

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

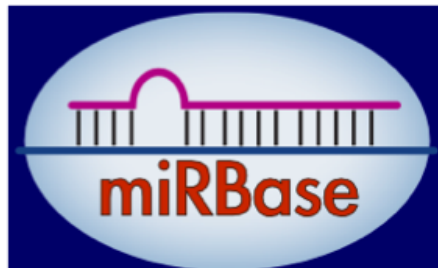
- review of Enrichment map
- looking at Pathways in depth - Reactome app, Pathway commons, String and GeneMania
- **Post analysis**
- **Enrichment Map Dark Matter**

Post/signature analysis

Drugs

 **DRUGBANK**

Regulators



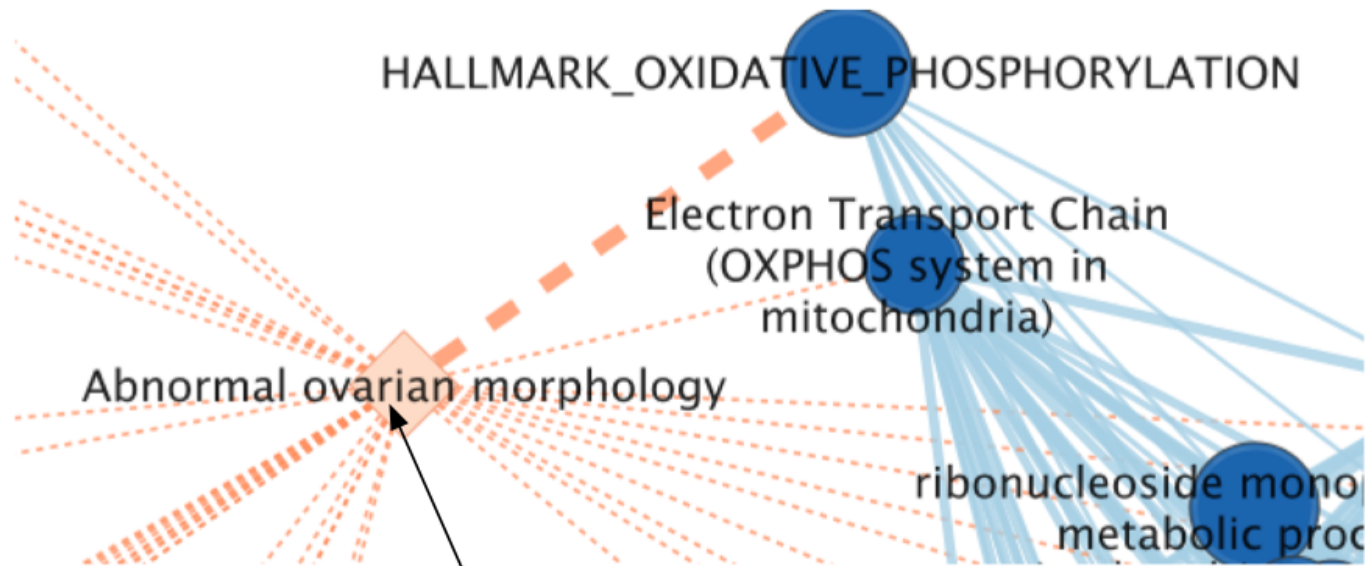
 **TRANSFAC**
database

Disease Genes/ signatures



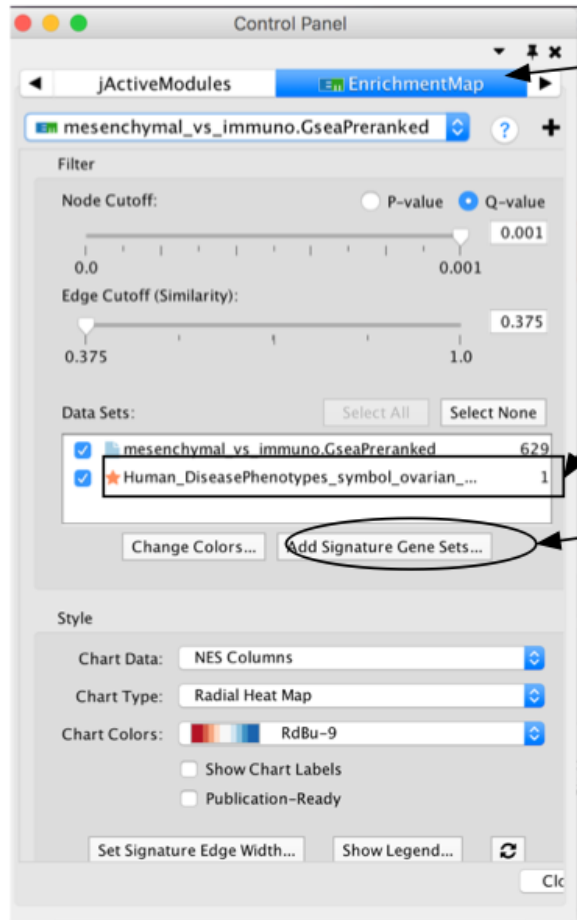
 **MSigDB**
Molecular Signatures
Database

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

EnrichmentMap: Add Signature Gene Sets (Post-Analysis)

Signature Gene Sets

Import	Name	Genes	Largest Overlap	Mann-Whitney (Two-Sided)
<input checked="" type="checkbox"/>	ABNORMALITY OF THE NERVOUS SYSTEMHP0000707	3206	109	0.0
<input checked="" type="checkbox"/>	MORPHOLOGICAL ABNORMALITY OF THE CENTRAL NERVOUS...	2245	81	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF THE CURVATURE OF THE VERTEBRAL COLU...	946	53	0.0
<input checked="" type="checkbox"/>	ABNORMAL JOINT MORPHOLOGYHP0001367	913	48	0.0
<input checked="" type="checkbox"/>	ABNORMAL MUSCLE PHYSIOLOGYHP0011804	2022	85	0.0
<input checked="" type="checkbox"/>	ABNORMAL APPENDICULAR SKELETON MORPHOLOGYHP0011406	1406	71	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF JOINT MOBILITYHP0011729	971	56	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF CARDIOVASCULAR SYSTEM MORPHOLOGY...	1391	87	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF THE SKELETAL SYSTEMHP0000924	2607	102	0.0
<input checked="" type="checkbox"/>	ABNORMAL ORAL MORPHOLOGYHP0011816	1694	77	0.0
<input checked="" type="checkbox"/>	AUTOSOMAL RECESSIVE INHERITANCEHP0000007	2543	83	0.0
<input checked="" type="checkbox"/>	ABNORMAL AXIAL SKELETON MORPHOLOGYHP0009121	2130	90	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF FINGERHP0001167	931	54	0.0
<input checked="" type="checkbox"/>	ABNORMAL EYE MORPHOLOGYHP0012372	1982	122	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF THE EYEHP0000478	2508	132	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF LIMBSHP00040064	1801	83	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF BRAIN MORPHOLOGYHP0012443	2046	78	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF THE SKINHP0000951	1735	74	0.0
<input checked="" type="checkbox"/>	ABNORMALITY OF THE EARSHP0000598	1825	98	0.0
<input checked="" type="checkbox"/>	GROWTH ABNORMALITYHP0001507	2122	81	0.0

8941 gene sets loaded, 3458 passed cutoff, 3458 selected for import

Edge Weight Parameters

Test: Mann-Whitney (Two-Sided)

Cutoff: 0.05

Data Set: All Data Sets

Ranks to use for Mann-Whitney test: mesenchymal_vs_immuno.Cse4Pre-ranked

Data Set Name: Use Default Human_DiseasePhenotypes_symbol

Cancel Finish

Using the entire Human Phenotype Database

- * What phenotypes have significant overlap with my expression data?
- * Depending on the size of network and the size of the signature set can take a long time

EnrichmentMap: Add Signature Gene Sets (Post-Analysis)

Signature Gene Sets

Import	Name	Genes	Largest Overlap	Mann-Whitney (Two-Sided)
<input checked="" type="checkbox"/>	ABNORMAL OVARIAN MORPHOLOGYHP0011065	126	11	0.0026
<input checked="" type="checkbox"/>	OVARIAN CYSTHP0000138	67	6	0.0101
<input checked="" type="checkbox"/>	ABNORMAL OVARIAN PHYSIOLOGYHP0011066	77	6	0.0136
<input checked="" type="checkbox"/>	OVARIAN NEOPLASMP0000615	63	16	0.0275
<input checked="" type="checkbox"/>	OVARIAN PAPILLARY ADENOCARCINOMASHP0006774	7	4	0.0343
<input checked="" type="checkbox"/>	OVARIAN CONADOBLASTOMASHP0000149	11	3	0.041
<input checked="" type="checkbox"/>	ATLASIA/HYPOPLASIA OF THE THYMUSHP0010515	32	3	0.0431
<input checked="" type="checkbox"/>	PREMATURE OVARIAN INSUFFICIENCYHP0008289	52	5	0.0563
<input checked="" type="checkbox"/>	OVARIAN CARCINOMASHP0025118	13	5	0.0783
<input checked="" type="checkbox"/>	OVARIAN MUCINOUS TUMORHP0001494	1	1	0.0919
<input checked="" type="checkbox"/>	OVARIAN FIBROMASHP0010618	3	2	0.1194
<input checked="" type="checkbox"/>	OVARIAN SEX CORIO-STROMAL TUMORHP0011918	1	1	0.2313
<input checked="" type="checkbox"/>	OVARIAN THECOMASHP0001083	1	1	0.2313
<input checked="" type="checkbox"/>	OVARIAN SEROUS CYSTADENOMASHP0012887	2	1	0.3025
<input checked="" type="checkbox"/>	OVARIAN TERATOMASHP0012226	1	1	0.6155
<input checked="" type="checkbox"/>	OVARIAN DERMOID CYSTHP0001274	1	1	0.6155
<input checked="" type="checkbox"/>	HEMORRHAGE, OVARIAN CYSTHP0001286	2	0	1

17 gene sets loaded, 7 passed cutoff, 7 selected for import

Edge Weight Parameters

Test: Mann-Whitney (Two-Sided)

Cutoff: 0.05

Data Set: All Data Sets

Ranks to use for Mann-Whitney test: mesenchymal_vs_immuno.Cse4Pre-ranked

Data Set Name: Use Default Human_DiseasePhenotypes_symbol_ovarian_only

Cancel Finish

Using the **subset** Human Phenotype Database

Where do specific signatures overlap with our data?

Will be relatively fast to compute.

Signature Set Metrics

EnrichmentMap: Add Signature Gene Sets (Post-Analysis)

Signature Gene Sets

Import	Name	Genes	Largest Overlap	Mann-Whitney (Two-Si...
<input checked="" type="checkbox"/>	BEVACIZUMAB%DRUGBANK%DB00112	11	5	0.0061
<input checked="" type="checkbox"/>	OLAPARIB%DRUGBANK%DB09074	3	3	0.0432
<input type="checkbox"/>	PACLITAXEL%DRUGBANK%DB01229	6	2	0.0832
<input type="checkbox"/>	DOXORUBICIN%DRUGBANK%DB00997	1	1	0.1991

4 gene sets loaded, 2 passed cutoff, 2 selected for import

Edge Weight Parameters

Test: **Mann-Whitney (Two-Sided)**

Cutoff: **Mann-Whitney (One-Sided Greater)**

Data Set: **Overlap has at least X genes**

Ranks to use for Mann-Whitney test

mesenchymal_vs_immuno.GseaPreranked: GSEARanking

Data Set Name: ☒ Use Default ovarian_cancer_drugs(1)

Cancel Finish

Different tests to calculate significance of overlap:
Mann Whitney - two sided, one sided greater or lesser
hypergeometric test
overlap has X number of genes
overlap has X percentage of geneset genes (EM or sig)

Load from file or from web.

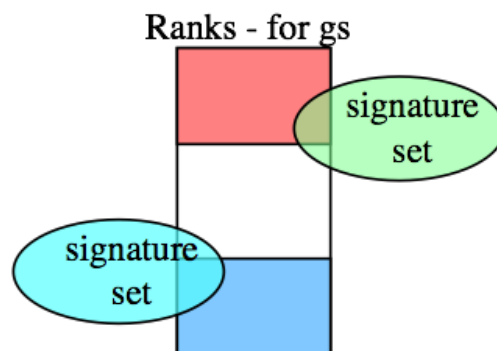
Web pulls gmt files directly from download.baderlab.org but files are large!

Mann Whitney

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

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

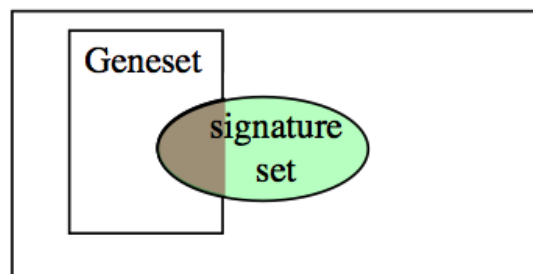
Are the genes in the signature set found mostly at the bottom of the rank list, mostly at the top of the rank list or simply ranked highly (irrespective of direction?)



Hypergeometric

Ranks are irrelevant

Is there a significant overlap size



Overlap

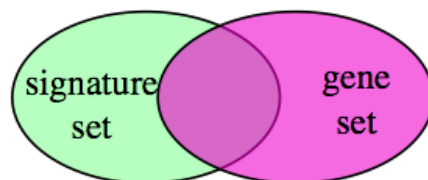
number of genes in overlap

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

EnrichmentMap: Add Signature Gene Sets (Post-Analysis)

Signature Gene Sets

Import	Name	Genes	Largest Overlap	Mann-Whitney (Greater)
<input checked="" type="checkbox"/>	MARIMASTAT%DRUGBANK%DB00786	23	22	1.7234E-05
<input checked="" type="checkbox"/>	REGORAFENIB%DRUGBANK%DB08896	18	7	0.0001
<input checked="" type="checkbox"/>	VERAPAMIL%DRUGBANK%DB00661	16	11	0.0001
<input checked="" type="checkbox"/>	SUNITINIB%DRUGBANK%DB01268	8	5	0.0002
<input checked="" type="checkbox"/>	PAZOPANIB%DRUGBANK%DB06589	10	6	0.0002
<input checked="" type="checkbox"/>	ZONISAMIDE%DRUGBANK%DB00909	31	11	0.0003
<input checked="" type="checkbox"/>	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM%DRUGBANK%DB...	4	4	0.0003
<input checked="" type="checkbox"/>	CINNARIZINE%DRUGBANK%DB00568	11	6	0.0004
<input checked="" type="checkbox"/>	IMATINIB%DRUGBANK%DB00619	9	4	0.0005
<input checked="" type="checkbox"/>	PONATINIB%DRUGBANK%DB08901	15	6	0.0006
<input checked="" type="checkbox"/>	DASATINIB%DRUGBANK%DB01254	10	4	0.0008
<input checked="" type="checkbox"/>	SORAFENIB%DRUGBANK%DB00398	10	5	0.0011
<input checked="" type="checkbox"/>	NIFEDIPINE%DRUGBANK%DB01115	8	6	0.0016
<input checked="" type="checkbox"/>	NITRENDIPINE%DRUGBANK%DB01054	8	7	0.0016
<input checked="" type="checkbox"/>	ISRADIPINE%DRUGBANK%DB00270	7	6	0.0016
<input checked="" type="checkbox"/>	DRONEDARONE%DRUGBANK%DB04855	18	11	0.0017
<input checked="" type="checkbox"/>	NIMODIPINE%DRUGBANK%DB00393	10	8	0.0017
<input checked="" type="checkbox"/>	FELODIPINE%DRUGBANK%DB01023	13	8	0.0019
<input checked="" type="checkbox"/>	AMLODIPINE%DRUGBANK%DB00381	10	7	0.0027
<input checked="" type="checkbox"/>	TENECTEPLASE%DRUGBANK%DB00031	11	5	0.0028

1456 gene sets loaded, 98 passed cutoff, 98 selected for import

Edge Weight Parameters

Test: **Mann-Whitney (One-Sided Greater)**

Cutoff: 0.05

Data Set: -- All Data Sets --

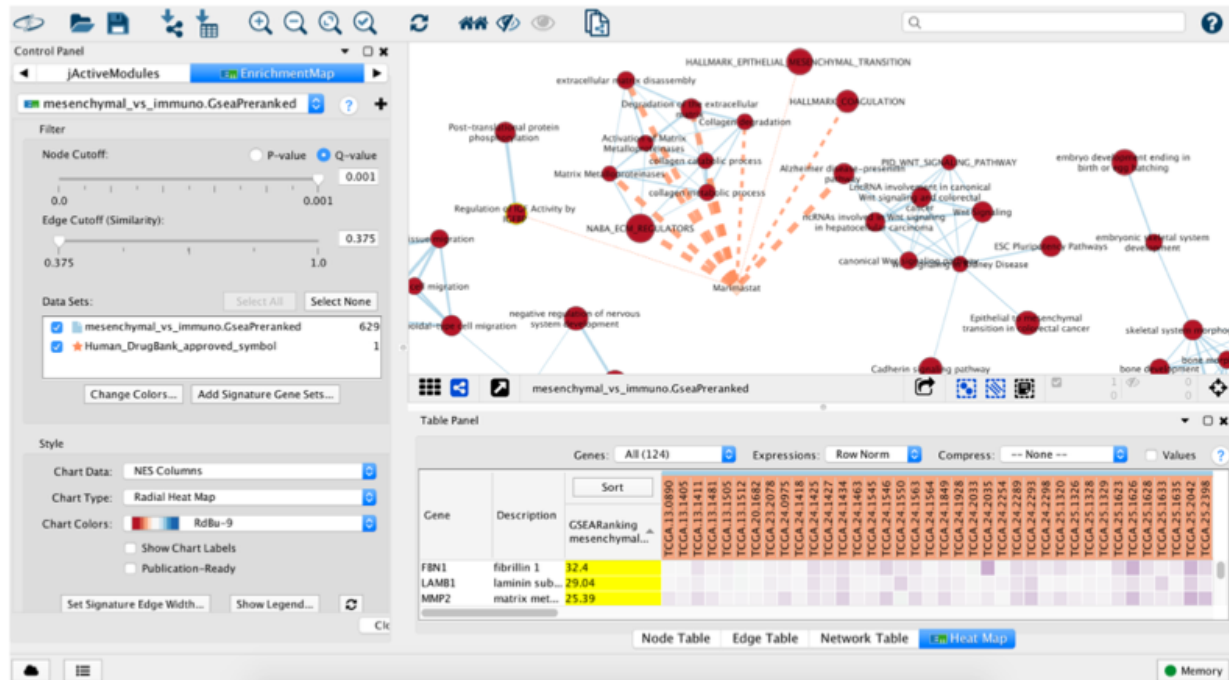
Ranks to use for Mann-Whitney test

mesenchymal_vs_immuno.GseaPreranked: GSEARanking

Data Set Name: ☒ Use Default Human_DrugBank_approved_symbol

Cancel Finish

Example Post Analysis - Drugs that can target Mesenchymal



Example Post Analysis - Drugs that can target Mesenchymal

The screenshot displays the Gene Set Enrichment Analysis (GSEA) software interface. On the left, the 'Control Panel' shows the 'ActiveModules' list with 'mesenchymal_vs_immuno.GseaPreranked' selected. The 'Filter' section shows 'Node Cutoff' set to 0.001 and 'Edge Cutoff (Similarity)' set to 0.375. The 'Data Sets' section shows 'mesenchymal_vs_immuno' and 'Human_DrugBank_approved' selected. The 'Style' section shows 'Chart Data' as 'NES Columns', 'Chart Type' as 'Radial Heat Map', and 'Chart Colors' as 'RedBlue'. The main window shows a network diagram with nodes and edges, representing biological pathways. A pink overlay from DrugBank is visible, showing the entry for Marimastat.

DRUGBANK Browse Search Downloads Commercial Data Help

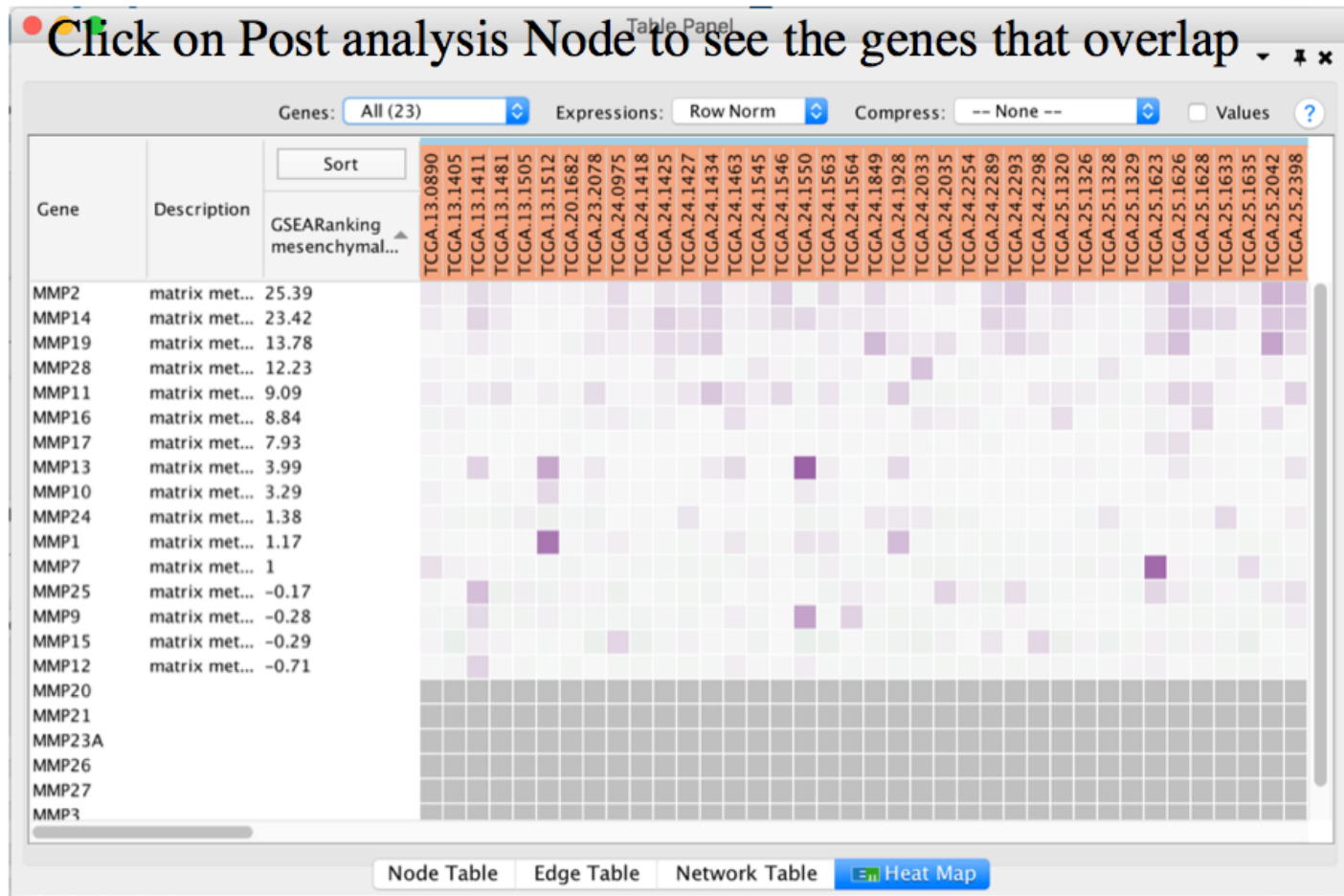
Drugs

Marimastat Targets (23) BioInteractions (23)

IDENTIFICATION

Name	Marimastat
Accession Number	DB00786 (APRD00559)
Type	Small Molecule
Groups	Investigational
Description	Used in the treatment of cancer, Marimastat is an angiogenesis and metastasis inhibitor. As an angiogenesis inhibitor it limits the growth and production of blood vessels. As an antimetastatic agent it prevents malignant cells from breaching the basement membranes.
Structure	

Example Post Analysis - Drugs that can target Mesenchymal

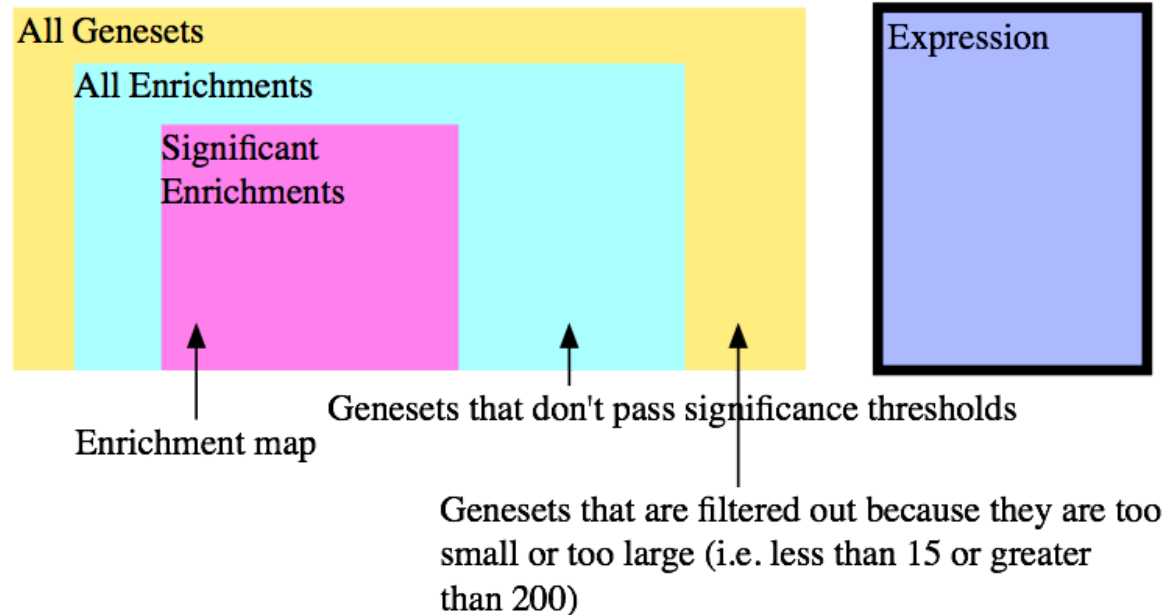


Post Analysis Summary

Dark Matter

What is dark matter?

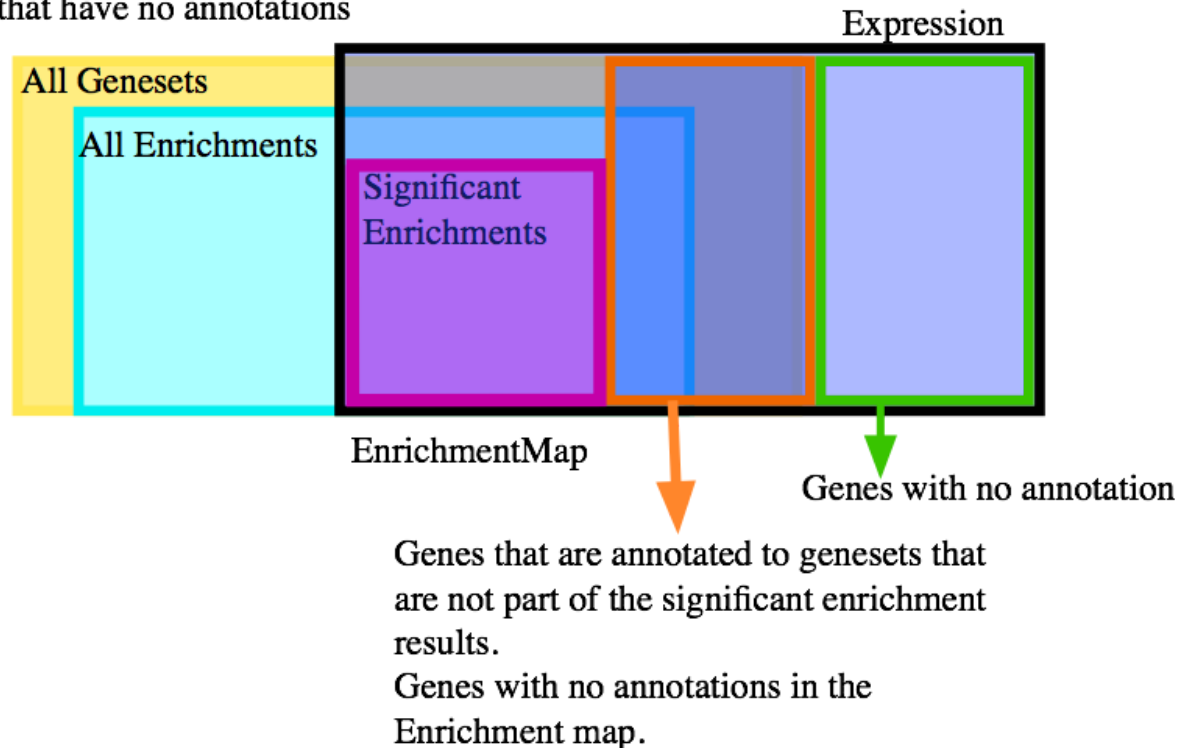
Sometimes the genes we don't see in our pathway results are just as important as the genes that we do see.



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 pathways (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 pathways (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.

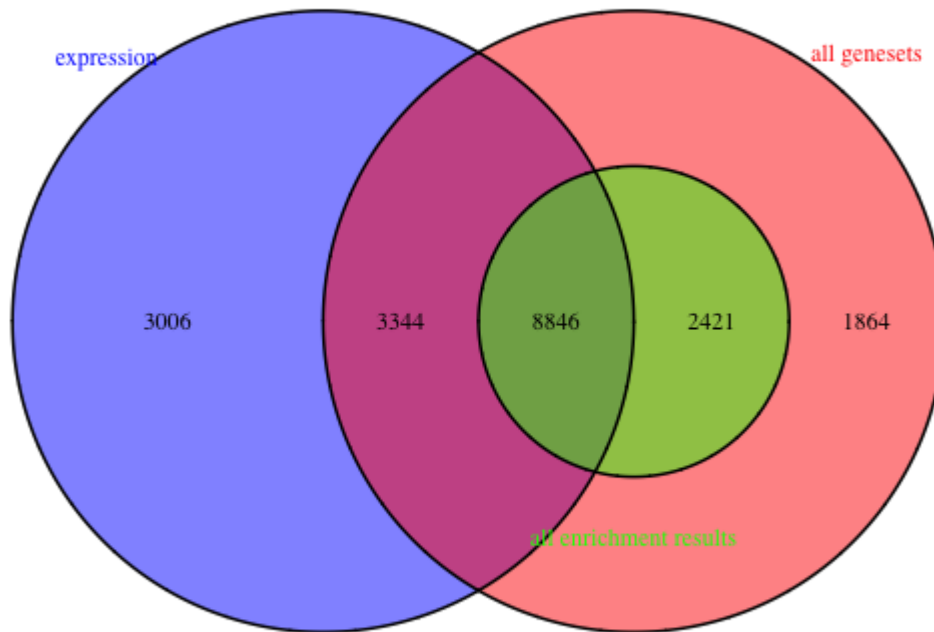
Venn Diagram of Dark Matter Overlaps

```
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

IGDCC3 - Immunoglobulin superfamily DCC subclass member 3

uniprot.org/uniprot/Q8IVU1

UniProtKB

BLAST Align Retrieve/ID mapping Peptide search

UniProtKB - Q8IVU1 (IGDCC3_HUMAN)

Display

Entry

Publications

Feature viewer

Feature table

Protein | Immunoglobulin superfamily DCC subclass member 3

Gene | IGDCC3

Organism | Homo sapiens (Human)

Status | Reviewed - Annotation score: ●●●○○ - Experimental evidence at transcript levelⁱ

Functionⁱ

GO - Biological processⁱ

- neuromuscular process controlling balance Source: Ensembl

Complete GO annotation on QuickGO ...

Names & Taxonomyⁱ

Protein names ⁱ	Recommended name: Immunoglobulin superfamily DCC subclass member 3 Alternative name(s): <ul style="list-style-type: none">Putative neuronal cell adhesion molecule
Gene names ⁱ	Name: IGDCC3 Synonyms: PUNC
Organism ⁱ	Homo sapiens (Human)
Taxonomic identifier ⁱ	9606 [NCBI]
Taxonomic lineage ⁱ	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo [DB]
Proteomes ⁱ	UP000005640 Component ⁱ : Chromosome 15

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

