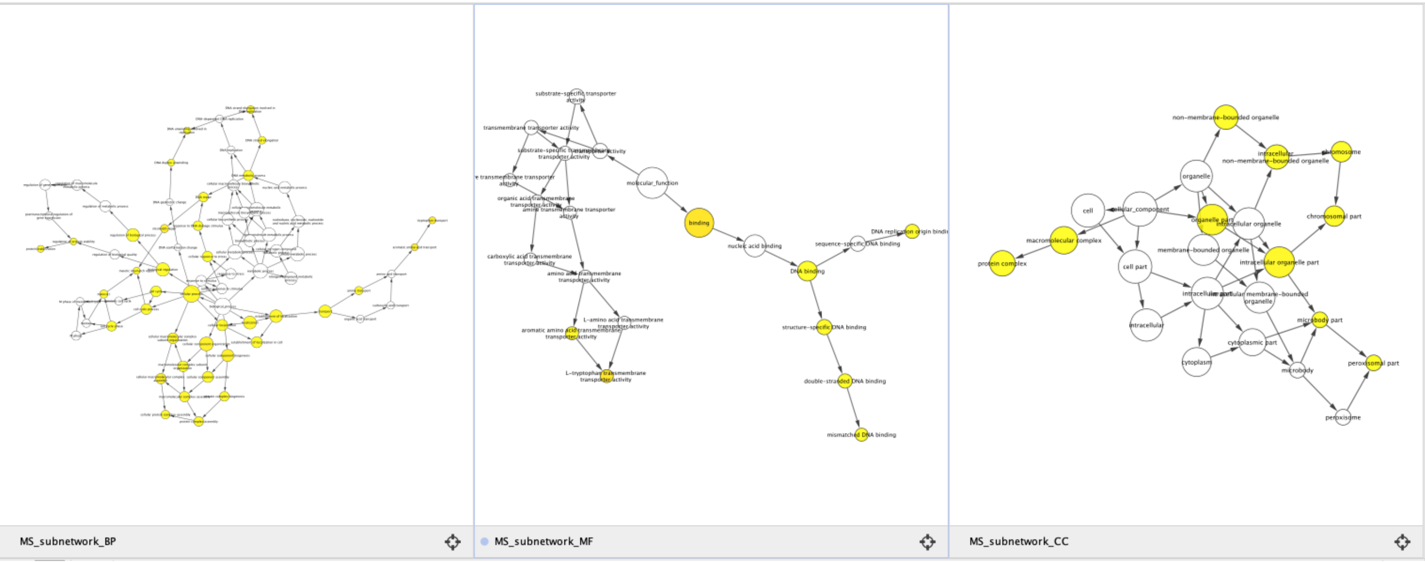
ANS: There should be several immune-related GO significantly enriched in the MS network. Not much (if anything) in the HT net.

ANS: In the total PPI network the top overrepresented GO terms are: cellular process (BP), cellular macromolecule metabolic process (BP), general RNA polymerase II transcription factor activity (MF), macromolecular complex (CC), and protein complex (CC). Overall not particularly informative.



shared name	name	pValue_MainCluster	adjustedPValue_MainClu:	xx_Main(	X_MainCl	nn_Main(	N_MainC	description_MainCluster	nodeFillC	nodeSize	nodeTyp	nod
9987	9987	3.6003E-36	5.8685E-33	502	513	4973	6197	cellular process	32.2314...	44.8107...	ellipse	
44260	44260	8.4691E-25	6.9023E-22	359	513	3007	6197	cellular macromolecule metabolic process	21.1610...	37.8945...	ellipse	
43170	43170	1.5587E-23	8.4691E-21	362	513	3083	6197	macromolecule metabolic process	20.0721...	38.0525...	ellipse	
90304	90304	1.2240E-22	4.9877E-20	216	513	1457	6197	nucleic acid metabolic process	19.3021...	29.3938...	ellipse	
44237	44237	8.7728E-20	2.8599E-17	408	513	3818	6197	cellular metabolic process	16.5436...	40.3980...	ellipse	
65007	65007	2.3702E-19	6.4391E-17	240	513	1791	6197	biological regulation	16.1911...	30.9838...	ellipse	
44085	44085	3.2363E-19	7.5361E-17	150	513	906	6197	cellular component biogenesis	16.1228...	24.4948...	ellipse	

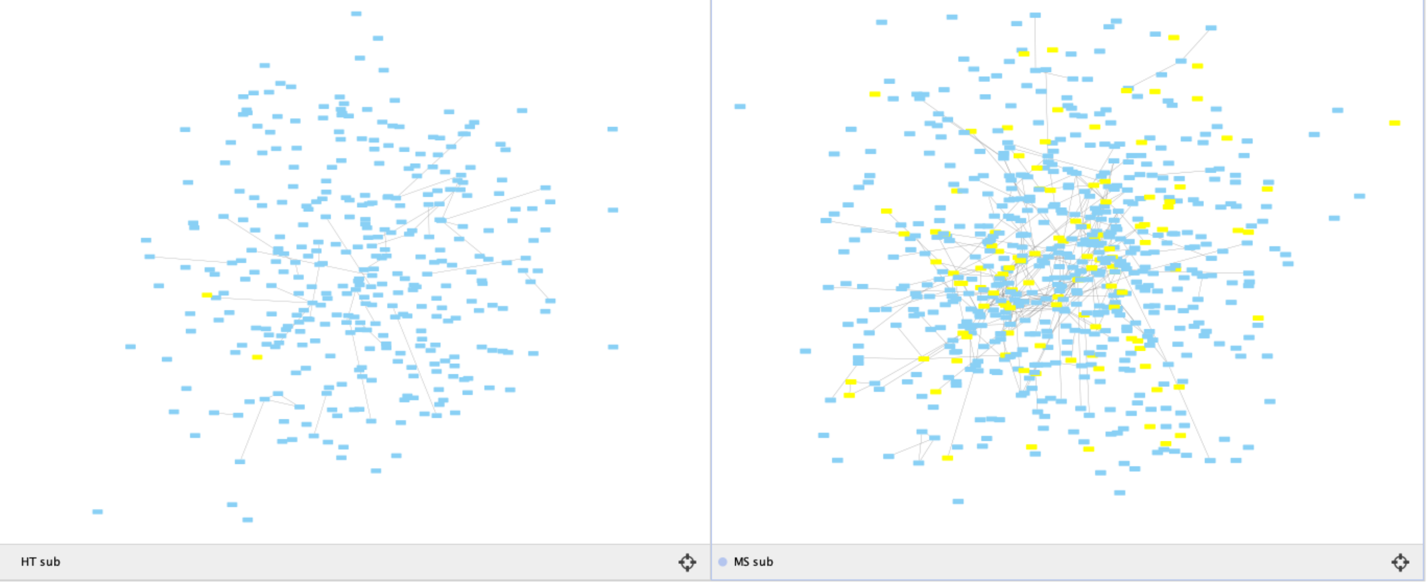
The first order network for the MS dataset shows enrichment for mismatch repair (BP), binding (MF), organelle part (CC) and quite a few terms about aromatic or tryptophan activity but this is mainly driven by the two genes “TAP2|TAP1”.


substrate-specific transporter activity
transmembrane transporter activity
substrate-specific transmembrane transporter activity
transporter activity
transmembrane transporter activity
organic acid transmembrane transporter activity
amine transmembrane transporter activity
molecular_function
binding
carboxylic acid transmembrane transporter activity
amino acid transmembrane transporter activity
L-amino acid transmembrane transporter activity
aromatic amino acid transmembrane transporter activity
L-tryptophan transmembrane transporter activity
nucleic acid binding
DNA replication origin bindi
sequence-specific DNA binding
DNA binding
structure-specific DNA binding
double-stranded DNA binding
mismatched DNA binding
non-membrane-bounded organelle
intracellular non-membrane-bounded organelle
chromosome
organelle
cell
cellular_component
organelle part
intracellular organelle
chromosomal part
macromolecular complex
protein complex
membrane-bounded organelle
intracellular organelle part
cell part
intracellular part
intracellular membrane-bounded organelle
intracellular
microbody part
cytoplasmic part
cytoplasm
microbody
peroxisomal part
peroxisome
MS_subnetwork_BP
MS_subnetwork_MF
MS_subnetwork_CC

The first order network for the HT dataset returned no significantly enriched GO terms.

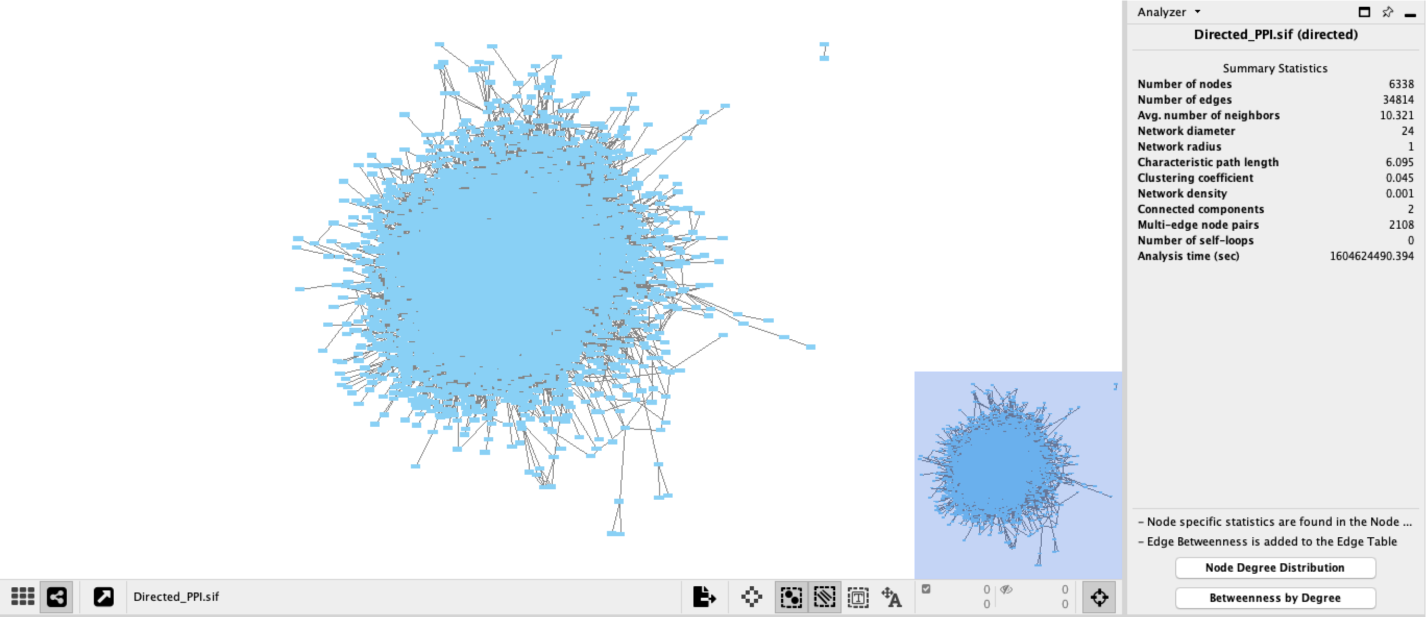
Q6: Map and color known MS and HT genes onto their respective first order nets. Interpret results.

ANS: The MS net contains more known genes. Perhaps the HT GWAS has a high false negative rate due to low power.



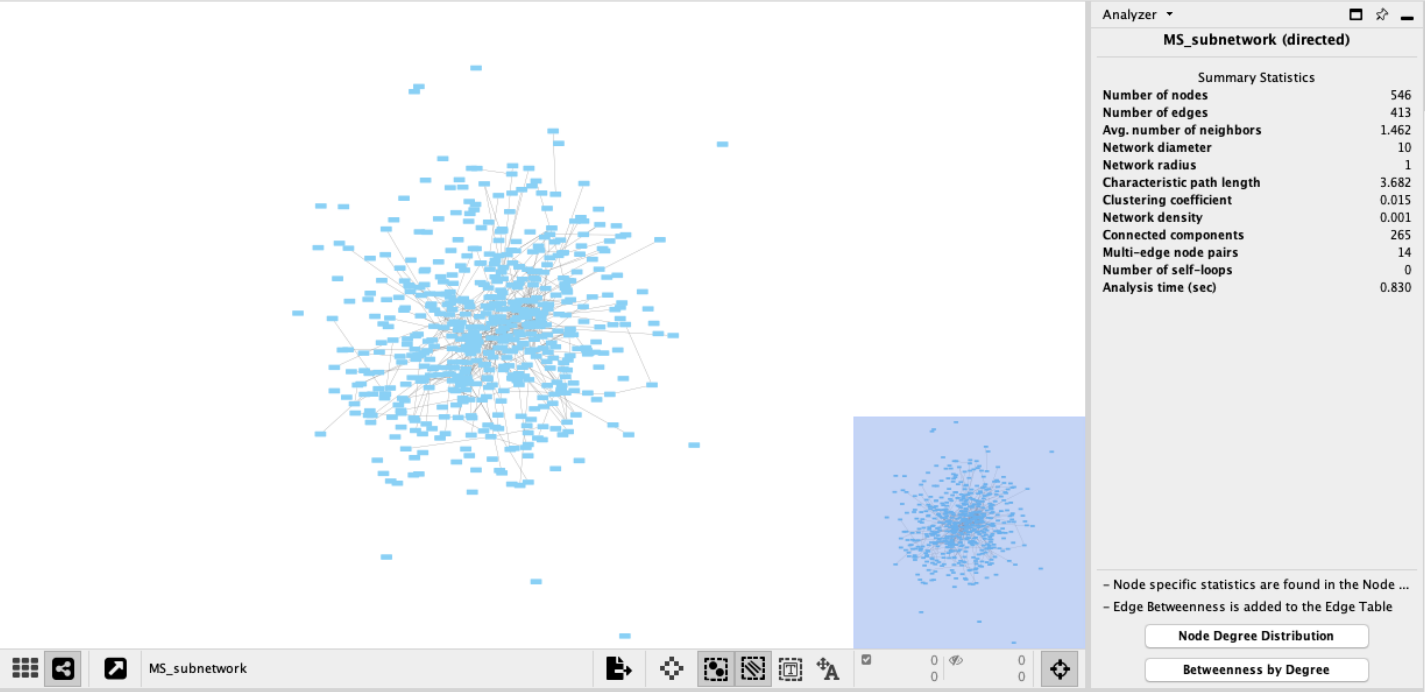
Q7: Repeat steps 3 and 4 with directed protein network from PNAS paper.

ANS: Compared to the HT first order network, the MS first order network is much more connected with a large connected component that exceeds what is expected by chance using a permutation null distribution.



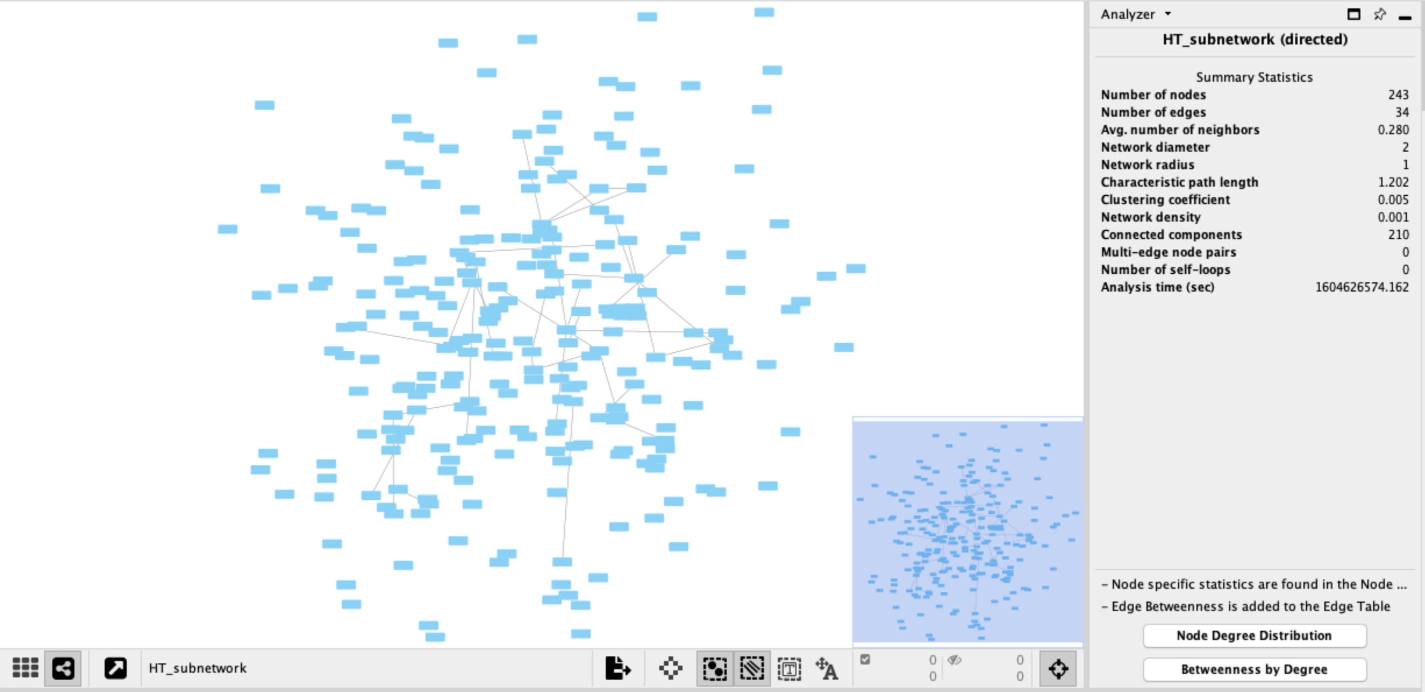
```
Directed_PPI.sif (directed)

Summary Statistics
Number of nodes                         6338
Number of edges                        34814
Avg. number of neighbors              10.321
Network diameter                          24
Network radius                             1
Characteristic path length             6.095
Clustering coefficient                 0.045
Network density                        0.001
Connected components                       2
Multi-edge node pairs                   2108
Number of self-loops                       0
Analysis time (sec)           1604624490.394

- Node specific statistics are found in the Node ...
- Edge Betweenness is added to the Edge Table
```

The MS first order network contains 546 nodes and 413 edges.



```
MS_subnetwork (directed)

Summary Statistics
Number of nodes                  546
Number of edges                  413
Avg. number of neighbors       1.462
Network diameter                  10
Network radius                     1
Characteristic path length     3.682
Clustering coefficient         0.015
Network density                0.001
Connected components             265
Multi-edge node pairs             14
Number of self-loops               0
Analysis time (sec)            0.830

- Node specific statistics are found in the Node ...
- Edge Betweenness is added to the Edge Table
```

The HT first order network contains 243 nodes and 34 edges. This is much fewer edges than expected leaving quite a few orphan nodes. This is reflected in the large reduction in the average number of neighbors (0.28)



HT_subnetwork (directed)

Summary Statistics

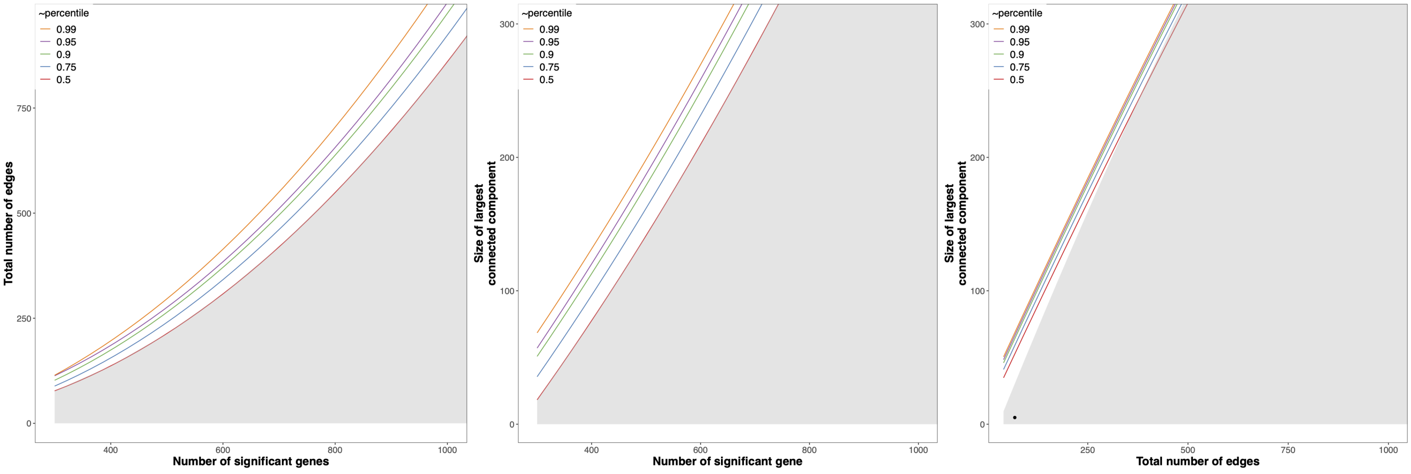
Number of nodes	243
Number of edges	34
Avg. number of neighbors	0.280
Network diameter	2
Network radius	1
Characteristic path length	1.202
Clustering coefficient	0.005
Network density	0.001
Connected components	210
Multi-edge node pairs	0
Number of self-loops	0
Analysis time (sec)	1604626574.162

- Node specific statistics are found in the Node ...
- Edge Betweenness is added to the Edge Table

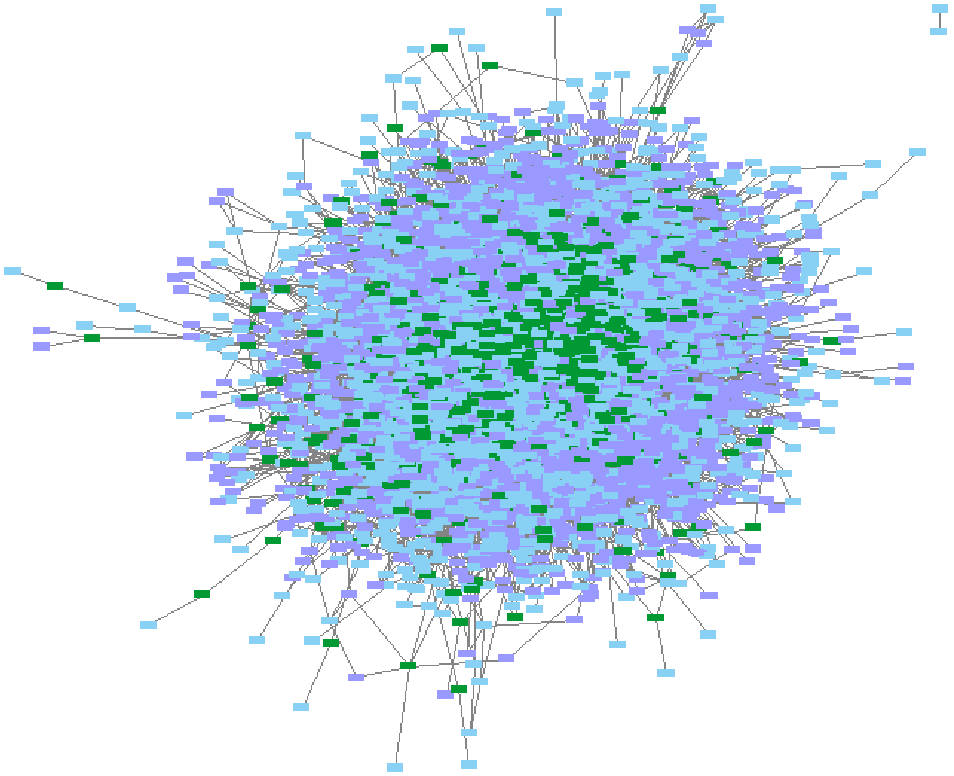
Compared to the undirected network, all numbers for the HT network are reduced. Possibly because the lack of connectivity. For the MS-directed first order network, we see an increase in the number of edges and largest node.

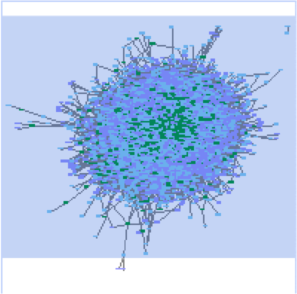
```
source("Pathway_permutation.r")

UNDIRECTED
  Region Extracted_nodes Edges largest_nodes
1     HT             349    80             7
2     MS             926   940           273

VS

DIRECTED
  Region Extracted_nodes Edges largest_nodes
1     HT             243    68             5
2     MS             742  1354           379
```

Similar to before we see that the MS-directed first order network is more connected than our permuted distribution whereas the HT falls well within the bounds.



Q8: Color nodes by controllability category (dispensable, indispensable, neutral).

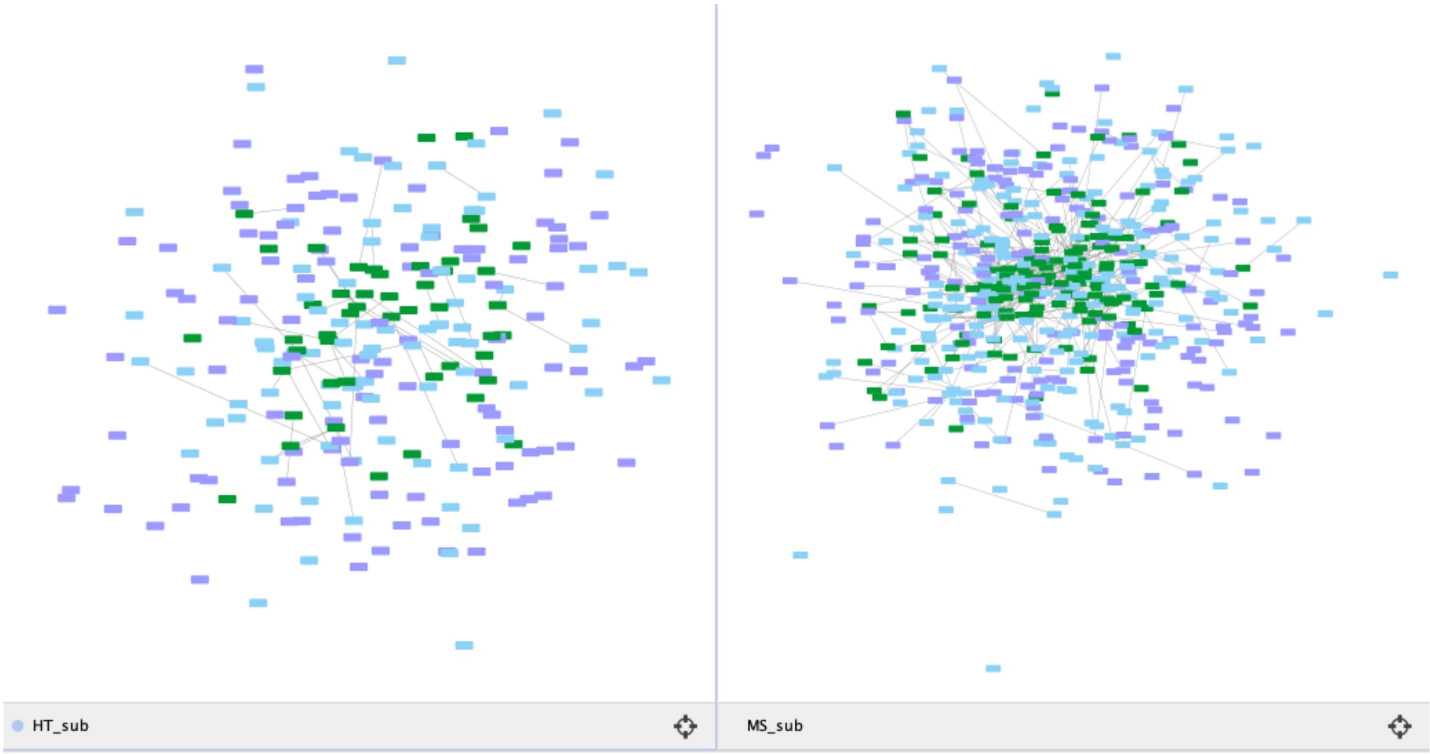




Q9: Repeat step 6. Are MS-associated genes more enriched in any controllability category? Interpret.

ANS: There is no significant enrichment of MS (or HT) associated genes in any controllability category. This is in line with Fig. 2B in the controllability paper (PMID: 27091990).

further: For the nodes in the directed MS network with a pval < 0.05, 229/546 are Neutral, 175/546 are Dispensable, and 142/546 are Indispensable. Via hypergeometric test comparing the amount of total proteins in each controlability group versus the MS subnetwork, we see there is enrichment for both “Indispensable” and “Neutral” proteins but not “Dispensable”. One way this could be biologically explained is if the MS GWAS picked up on highly critical genes (as expected) as well as their more neutral interactors. Dispensable genes were not picked up through the GWAS as these may be more likely to vary in the “control” population as well.



```
#phyper(neutral in subnet, subnet_size, total_size-subnet size, total neutral genes)
#Note: I removed the 8 genes with missing controllability categories

#Test for over-representation
phyper(229, 546, 6330-546, 2261, lower.tail = F) #Neutral
```

```
## [1] 0.0007079792
```

```
phyper(175, 546, 6330-546, 2343, lower.tail = F) #Dispensable
```

```
## [1] 0.9935267
```

```
phyper(142, 546, 6330-546, 1326, lower.tail = F) #Indispensable
```

```
## [1] 0.001234459
```

```
knitr::knit_exit()
```