BCB420 - Computational Systems Biology

Lecture 11 - Enrichment Map and other Cytoscape Apps cont'd

Ruth Isserlin

2020-03-22

Assignment #3

- Data set Pathway and Network Analysis
- Due April 3, 2020! @ 13:00

What to hand in?

- **html rendered RNotebook** - you should submit this through quercus
- Make sure the notebook and all associated code is checked into your github repo as I will be pulling all the repos at the deadline and using them to compile your code. - Your checked in code must replicate the handed in notebook.
- Document your work and your code directly in the notebook.
- **Reference the paper associated with your data!**
- **Introduce your paper and your data again**
- You are allowed to use helper functions or methods but make sure when you source those files the paths to them are relative and that they are checked into your repo as well.

Outline for Today's lecture

- review of Enrichment map
- looking at Pathways in depth - Reactome app, Pathway commons, String and GeneMania
- **Post analysis**
- **Enrichment Map Dark Matter**

Post/signature analysis

Drugs

DRUGBANK

Regulators

TargetScanHuman
Prediction of microRNA targets
Release 7.2: March 2018
Agarwal et al., 2015

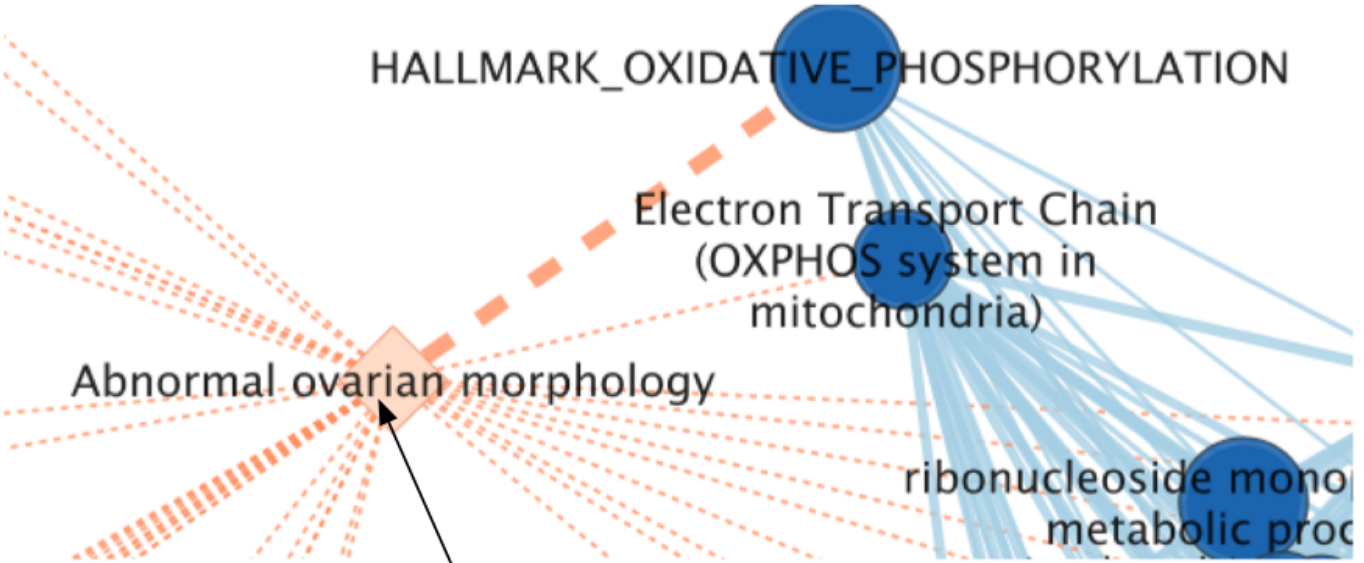
TRANSFAC
database

Disease Genes/ signatures

human phenotype ontology

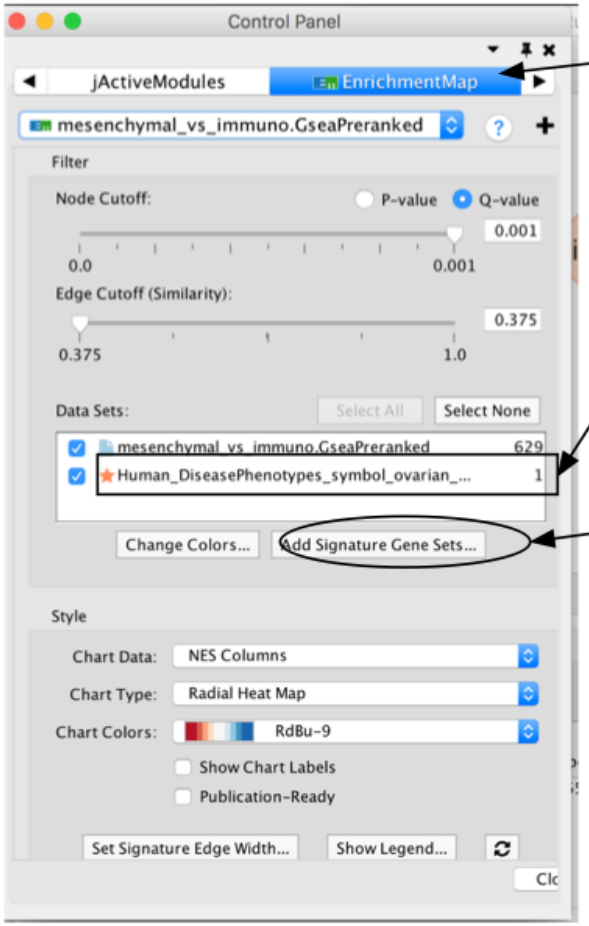
MSigDB
Molecular Signatures Database

Post/signature analysis



Adds signature node to network
- different shape and colour
- implemented using cytoscape bypass (for some of the attributes)

How to add a signature set



In the Enrichment Map Control panel

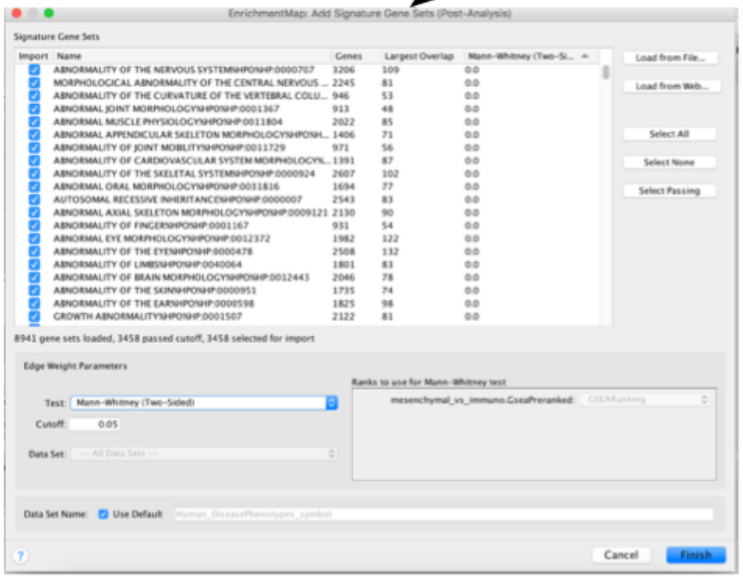
Once you have created a signature set it will appear in the Data set Panel

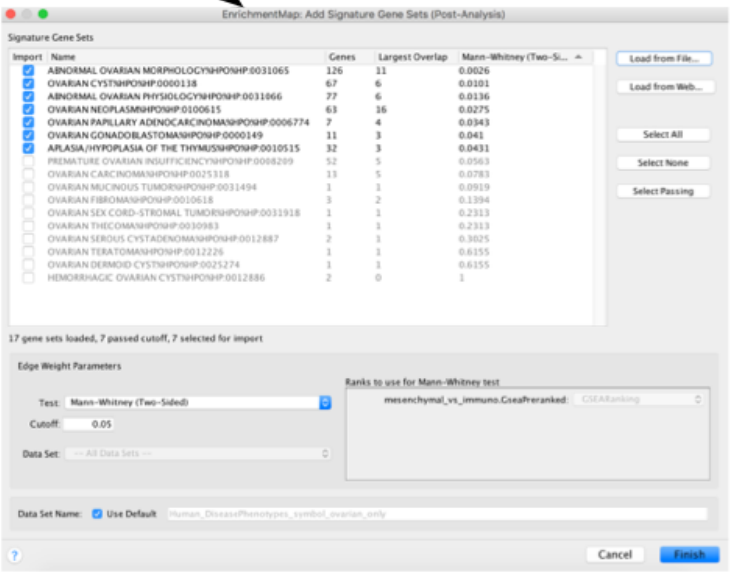
To Add signature gene sets

Post/signature analysis

Exploratory

Known signatures

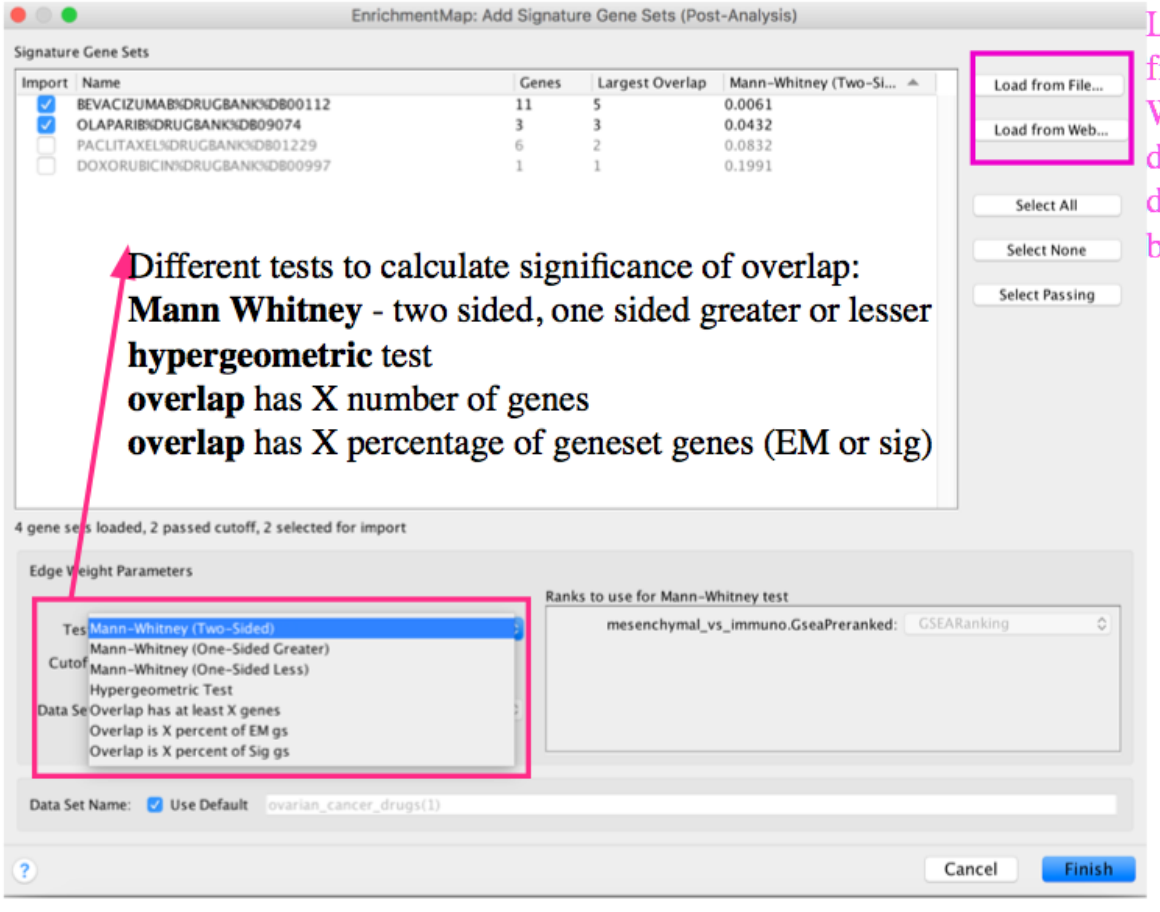




Using the entire Human Phenotype Database
* What phenotypes have significant overlap with my expression data?
* Depedning on the size of network and the the size of the signature set can take a long time

Using the **subset** Human Phenotype Database
Where do specific signatures overlap with our data?
Will be relatively fast to compute.

Signature Set Metrics



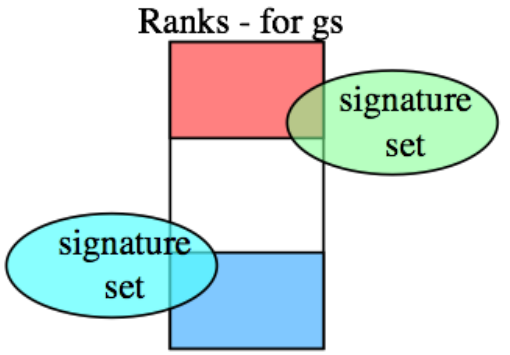
Load from file or from web.
Web pulls gmt files directrly from download.baderlab.org but files are large!

Mann Whitney

many different names - wilcoxon rank sum test, mann whitney u test

test can be run multiple ways - lesser, greater, either

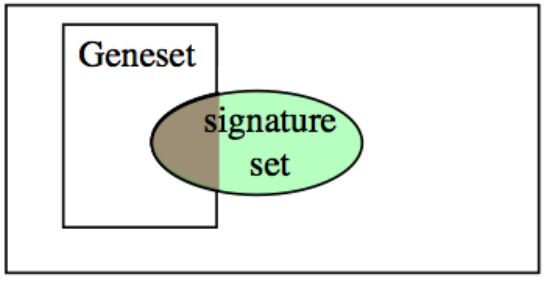
Are the genes in the signature set found mostly at the bottom of the rank list, mostly at the top of the rank list or simply ranked highly (irrespective of direction?



Hypergeometric

Ranks are irrevelant

Is there a significant overlap size

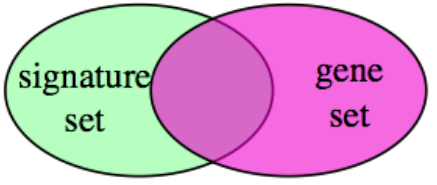


Overlap

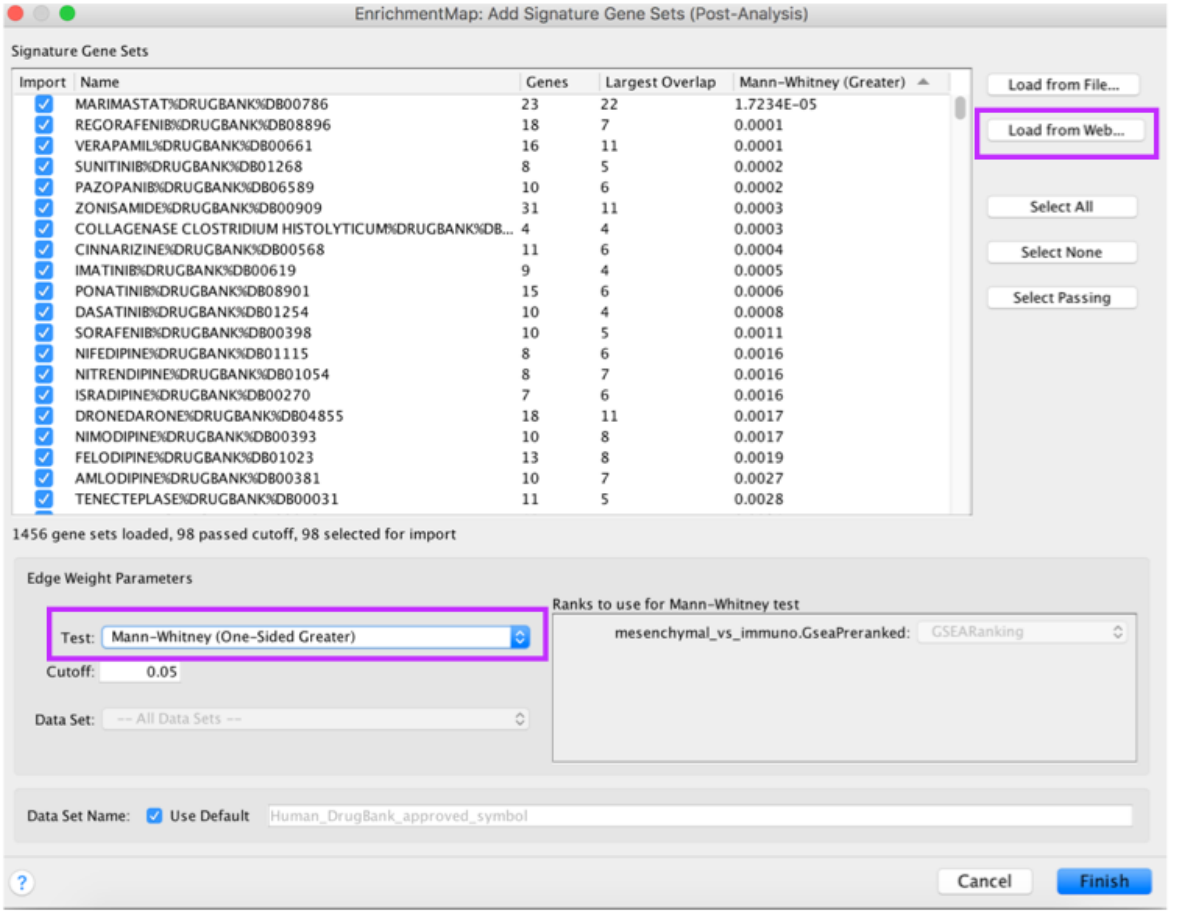
number of genes in overlap

percentage of genes in overlap as compared to the signature set

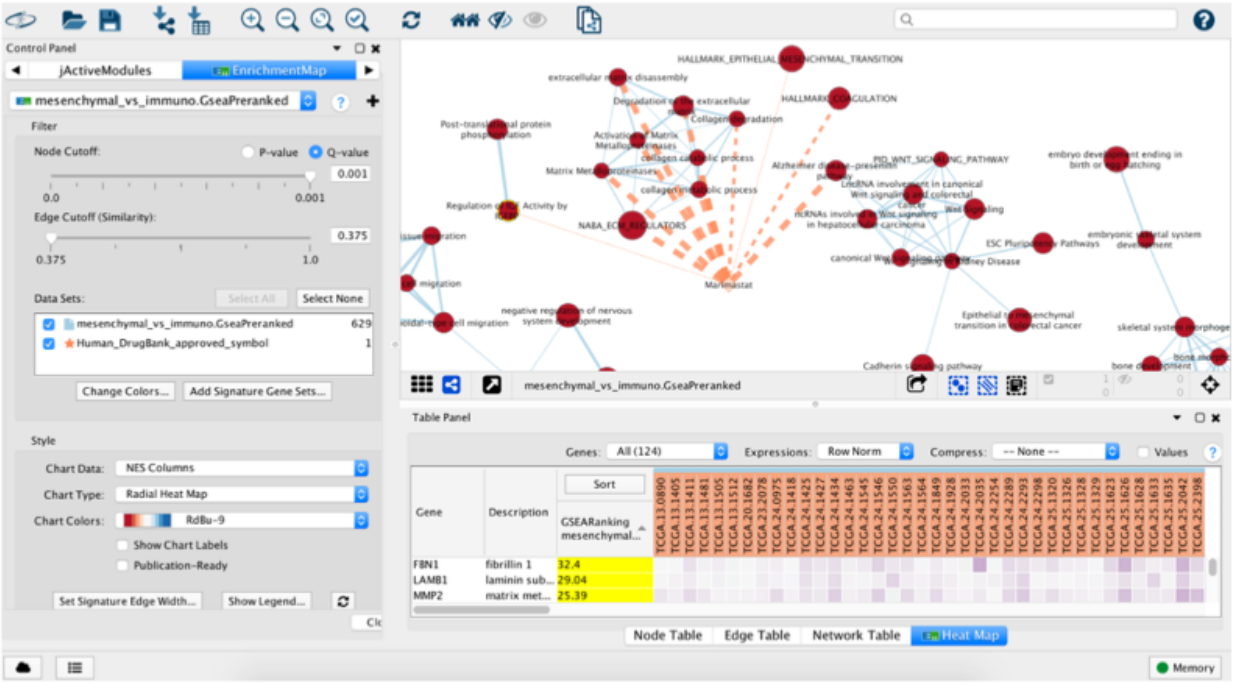
percentage of genes in overlap compared to the gene set



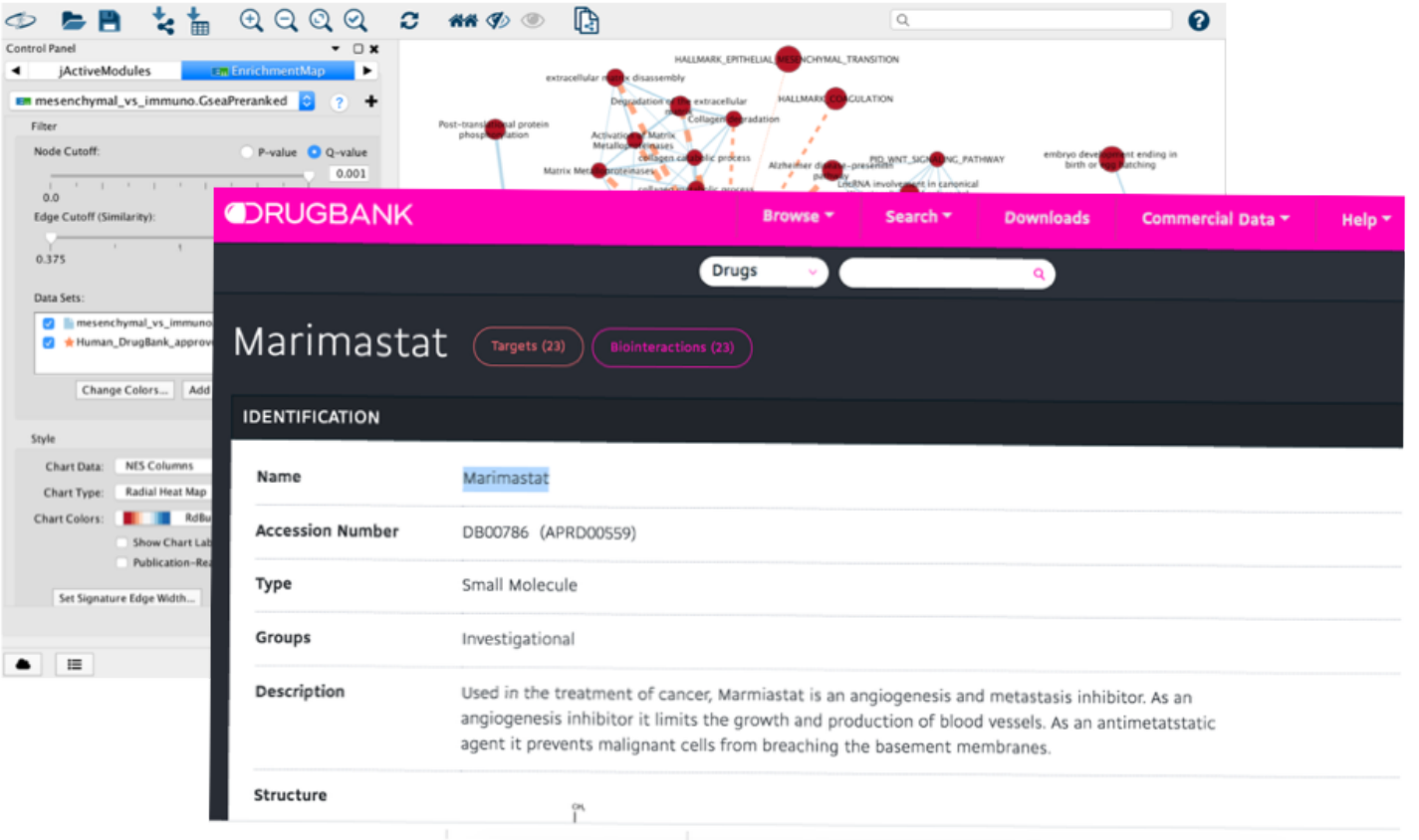
Example Post Analysis - Drugs that can target Mesenchymal



Example Post Analysis - Drugs that can target Mesenchymal

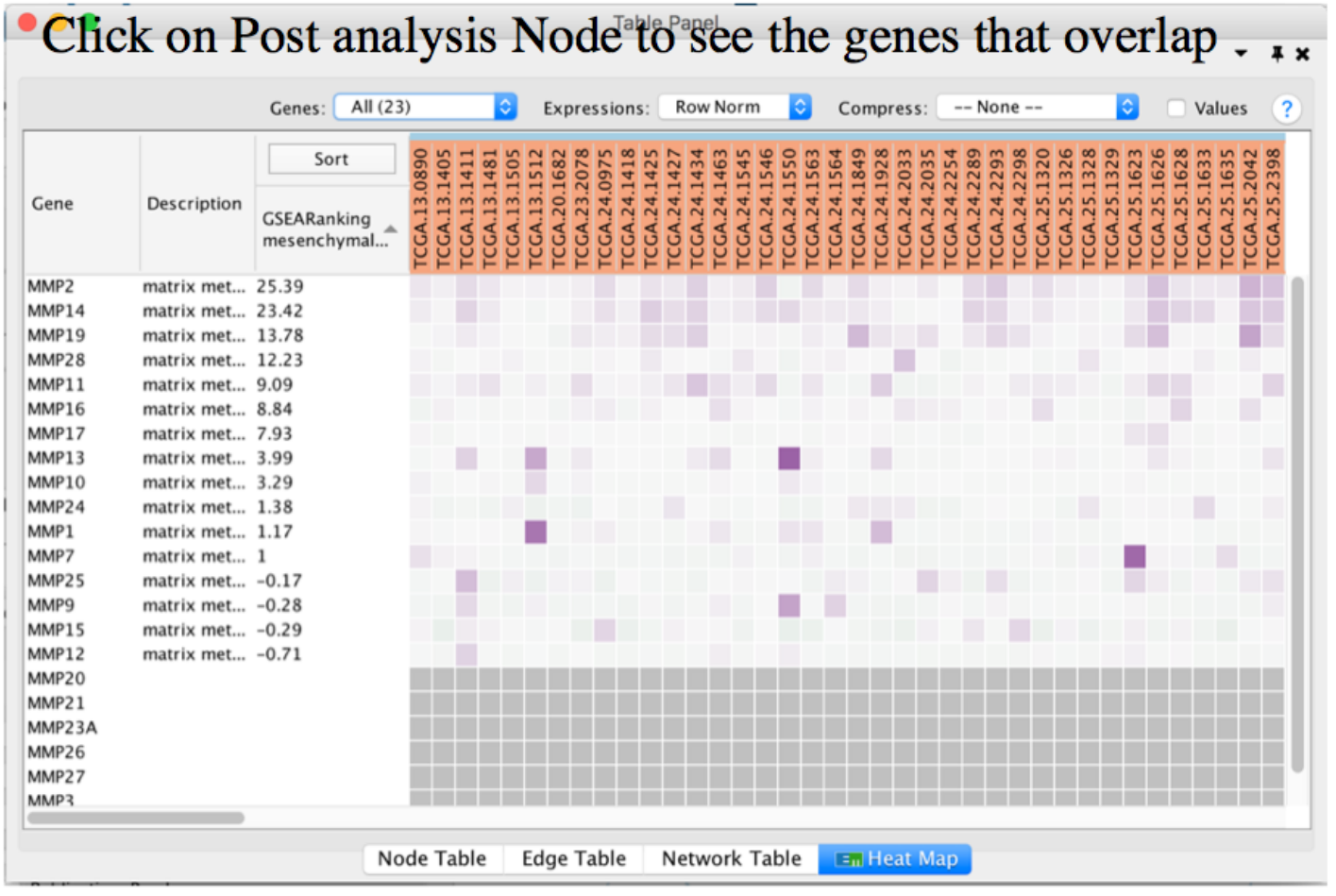


Example Post Analysis - Drugs that can target Mesenchymal



Example Post Analysis - Drugs that can target Mesenchymal

Click on Post analysis Node to see the genes that overlap

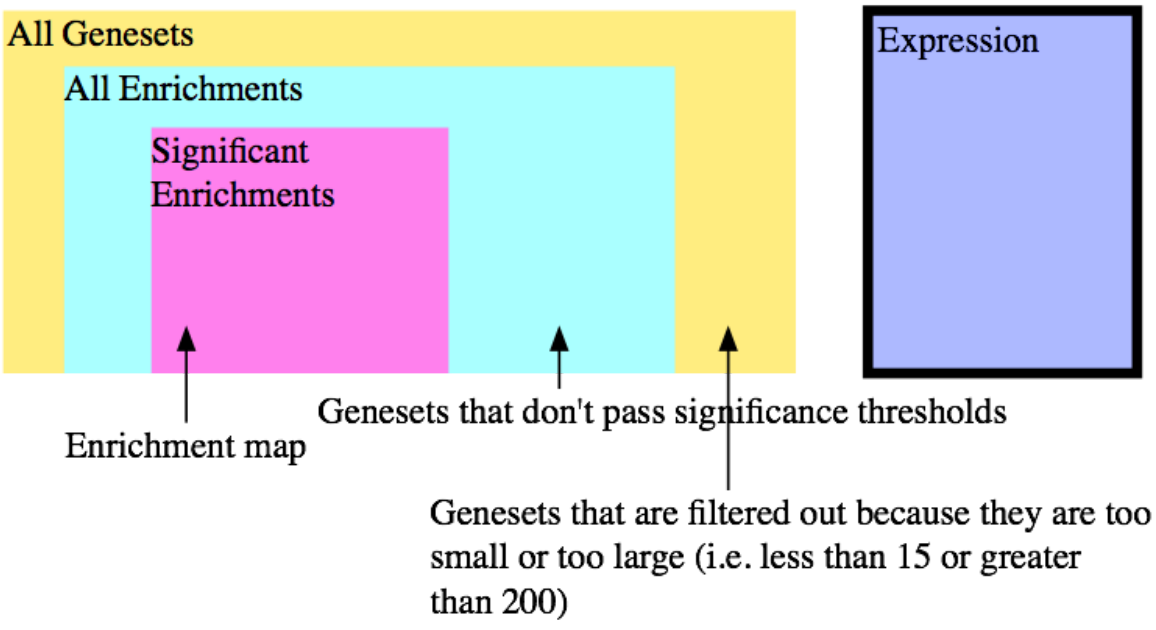


Post Analysis Summary

Dark Matter

What is dark matter?

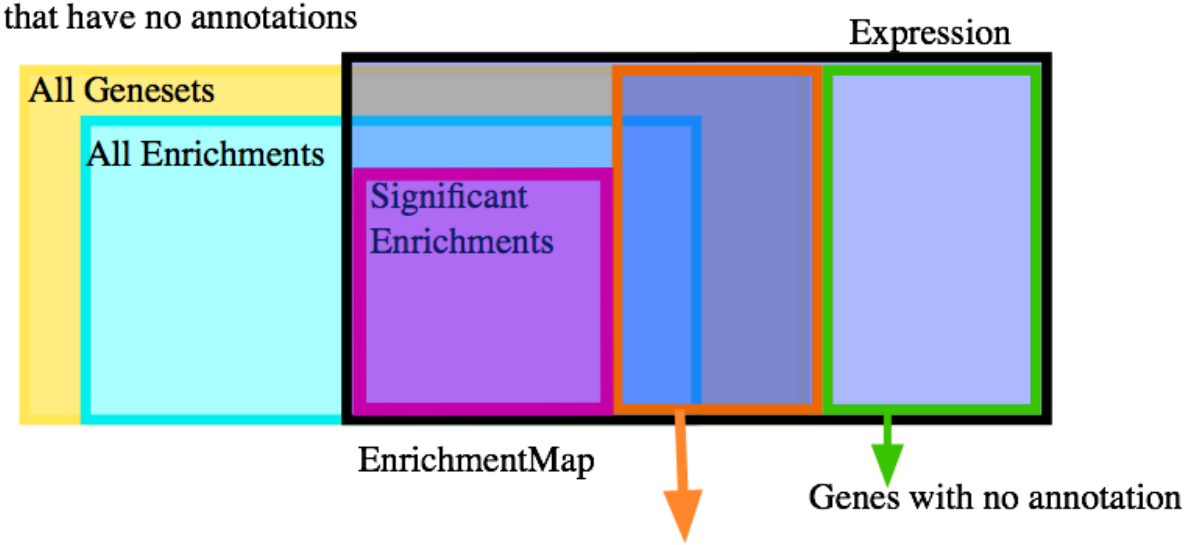
Sometimes the genes we don't see in our pathway results are just as important as the genes that we do see.



Dark Matter

Different types of dark matter:

1. Genes that are annotated but the functions they are annotated to get filtered out because they are either too large or too small
2. Genes that have no annotations



Dark Matter

Files needed in order to conduct a dark matter analysis:

1. Definitions of the genesets used in the analysis - gmt file.

```
library(GSA)
gmt_file <- file.path(getwd(),"data",

"Human_GOBP_AllPathways_no_GO_iea_February_01_2020_symbol.gmt")

capture.output(genesets<-
GSA.read.gmt(gmt_file),file="gsa_load.out")

names(genesets$genesets) <- genesets$geneset.names
```

Dark Matter

Files needed in order to conduct a dark matter analysis:

1. Definitions of the genesets used in the analysis - gmt file.
2. Expression file + rank file

```
expression <- read.table(file.path(getwd(),"data",

"Supplementary_Table6_TCGA_OV_RNAseq_expression.txt"),
                         header = TRUE, sep = "\t", quote="\"",
                         stringsAsFactors = FALSE)
ranks <-
read.table(file.path(getwd(),"data","Supplementary_Table2_MesenvsImmu

                     header = TRUE, sep = "\t", quote="\"",
                     stringsAsFactors = FALSE)
```

Dark Matter

Files needed in order to conduct a dark matter analysis:

1. Definitions of the genesets used in the analysis - gmt file.
2. Expression file
3. GSEA results files - the na_pos and na_neg spreadsheets in GSEA results directories

```
#get all the GSEA directories
gsea_directories <- list.files(path = file.path(getwd(),"data"),
                               pattern = "\\.GseaPreranked")
if(length(gsea_directories) == 1){
  gsea_dir <- file.path(getwd(),"data",gsea_directories[1])
  #get the gsea result files
  gsea_results_files <- list.files(path = gsea_dir,
                               pattern = "gsea_report_*.*.xls")
  #there should be 2 gsea results files
  enr_file1 <-
read.table(file.path(gsea_dir,gsea_results_files[1]),
                         header = TRUE, sep = "\t", quote="\"",
                         stringsAsFactors = FALSE,row.names=1)
  enr_file2 <-
read.table(file.path(gsea_dir,gsea_results_files[1]),
                         header = TRUE, sep = "\t", quote="\"",
                         stringsAsFactors = FALSE,row.names=1)
}
```

Dark Matter

Collect the Data we need to calculate the dark matter from the above files:

1. all genes in the expression set - already loaded above
2. all genes in the enrichment results

```
#get the genes from the set of enriched pathwasy (no matter what
threshold)
all_enr_genesets<- c(rownames(enr_file1), rownames(enr_file2))
genes_enr_gs <- c()
for(i in 1:length(all_enr_genesets)){
  current_geneset <-
unlist(genesets$genesets[which(genesets$geneset.names %in%
all_enr_genesets[i])])
  genes_enr_gs <- union(genes_enr_gs, current_geneset)
}
```

Dark Matter

Data we need to calculate the dark matter:

1. all genes in the expression set - row names of the expression matrix
2. all genes in the enrichment results
3. all genes in the **significant enrichment results** - define your thresholds

```
FDR_threshold <- 0.001
#get the genes from the set of enriched pathwasy (no matter what
threshold)
all_sig_enr_genesets<- c(rownames(enr_file1)
[which(enr_file1[,"FDR.q.val"]<=FDR_threshold)],
rownames(enr_file2)[which(enr_file2[,"FDR.q.val"]<=FDR_threshold)])
genes_sig_enr_gs <- c()
for(i in 1:length(all_sig_enr_genesets)){
  current_geneset <-
unlist(genesets$genesets[which(genesets$geneset.names %in%
all_sig_enr_genesets[i])])
  genes_sig_enr_gs <- union(genes_sig_enr_gs, current_geneset)
}
```

Dark Matter

Data we need to calculate the dark matter:

1. all genes in the expression set - row names of the expression matrix
2. all genes in the enrichment results
3. all genes in the significant enrichment results - define your thresholds
4. all genes in geneset file

```
genes_all_gs <- unique(unlist(genesets$genesets))
```

Dark Matter

Data we need to calculate the dark matter:

1. all genes in the expression set - row names of the expression matrix - There are 15196 unique genes in the expression file.
2. all genes in the enrichment results - There are 11267 unique genes in the enrichment results.
3. all genes in the significant enrichment results - There are 4773 unique genes in the enrichment results.
4. all genes in geneset file - There are 16475 unique genes in the geneset file.

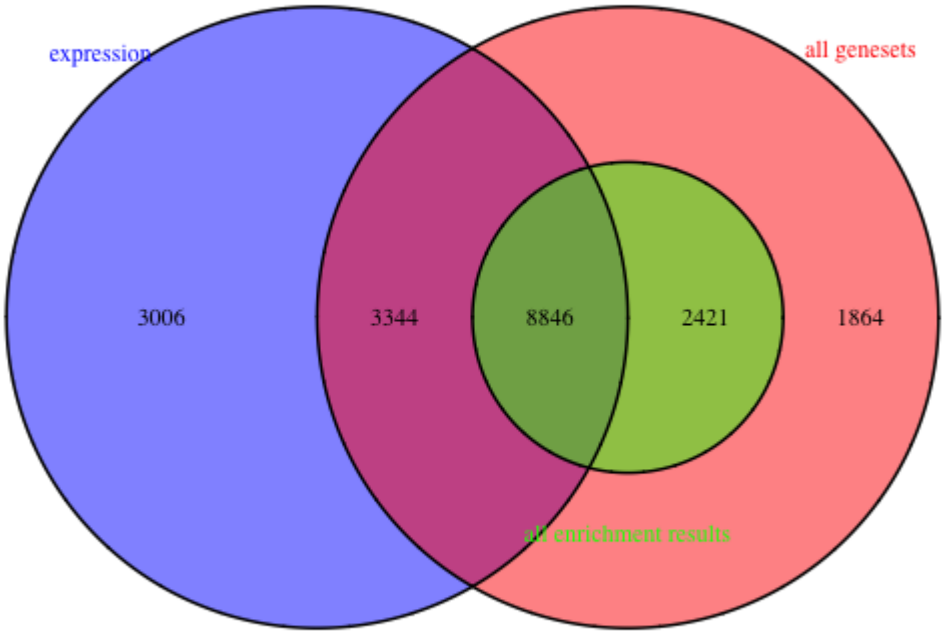
Venn Diagram of Dark Matter Overlaps

```
library(VennDiagram)

A <- genes_all_gs
B <- genes_enr_gs
C <- expression[,1]
png(file.path(getwd(),"data","dark_matter_overlaps.png"))
draw.triple.venn( area1=length(A), area2=length(B), area3 =
length(C),
                  n12 = length(intersect(A,B)),
n13=length(intersect(A,C)),
                  n23 = length(intersect(B,C)),
                  n123 = length(intersect(A,intersect(B,C))),
                  category = c("all genesets","all enrichment
results","expression"),
                  fill = c("red","green","blue"),
                  cat.col = c("red","green","blue")
)
```

```
## (polygon[GRID.polygon.1], polygon[GRID.polygon.2],
polygon[GRID.polygon.3], polygon[GRID.polygon.4],
polygon[GRID.polygon.5], polygon[GRID.polygon.6], text[GRID.text.7],
text[GRID.text.8], text[GRID.text.9], text[GRID.text.10],
```

Dark matter - overlaps



Dark Matter

Get the set of genes that have no annotation

```
genes_no_annotation <- setdiff(expression[,1], genes_all_gs)
```

Get the top ranked genes that have no annotation

```
ranked_gene_no_annotation <- ranks[which(ranks[,1] %in%
genes_no_annotation),]
```

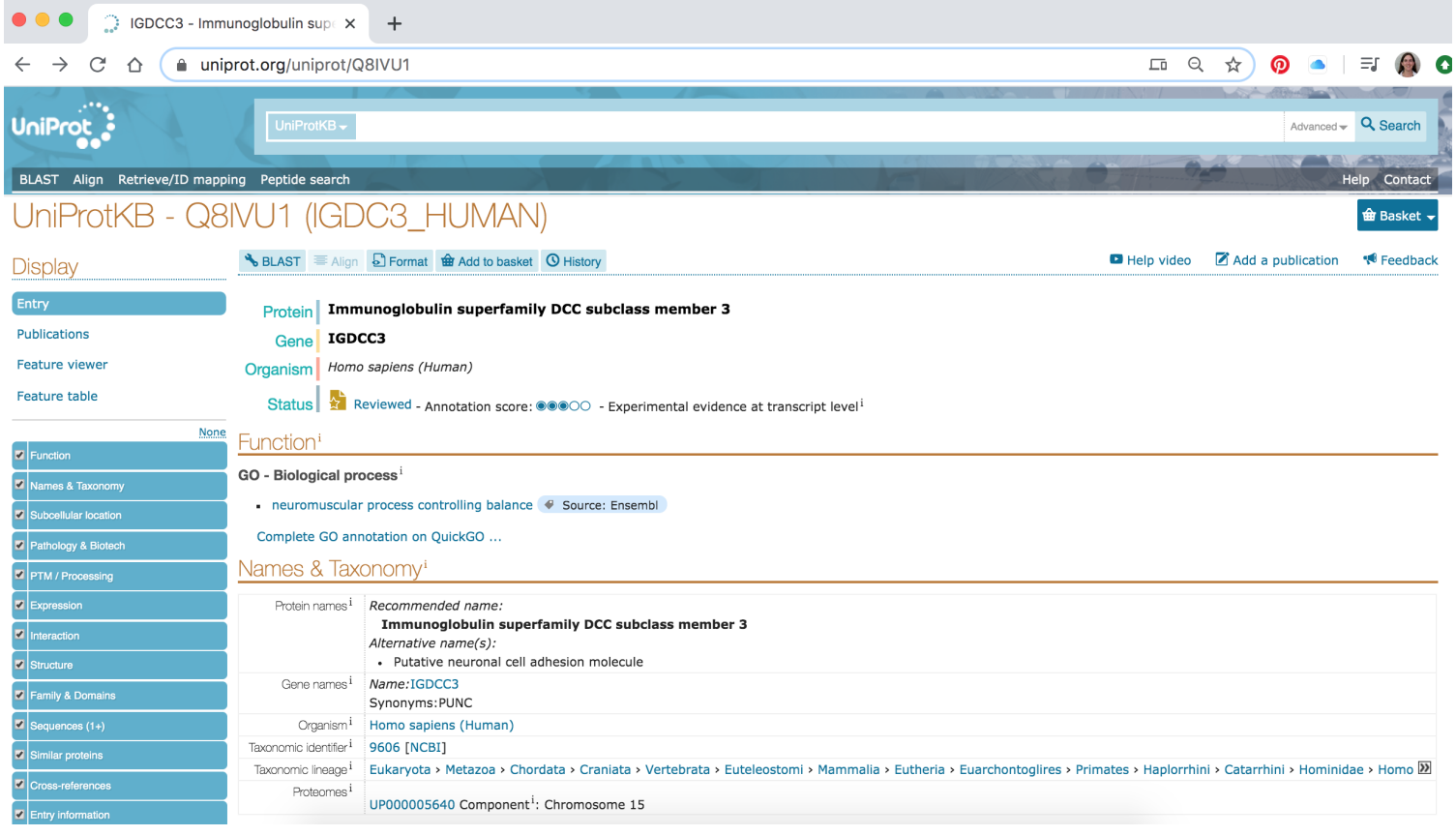
Top ten Mesenchymal Dark matter genes

```
ranked_gene_no_annotation[1:10,]
```

```
##    GeneName     rank
## 1    IGDCC3 36.32958
## 14   ZNF469 28.83028
## 40   GLT8D2 24.77158
## 53 KIAA1644 23.58145
## 61  TSPAN18 22.71841
## 74     LHFP 21.54415
## 77    VGLL3 21.34833
## 86    MEIS3 20.77773
## 88  ZCCHC24 20.71234
## 90  FAM198B 20.49151
```

IGDCC3 - Immunoglobulin superfamily DCC subclass member 3

Uniprot reference



Dark Matter

Different types of dark matter:

1. Genes that are annotated but the functions they are annotated to get filtered out because they are either too large or too small
2. Genes that have no annotations