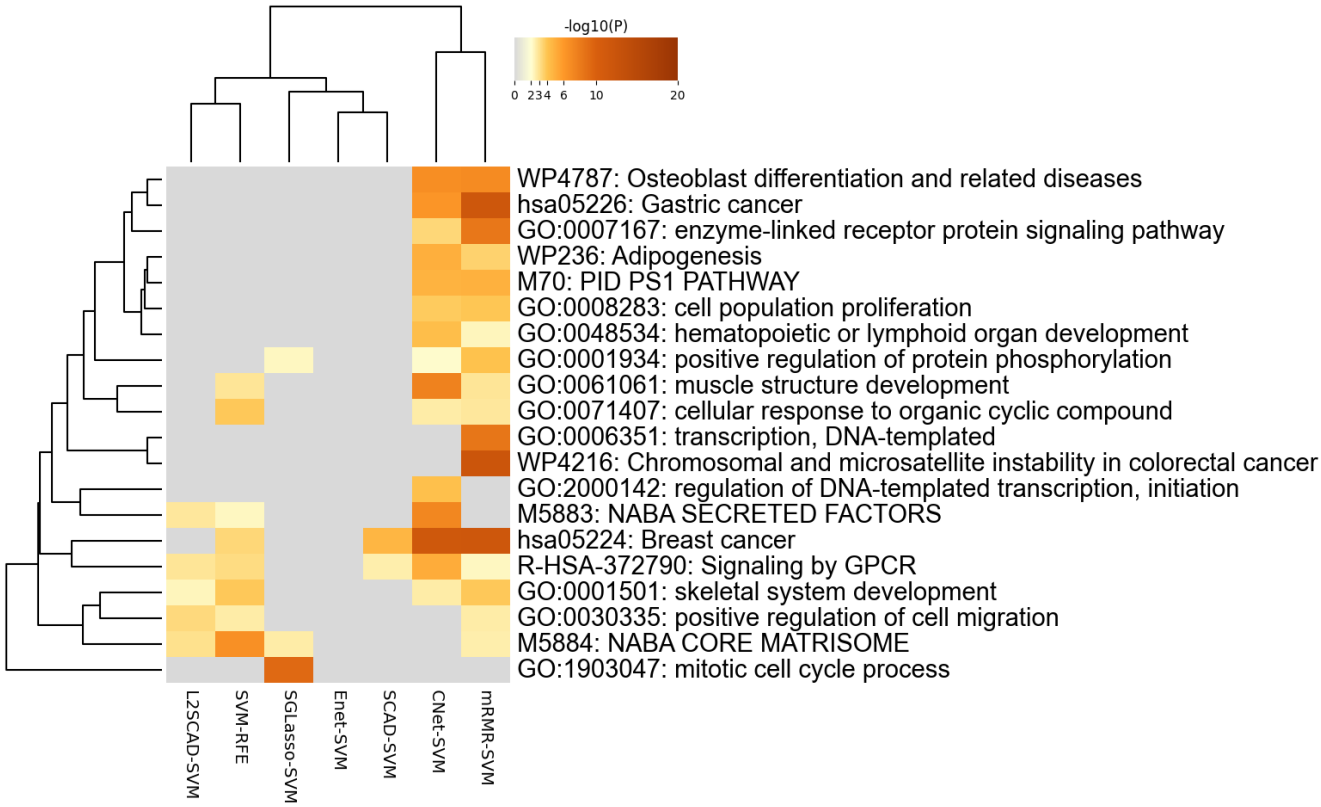
Metascape Gene List Analysis Report

metascape.org (http://metascape.org)[1]

Heatmap Summary

Figure 1. Heatmap of enriched terms across input gene lists, colored by p-values.

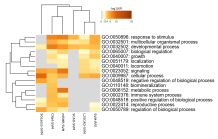


(./Enrichment_heatmap/HeatmapSelectedGO.pdf)

Metascape only visualizes the top 20 clusters. Up to 100 enriched clusters can be viewed here.

(./Enrichment_heatmap/HeatmapSelectedGOTop100.png) (./Enrichment_heatmap/HeatmapSelectedGOTop100.pdf)

The top-level Gene Ontology biological processes can be viewed here.



(./Enrichment_heatmap/HeatmapSelectedGOParent.png) (./Enrichment_heatmap/HeatmapSelectedGOParent.pdf)

The heatmap can be interactively viewed using JTreeView (http://jtreeview.sourceforge.net)[2] (.cdt, .gtr and .atr files can be found in the Zip package).

Gene Lists

User-provided gene identifiers are first converted into their corresponding H. sapiens Entrez gene IDs using the latest version of the database (last updated on 2022-04-22). If multiple identifiers correspond to the same Entrez gene ID, they will be considered as a single Entrez gene ID in downstream analyses. Each gene list is assigned a unique color, which is used throughout the analysis. The gene lists are summarized in Table 1.

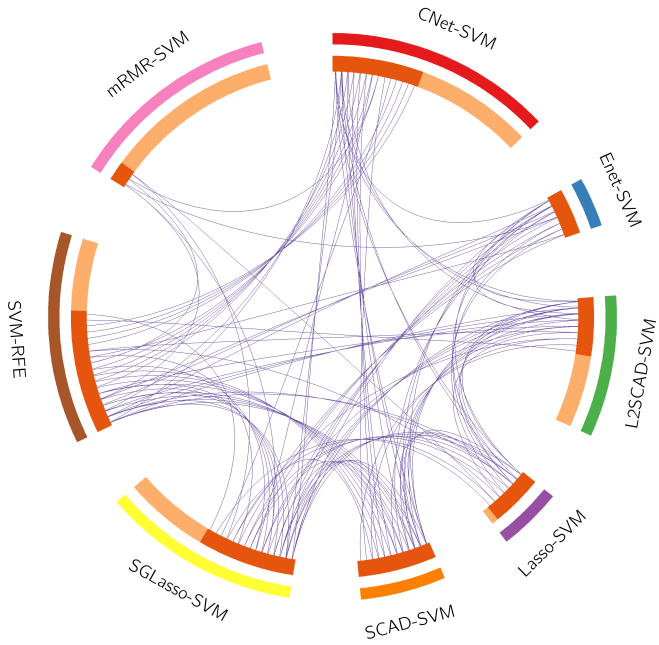
Table 1. Statistics of input gene lists.

Name	Total	Unique	Color Code
CNet-SVM	32	32	
Enet-SVM	7	7	
L2SCAD-SVM	20	20	
Lasso-SVM	9	9	
SCAD-SVM	12	12	

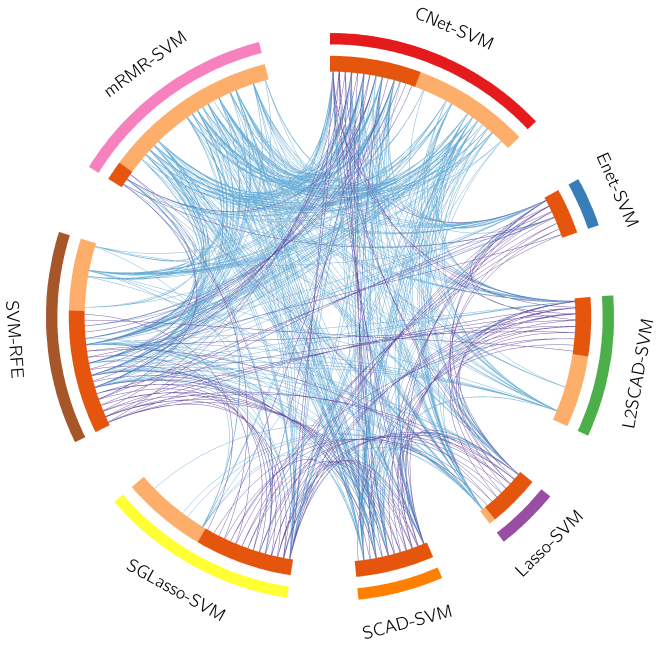
Name	Total	Unique	Color Code
SGLasso-SVM	28	28	
SVM-RFE	30	30	
mRMR-SVM	30	30	

The overlaps between these lists are shown in a Circos (http://circos.ca)[3] plot (Figure 2.a). Another useful representation is to overlap genes based on their functions or shared pathways. The overlaps between gene lists can be significantly improved by considering overlaps between genes sharing the same enriched ontology term(s) (Figure 2.b). Only ontology terms that contain less than 100 genes are used to calculate functional overlaps to avoid linking genes using very general annotation. (We do not want to link all genes, only genes that belong to specific biological processes.)

Figure 2. Overlap between gene lists: (a) only at the gene level, where purple curves link identical genes; (b) including the shared term level, where blue curves link genes that belong to the same enriched ontology term. The inner circle represents gene lists, where hits are arranged along the arc. Genes that hit multiple lists are colored in dark orange, and genes unique to a list are shown in light orange. The publication-quality version of the figures is included in the Zip package as a .svg file under the Overlap_circos folder (readable by popular web browsers and Adobe Illustrator).



(./Overlap_circos/CircosOverlapByGene.svg)



(./Overlap_circos/CircosOverlapByGO.svg)

Gene Annotation

The following are the list of annotations retrieved from the latest version of the database (last updated on 2022-04-22) (Table 2).

Table 2. Gene annotations extracted

Name	Type	Description
Gene Symbol	Description	Primary HUGO gene symbol.
Description	Description	Short description.
Biological Process (GO)	Function/Location	Descriptions summarized based on gene ontology database, where up to three most informative GO terms are kept.
Kinase Class (UniProt)	Function/Location	Detailed kinase classes.
Protein Function (Protein Atlas)	Function/Location	Protein Function (Protein Atlas)
Subcellular Location (Protein Atlas)	Function/Location	Sucellular Location (Protein Atlas)
Drug (DrugBank)	Genotype/Phenotype/Disease	Drug information for the given gene as target.
Canonical Pathways	Ontology	Canonical Pathways
Hallmark Gene Sets	Ontology	Hallmark Gene Sets

Pathway and Process Enrichment Analysis

For each given gene list, pathway and process enrichment analysis has been carried out with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, Cell Type Signatures, CORUM, TRRUST, DisGeNET, PaGenBase, Transcription Factor Targets, WikiPathways, PANTHER Pathway and COVID. All genes in the genome have been used as the enrichment background. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities. More specifically, p-values are calculated based on the cumulative hypergeometric distribution[4], and q-values are calculated using the Benjamini-Hochberg procedure to account for multiple testings[5]. Kappa scores[6] are used as the similarity metric when performing hierarchical clustering on the enriched terms, and sub-trees with a similarity of > 0.3 are considered a cluster. The most statistically significant term within a cluster is chosen to represent the cluster.
When multiple gene lists are provided, all lists are merged into one list called "_FINAL". A term may be found enriched in several individual gene lists and/or in the _FINAL gene list, and the best p-value among them is chosen as the final p-value. The pathway/process clusters that are found to be of interest (either shared or unique based on specific list enrichment) are used to prioritize the genes that fall into those clusters (membership is presented as 1/0 binary columns in the Excel spreadsheet). Note that individual gene lists containing more than 3000 genes are ignored during the enrichment analysis to avoid superficial terms; this is because long gene lists are often not random and generally trigger too many terms that are not of direct relevance to the biology under study.

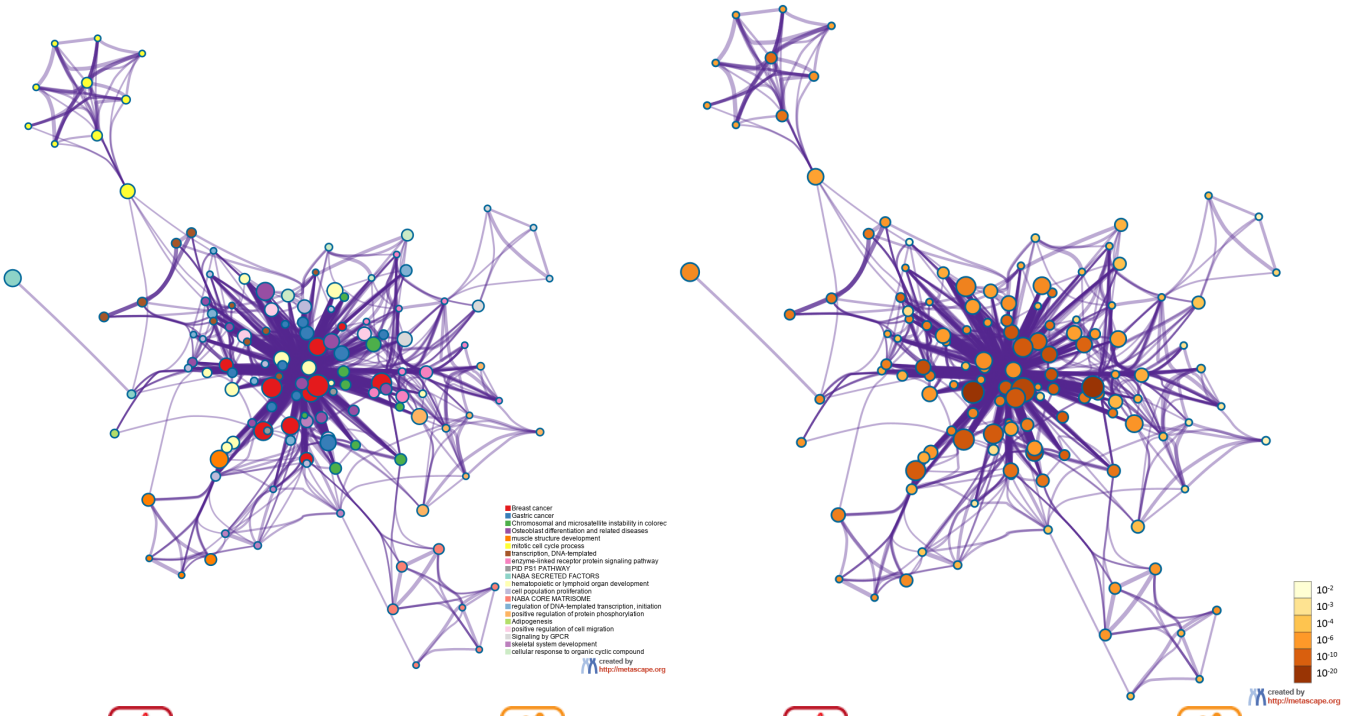
Table 3. Top 20 clusters with their representative enriched terms (one per cluster). "Count" is the number of genes in the user-provided lists with membership in the given ontology term. "%" is the percentage of all of the user-provided genes that are found in the given ontology term (only input genes with at least one ontology term annotation are included in the calculation). "Log10(P)" is the p-value in log base 10. "Log10(q)" is the multi-test adjusted p-value in log base 10. __PATTERN__ shows the color code used for the gene lists where the term is found statistically significant, i.e., multiple colors indicate a pathway/process that is shared across multiple lists.

PATTERN	GO	Category	Description	Count	%	Log10(P)	Log10(q)
	hsa05224	KEGG Pathway	Breast cancer	18	16.51	-22.04	-17.69
	hsa05226	KEGG Pathway	Gastric cancer	14	12.84	-15.54	-11.72
	WP4216	WikiPathways	Chromosomal and microsatellite instability in colorectal cancer	10	9.17	-12.84	-9.33
	WP4787	WikiPathways	Osteoblast differentiation and related diseases	11	10.09	-12.17	-8.73
	GO:0061061	GO Biological Processes	muscle structure development	16	14.68	-10.65	-7.47
	GO:1903047	GO Biological Processes	mitotic cell cycle process	9	32.14	-9.24	-6.04
	GO:0006351	GO Biological Processes	transcription, DNA-templated	8	26.67	-8.37	-5.26
	GO:0007167	GO Biological Processes	enzyme-linked receptor protein signaling pathway	9	30.00	-8.31	-5.24
	M70	Canonical Pathways	PID PS1 PATHWAY	6	5.50	-7.80	-5.17
	M5883	Canonical Pathways	NABA SECRETED FACTORS	7	21.88	-7.22	-4.41
	GO:0048534	GO Biological Processes	hematopoietic or lymphoid organ development	14	12.84	-6.97	-4.54
	GO:0008283	GO Biological Processes	cell population proliferation	12	11.01	-6.64	-4.26
	M5884	Canonical Pathways	NABA CORE MATRISOME	6	20.00	-6.57	-3.92
	GO:2000142	GO Biological Processes	regulation of DNA-templated transcription, initiation	6	5.50	-6.47	-4.13
	GO:0001934	GO Biological Processes	positive regulation of protein phosphorylation	14	12.84	-6.32	-4.01
	WP236	WikiPathways	Adipogenesis	7	6.42	-6.29	-3.99
	GO:0030335	GO Biological Processes	positive regulation of cell migration	12	11.01	-6.05	-3.80
	R-HSA-372790	Reactome Gene Sets	Signaling by GPCR	13	11.93	-5.76	-3.59

PATTERN	GO	Category	Description	Count	%	Log10(P)	Log10(q)
	GO:0001501	GO Biological Processes	skeletal system development	11	10.09	-5.65	-3.51
	GO:0071407	GO Biological Processes	cellular response to organic cyclic compound	11	10.09	-5.65	-3.51

To further capture the relationships between the terms, a subset of enriched terms have been selected and rendered as a network plot, where terms with a similarity > 0.3 are connected by edges. We select the terms with the best p-values from each of the 20 clusters, with the constraint that there are no more than 15 terms per cluster and no more than 250 terms in total. The network is visualized using Cytoscape (http://www.cytoscape.org)[7], where each node represents an enriched term and is colored first by its cluster ID (Figure 3.a) and then by its p-value (Figure 3.b). These networks can be interactively viewed in Cytoscape through the .cys files (contained in the Zip package, which also contains a publication-quality version as a PDF) or within a browser by clicking on the web icon. For clarity, term labels are only shown for one term per cluster, so it is recommended to use Cytoscape or a browser to visualize the network in order to inspect all node labels. We can also export the network into a PDF file within Cytoscape, and then edit the labels using Adobe Illustrator for publication purposes. To switch off all labels, delete the "Label" mapping under the "Style" tab within Cytoscape, and then export the network view.

Figure 3. Network of enriched terms: (a) colored by cluster ID, where nodes that share the same cluster ID are typically close to each other; (b) colored by p-value, where terms containing more genes tend to have a more significant p-value.

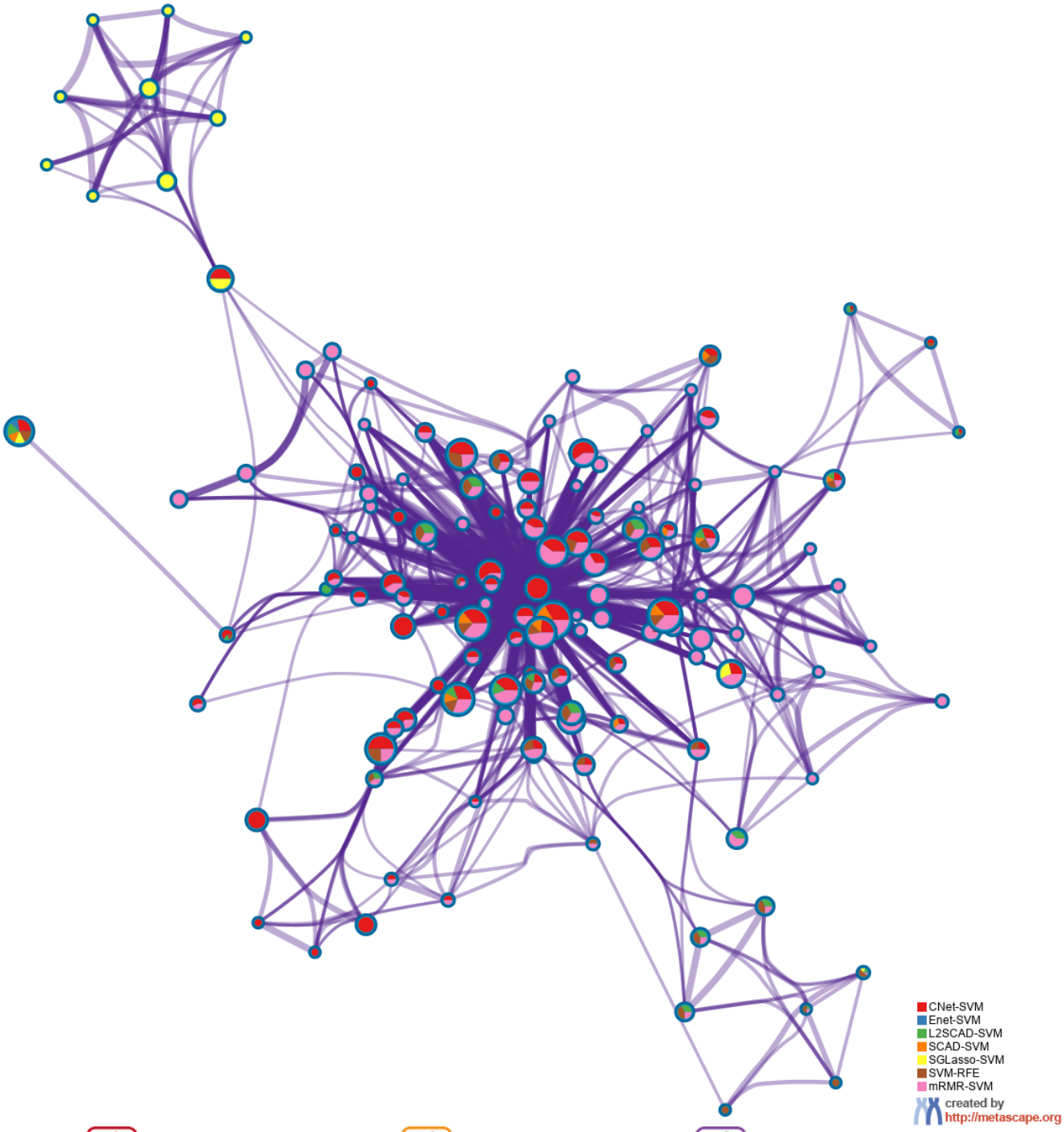


(./Enrichment_GO/ColorByCluster.pdf) (./Enrichment_GO/GONetwork.cys) (Enrichment_GO/GONetwork.html?Network=GONetwork&Style=ColorByCluster)

(./Enrichment_GO/ColorByPValue.pdf) (./Enrichment_GO/GONetwork.cys) (Enrichment_GO/GONetwork.html?Network=GONetwork&Style=ColorByPValue)

In the case of when multiple gene lists are provided, the nodes are represented as pie charts, where the size of a pie is proportional to the total number of hits that fall into that specific term. The pie charts are color-coded based on the gene list identities, where the size of a slice represents the percentage of genes under the term that originated from the corresponding gene list. This plot is particularly useful for visualizing whether the terms are shared by multiple lists or unique to a specific list, as well as for understanding how these terms associate with each other within the biological context of the meta study (Figure 4).

Figure 4. Network of enriched terms represented as pie charts, where pies are color-coded based on the identities of the gene lists.



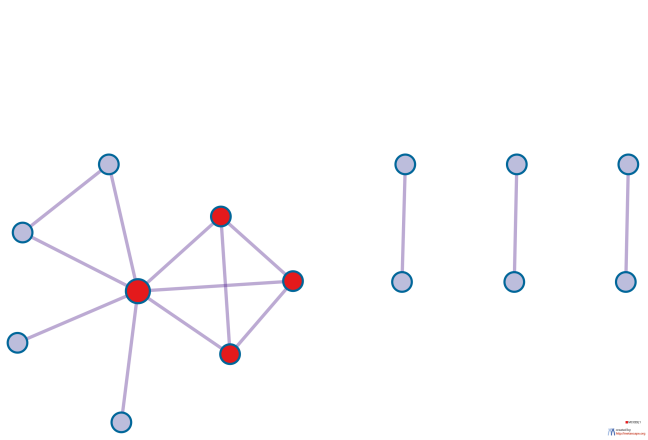
(./Enrichment_GO/ColorByCounts.pdf) (./Enrichment_GO/GONetwork.cys) (Enrichment_GO/GONetwork.html?Network=GONetwork&Style=ColorByCounts)

Protein-protein Interaction Enrichment Analysis

For each given gene list, protein-protein interaction enrichment analysis has been carried out with the following databases: STRING[8], BioGrid[9], OmniPath[10], InWeb_IM[11].Only physical interactions in STRING (physical score > 0.132) and BioGrid are used (details (http://metascape.org/blog/?p=219)). The resultant network contains the subset of proteins that form physical interactions with at least one other member in the list. If the network contains between 3 and 500 proteins, the Molecular Complex Detection (MCODE) algorithm[12] has been applied to identify densely connected network components. The MCODE networks identified for individual gene lists have been gathered and are shown in Figure 5.

Pathway and process enrichment analysis has been applied to each MCODE component independently, and the three best-scoring terms by p-value have been retained as the functional description of the corresponding components, shown in the tables underneath corresponding network plots within Figure 5.

Figure 5. Protein-protein interaction network and MCODE components identified in the gene lists.



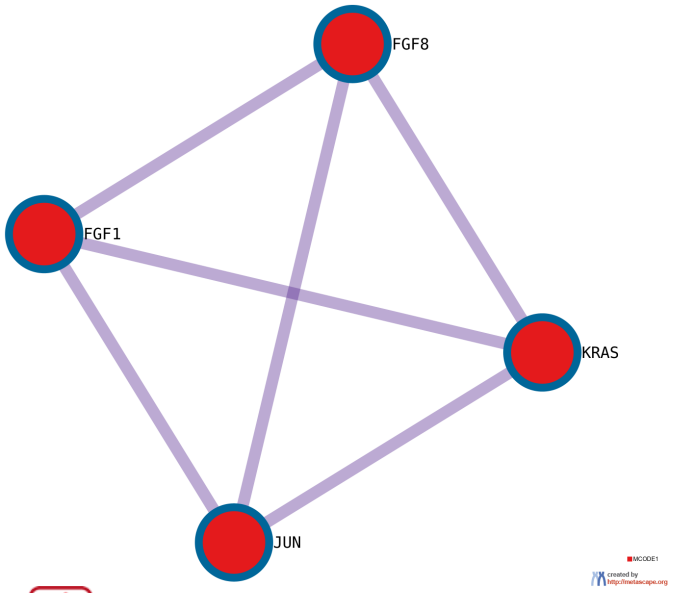
(./Enrichment_PPI/CNet-SVM_PPIColorByCluster.pdf)

(./Enrichment_PPI/MCODE_PPI.cys)

(Enrichment_PPI/PPINetwork.html?Network=CNet-SVM_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

CNet-SVM (Full Connection)

GO	Description	Log10(P)
WP3931	Embryonic stem cell pluripotency pathways	-11.0
hsa05224	Breast cancer	-10.5
WP4262	Breast cancer pathway	-10.3



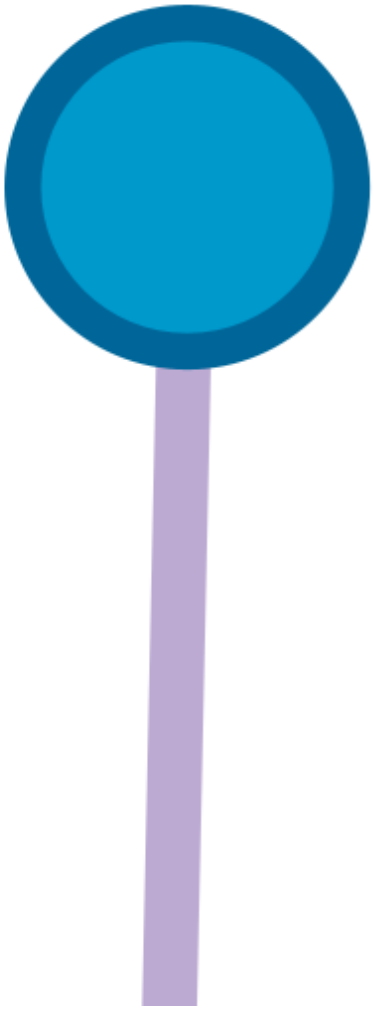
(./Enrichment_PPI/CNet-SVM_MCODE_ALL_PPIColorByCluster.pdf)

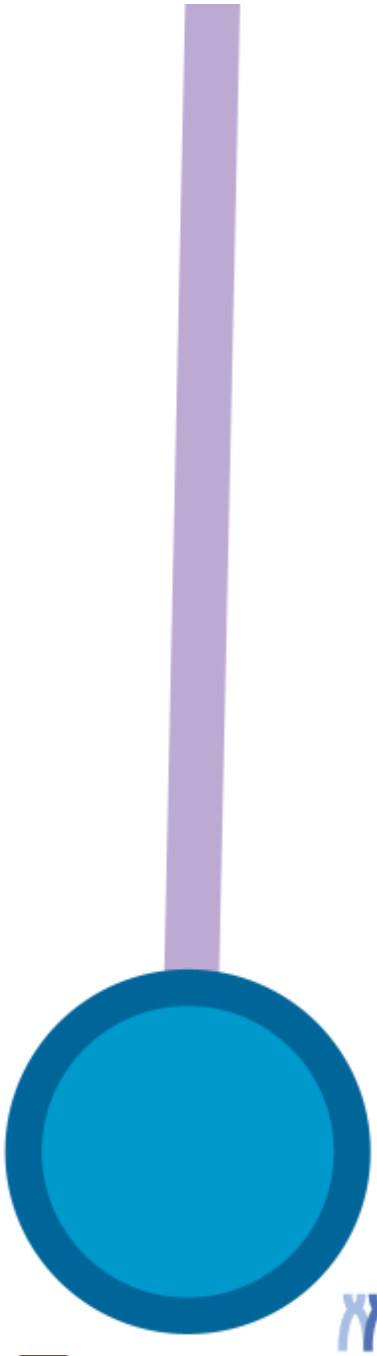
(./Enrichment_PPI/MCODE_PPI.cys)

(Enrichment_PPI/PPINetwork.html?Network=CNet-SVM_MCODE_ALL_PPIColorByCluster&Style=PPIColorByCluster&isPPI=True)

CNet-SVM (Keep MCODE Nodes Only)

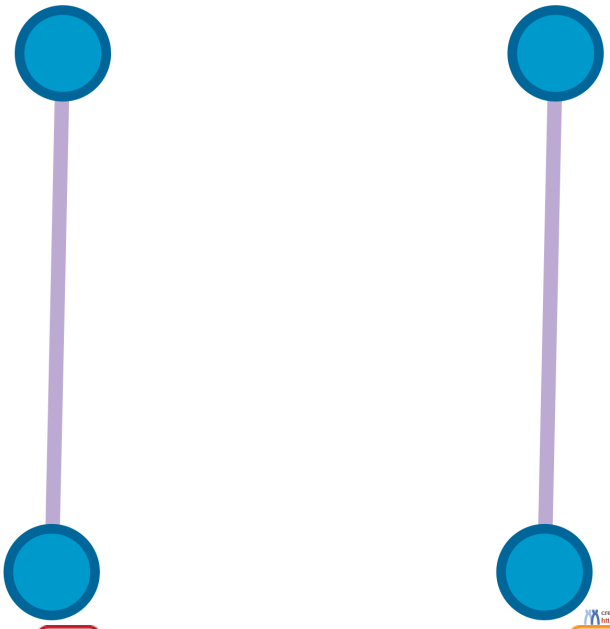
Color	MCODE	GO	Description	Log10(P)
	MCODE_1	hsa05224	Breast cancer	-9.3
	MCODE_1	WP4262	Breast cancer pathway	-9.2
	MCODE_1	R-HSA-5654704	SHC-mediated cascade:FGFR3	-9.1





(./Enrichment_PPI/L2SCAD-SVM_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=L2SCAD-SVM_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

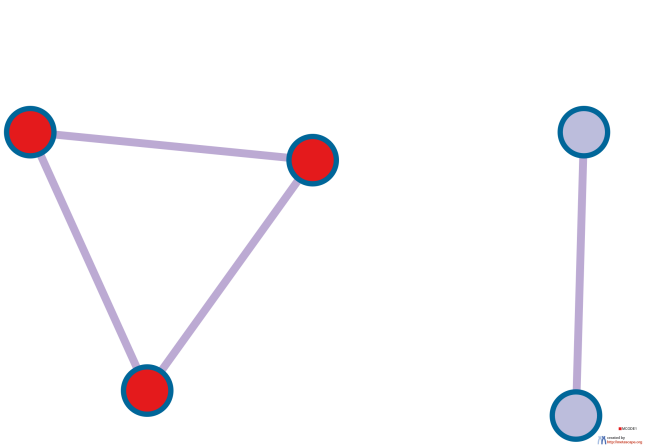
L2SCAD-SVM (Full Connection)



(./Enrichment_PPI/SCAD-SVM_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=SCAD-SVM_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

SCAD-SVM (Full Connection)

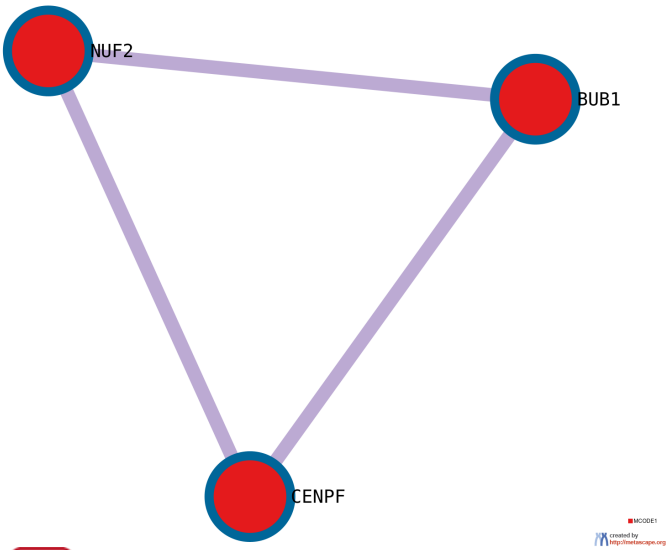
GO	Description	Log10(P)
R-HSA-388396	GPCR downstream signalling	-4.4
R-HSA-372790	Signaling by GPCR	-4.3



(./Enrichment_PPI/SGLasso-SVM_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=SGLasso-SVM_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

SGLasso-SVM (Full Connection)

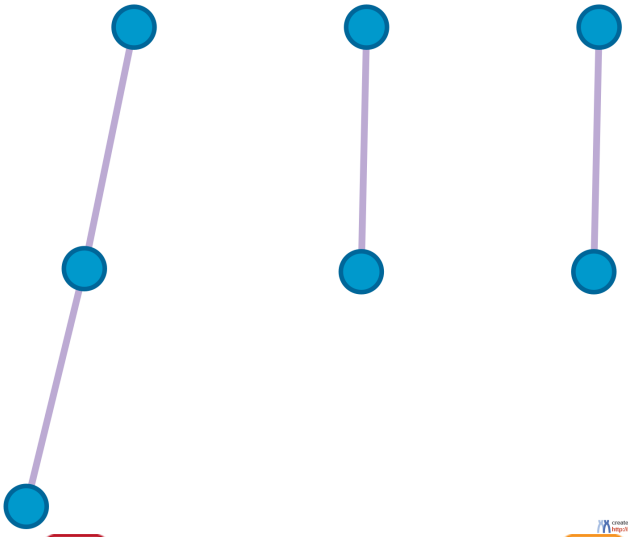
GO	Description	Log10(P)
GO:0007088	regulation of mitotic nuclear division	-9.0
GO:0051301	cell division	-8.9
GO:1903047	mitotic cell cycle process	-8.9



(./Enrichment_PPI/SGLasso-SVM_MCODE_ALL_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=SGLasso-SVM_MCODE_ALL_PPIColorByCluster&Style=PPIColorByCluster&isPPI=True)

SGLasso-SVM (Keep MCODE Nodes Only)

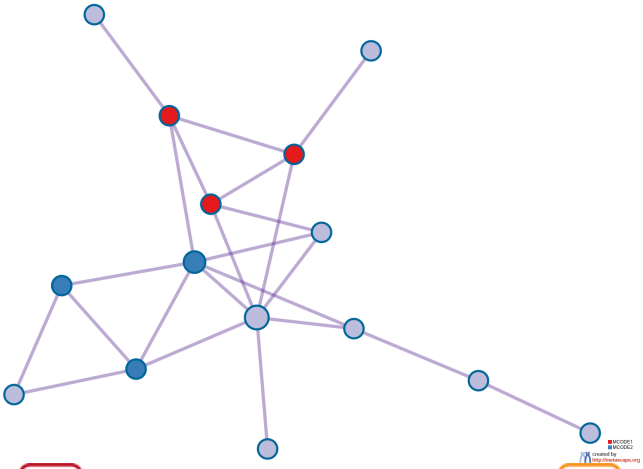
Color	MCODE	GO	Description	Log10(P)
	MCODE_1	GO:0007094	mitotic spindle assembly checkpoint signaling	-9.1
	MCODE_1	GO:0071174	mitotic spindle checkpoint signaling	-9.1
	MCODE_1	GO:0071173	spindle assembly checkpoint signaling	-9.1



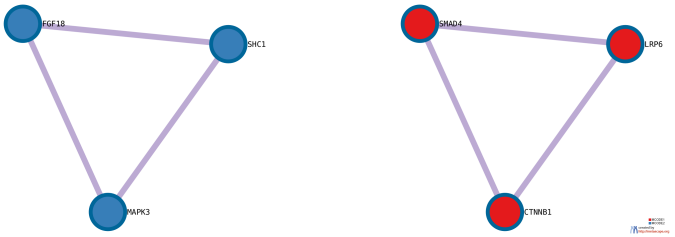
(./Enrichment_PPI/SVM-RFE_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=SVM-RFE_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

SVM-RFE (Full Connection)

GO	Description	Log10(P)
WP3932	Focal adhesion: PI3K-Akt-mTOR-signaling pathway	-4.4
WP4172	PI3K-Akt signaling pathway	-4.3
hsa04151	PI3K-Akt signaling pathway	-4.3



(./Enrichment_PPI/mRMR-SVM_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=mRMR-SVM_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)



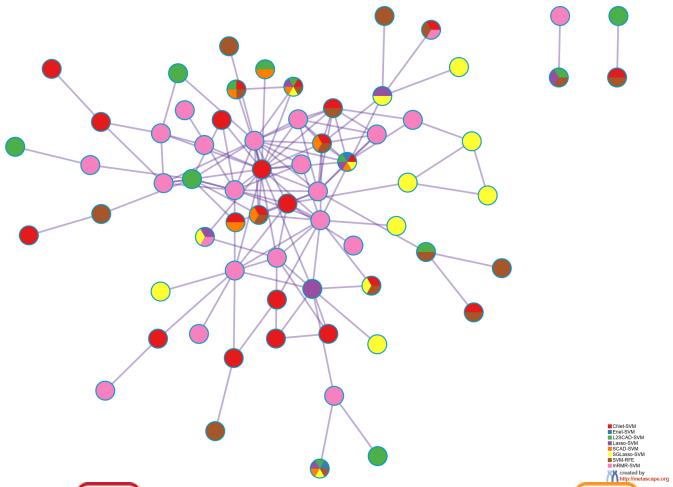
(./Enrichment_PPI/mRMR-SVM_MCODE_ALL_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=mRMR-SVM_MCODE_ALL_PPIColorByCluster&Style=PPIColorByCluster&isPPI=True)

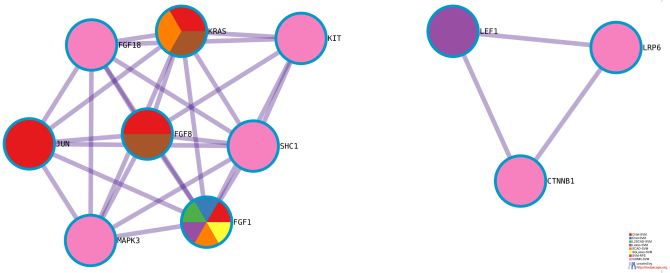
mRMR-SVM (Full Connection)

GO	Description	Log10(P)
R-HSA-5663202	Diseases of signal transduction by growth factor receptors and second messengers	-13.0
hsa05224	Breast cancer	-12.5
hsa05226	Gastric cancer	-12.4

mRMR-SVM (Keep MCODE Nodes Only)

Color	MCODE	GO	Description	Log10(P)
	MCODE_1	GO:0001837	epithelial to mesenchymal transition	-7.8
	MCODE_1	WP3931	Embryonic stem cell pluripotency pathways	-7.2
	MCODE_1	WP4787	Osteoblast differentiation and related diseases	-7.2
	MCODE_2	R-HSA-5654741	Signaling by FGFR3	-8.7
	MCODE_2	R-HSA-5654743	Signaling by FGFR4	-8.6
	MCODE_2	R-HSA-5654736	Signaling by FGFR1	-8.4





(./Enrichment_PPI/_FINAL_PPIColorByCounts.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=_FINAL_PPIColorByCluster&Style=PPIColorByCountsNoLabel&isPPI=True)

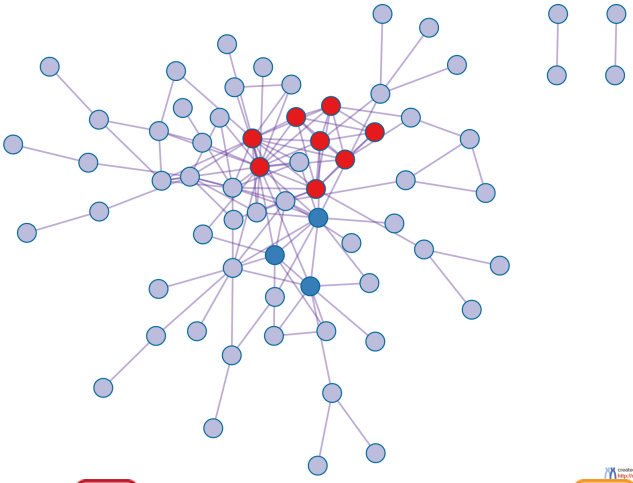
(./Enrichment_PPI/_FINAL_MCODE_ALL_PPIColorByCounts.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.h Network=_FINAL_MCODE_ALL_PPIColorByCluster&Style=PPIColorByCounts&isI

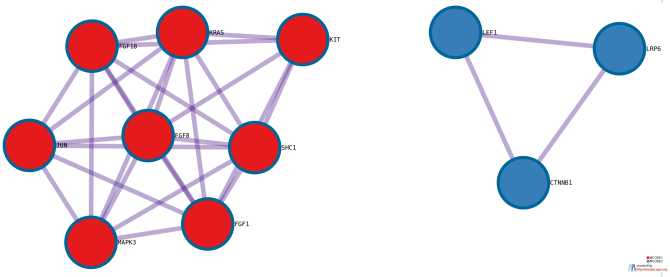
All lists merged Colored by Counts(Full Connection)

GO	Description	Log10(P)
hsa05224	Breast cancer	-24.4
WP4262	Breast cancer pathway	-23.9
hsa05200	Pathways in cancer	-19.0

All lists merged Colored by Counts(Keep MCODE Nodes Or

Color	MCODE	GO	Description	L
	MCODE_1	hsa05224	Breast cancer	-
	MCODE_1	WP4262	Breast cancer pathway	-
	MCODE_1	R-HSA-5654741	Signaling by FGFR3	-
	MCODE_2	WP3658	Wnt/beta-catenin signaling pathway in leukemia	-5
	MCODE_2	WP363	Wnt signaling pathway	-8
	MCODE_2	GO:0001837	epithelial to mesenchymal transition	-7





(./Enrichment_PPI/_FINAL_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.html?Network=_FINAL_PPIColorByCluster&Style=PPIColorByClusterNoLabel&isPPI=True)

(./Enrichment_PPI/_FINAL_MCODE_ALL_PPIColorByCluster.pdf) (./Enrichment_PPI/MCODE_PPI.cys) (Enrichment_PPI/PPINetwork.h Network=_FINAL_MCODE_ALL_PPIColorByCluster&Style=PPIColorByCluster&isI

All lists merged Colored by Cluster(Full Connection)

GO	Description	Log10(P)	Color
hsa05224	Breast cancer	-24.4	
WP4262	Breast cancer pathway	-23.9	
hsa05200	Pathways in cancer	-19.0	

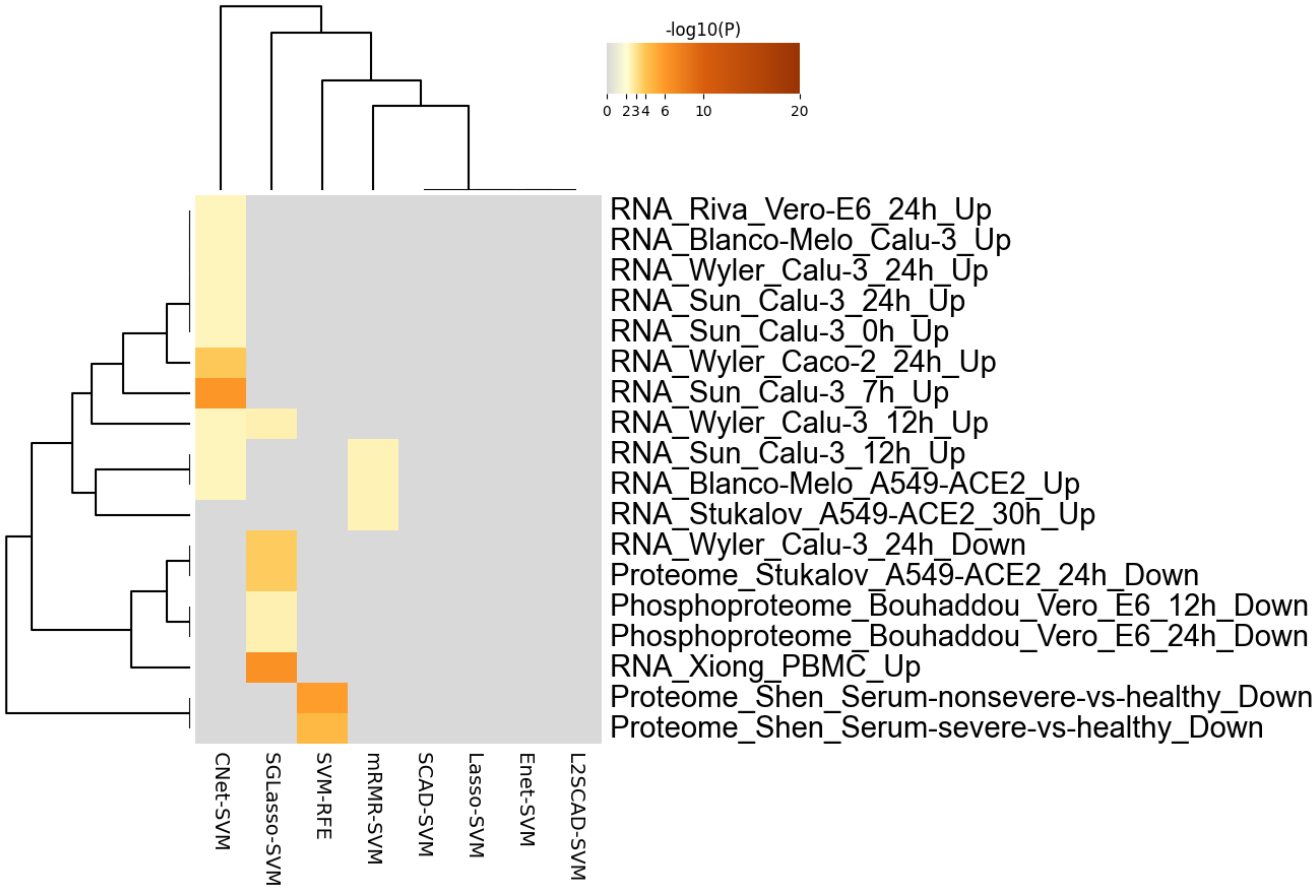
All lists merged Colored by Cluster(Keep MCODE Nodes Or

MCODE	GO	Description	L
MCODE_1	hsa05224	Breast cancer	-
MCODE_1	WP4262	Breast cancer pathway	-
MCODE_1	R-HSA-5654741	Signaling by FGFR3	-
MCODE_2	WP3658	Wnt/beta-catenin signaling pathway in leukemia	-9
MCODE_2	WP363	Wnt signaling pathway	-8
MCODE_2	GO:0001837	epithelial to mesenchymal transition	-7

Quality Control and Association Analysis

Gene list enrichments are identified in the following ontology categories: COVID, TRRUST, Transcription_Factor_Targets, Cell_Type_Signatures, DisGeNET, PaGenBase. All genes in the genome have been used as the enrichment background. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities. The top few enriched clusters (one term per cluster) are shown in the Figure 6-11. The algorithm used here is the same as that is used for pathway and process enrichment analysis.

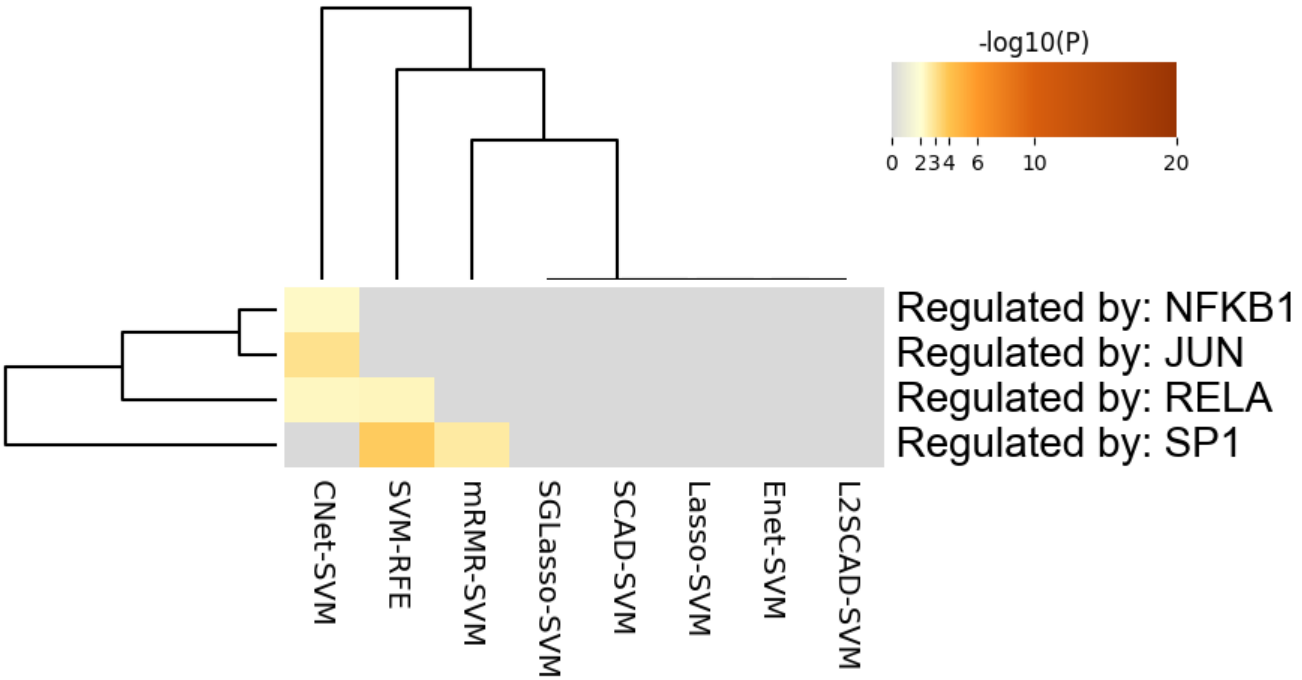
Figure 6. Summary of enrichment analysis in COVID[13].



(./Enrichment_QC/HeatmapSelectedGO_COVID.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	COVID054 (https://metascape.org/COVID#200_COVID054)	RNA_Xiong_PBMC_Up	6	21.00	-6.50	-3.40
	COVID042 (https://metascape.org/COVID#200_COVID042)	RNA_Sun_Calu-3_7h_Up	4	12.00	-6.20	-3.20
	COVID217 (https://metascape.org/COVID#200_COVID217)	Proteome_Shen_Serum-nonsevere-vs-healthy_Down	3	10.00	-5.80	-2.90
	COVID213 (https://metascape.org/COVID#200_COVID213)	Proteome_Shen_Serum-severe-vs-healthy_Down	3	10.00	-4.60	-2.10
	COVID046 (https://metascape.org/COVID#200_COVID046)	RNA_Wyler_Caco-2_24h_Up	3	9.40	-3.90	-1.60
	COVID134 (https://metascape.org/COVID#200_COVID134)	Proteome_Stukalov_A549-ACE2_24h_Down	4	14.00	-3.80	-1.60
	COVID049 (https://metascape.org/COVID#200_COVID049)	RNA_Wyler_Calu-3_24h_Down	4	14.00	-3.80	-1.60
	COVID048 (https://metascape.org/COVID#200_COVID048)	RNA_Wyler_Calu-3_12h_Up	3	11.00	-2.60	-0.83
	COVID057 (https://metascape.org/COVID#200_COVID057)	Phosphoproteome_Bouhaddou_Vero_E6_12h_Down	3	11.00	-2.60	-0.83
	COVID059 (https://metascape.org/COVID#200_COVID059)	Phosphoproteome_Bouhaddou_Vero_E6_24h_Down	3	11.00	-2.60	-0.83
	COVID010 (https://metascape.org/COVID#200_COVID010)	RNA_Blanco-Melo_A549-ACE2_Up	3	10.00	-2.50	-0.76
	COVID038 (https://metascape.org/COVID#200_COVID038)	RNA_Sun_Calu-3_12h_Up	3	10.00	-2.50	-0.76
	COVID193 (https://metascape.org/COVID#200_COVID193)	RNA_Stukalov_A549-ACE2_30h_Up	3	10.00	-2.50	-0.76
	COVID016 (https://metascape.org/COVID#200_COVID016)	RNA_Blanco-Melo_Calu-3_Up	3	9.40	-2.40	-0.71
	COVID036 (https://metascape.org/COVID#200_COVID036)	RNA_Sun_Calu-3_0h_Up	3	9.40	-2.40	-0.71
	COVID040 (https://metascape.org/COVID#200_COVID040)	RNA_Sun_Calu-3_24h_Up	3	9.40	-2.40	-0.71
	COVID050 (https://metascape.org/COVID#200_COVID050)	RNA_Wyler_Calu-3_24h_Up	3	9.40	-2.40	-0.71
	COVID243 (https://metascape.org/COVID#200_COVID243)	RNA_Riva_Vero-E6_24h_Up	3	9.40	-2.40	-0.71

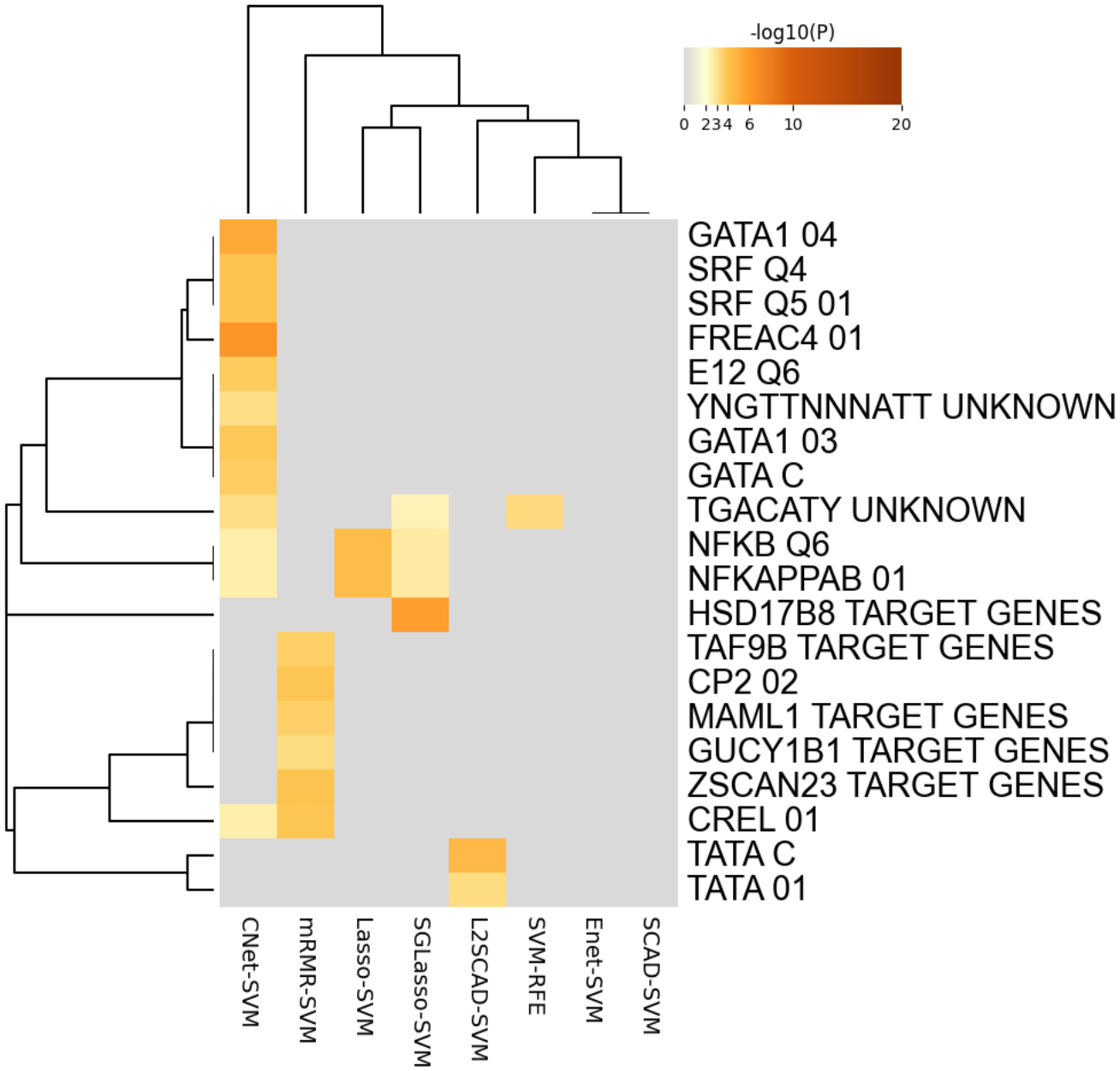
Figure 7. Summary of enrichment analysis in TRRUST.



(./Enrichment_QC/HeatmapSelectedGO_TRRUST.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	TRR01256	Regulated by: SP1	5	17.00	-3.80	-1.60
	TRR00645	Regulated by: JUN	3	9.40	-3.10	-1.20
	TRR01158	Regulated by: RELA	3	10.00	-2.40	-0.67
	TRR00875	Regulated by: NFKB1	3	9.40	-2.30	-0.59

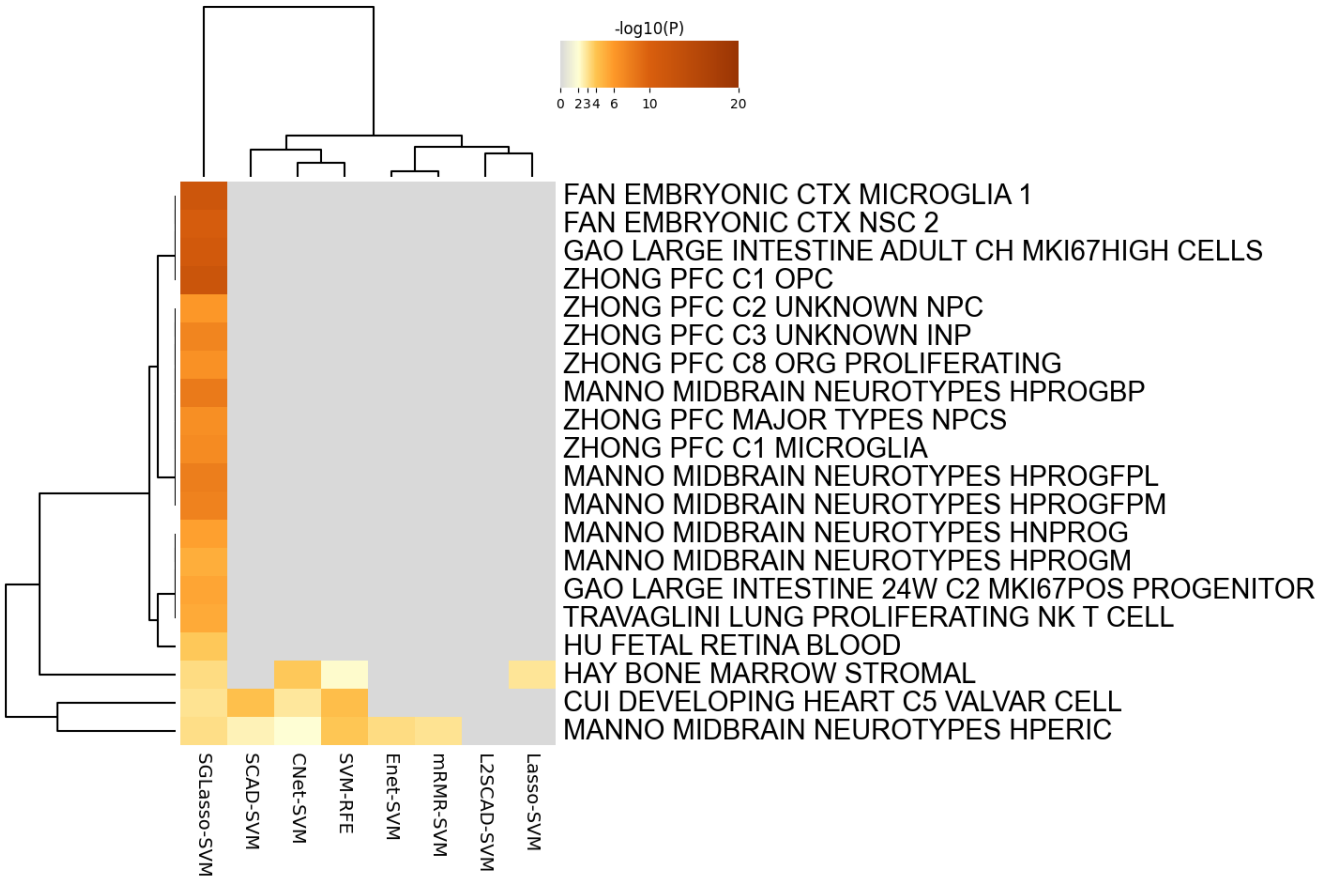
Figure 8. Summary of enrichment analysis in Transcription Factor Targets[14].



(./Enrichment_QC/HeatmapSelectedGO_Transcription_Factor_Targets.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	M14649 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M14649)	FREAC4 01	5	16.00	-6.20	-3.20
	M30019 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M30019)	HSD17B8 TARGET GENES	7	25.00	-5.70	-2.90
	M14012 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M14012)	GATA1 04	5	16.00	-5.20	-2.50
	M14357 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M14357)	TATA C	4	20.00	-4.50	-2.00
	M9949 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M9949)	NFKAPPAB 01	3	33.00	-4.30	-1.90
	M11921 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M11921)	NFKB Q6	3	33.00	-4.30	-1.90
	M30404 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M30404)	ZSCAN23 TARGET GENES	3	10.00	-4.10	-1.70
	M11934 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M11934)	SRF Q5 01	4	12.00	-4.00	-1.70
	M5479 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M5479)	SRF Q4	4	12.00	-4.00	-1.70
	M560 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M560)	CP2 02	4	13.00	-4.00	-1.70
	M10143 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M10143)	CREL 01	4	13.00	-3.90	-1.70
	M14186 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M14186)	GATA1 03	4	12.00	-3.90	-1.70
	M15183 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M15183)	E12 Q6	4	12.00	-3.80	-1.60
	M18920 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M18920)	GATA C	4	12.00	-3.70	-1.60
	M30190 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M30190)	TAF9B TARGET GENES	5	17.00	-3.70	-1.50
	M30054 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M30054)	MAML1 TARGET GENES	4	13.00	-3.60	-1.50
	M13849 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M13849)	TGACATY UNKNOWN	5	17.00	-3.30	-1.30
	M29986 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M29986)	GUCY1B1 TARGET GENES	5	17.00	-3.30	-1.30
	M3150 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M3150)	TATA 01	3	15.00	-3.20	-1.20
	M16291 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M16291)	YNGTTNNNATT UNKNOWN	4	12.00	-3.20	-1.20

Figure 9. Summary of enrichment analysis in Cell Type Signatures.

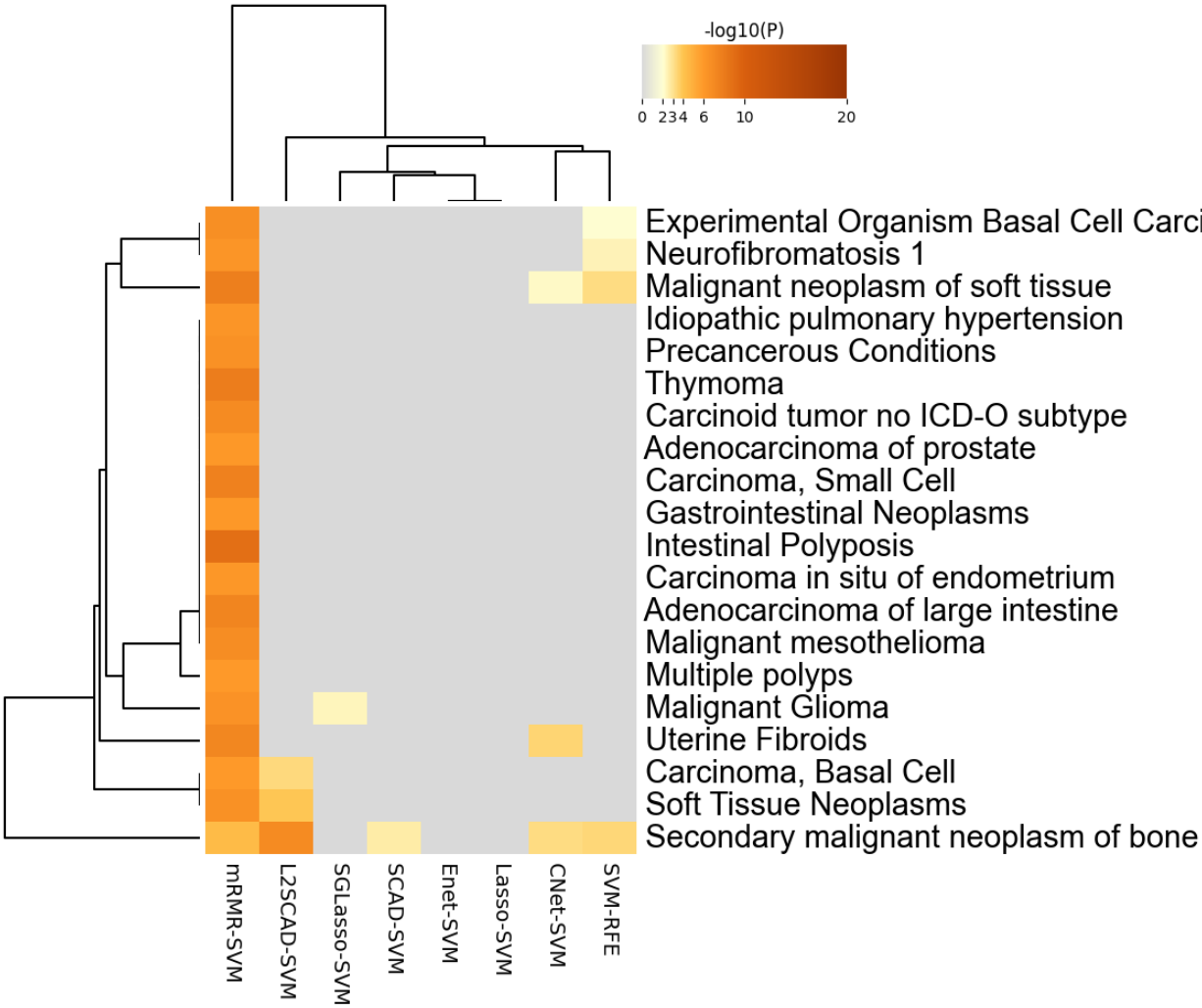


(./Enrichment_QC/HeatmapSelectedGO_Cell_Type_Signatures.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	M39096 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39096)	ZHONG PFC C1 OPC	9	32.00	-12.00	-7.70
	M39041 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39041)	FAN EMBRYONIC CTX MICROGLIA 1	8	29.00	-12.00	-7.70
	M39165 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39165)	GAO LARGE INTESTINE ADULT CH MKI67HIGH CELLS	7	25.00	-11.00	-7.20
	M39036 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39036)	FAN EMBRYONIC CTX NSC 2	8	29.00	-11.00	-6.60
	M39059 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39059)	MANNO MIDBRAIN NEUROTYPES HPROGBP	7	25.00	-8.10	-4.30
	M39060 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39060)	MANNO MIDBRAIN NEUROTYPES HPROGFPL	7	25.00	-7.80	-4.20
	M39061 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39061)	MANNO MIDBRAIN NEUROTYPES HPROGFPM	7	25.00	-7.50	-4.10
	M39083 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39083)	ZHONG PFC C3 UNKNOWN INP	4	14.00	-7.40	-4.00
	M39103 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39103)	ZHONG PFC C1 MICROGLIA	6	21.00	-6.90	-3.60
	M39078 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39078)	ZHONG PFC MAJOR TYPES NPCS	5	18.00	-6.70	-3.50

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	M39081 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39081)	ZHONG PFC C8 ORG PROLIFERATING	4	14.00	-6.50	-3.40
	M39087 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39087)	ZHONG PFC C2 UNKNOWN NPC	4	14.00	-6.10	-3.10
	M39062 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39062)	MANNO MIDBRAIN NEUROTYPES HNPROG	5	18.00	-5.70	-2.80
	M39153 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39153)	GAO LARGE INTESTINE 24W C2 MKI67POS PROGENITOR	4	14.00	-5.40	-2.60
	M41687 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M41687)	TRAVAGLINI LUNG PROLIFERATING NK T CELL	4	14.00	-5.20	-2.50
	M39058 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39058)	MANNO MIDBRAIN NEUROTYPES HPROGM	5	18.00	-5.10	-2.40
	M39302 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39302)	CUI DEVELOPING HEART C5 VALVAR CELL	4	13.00	-4.30	-1.90
	M39050 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39050)	MANNO MIDBRAIN NEUROTYPES HPERIC	6	20.00	-3.90	-1.70
	M39263 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39263)	HU FETAL RETINA BLOOD	4	14.00	-3.90	-1.70
	M39209 (https://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?systematicName=M39209)	HAY BONE MARROW STROMAL	6	19.00	-3.90	-1.70

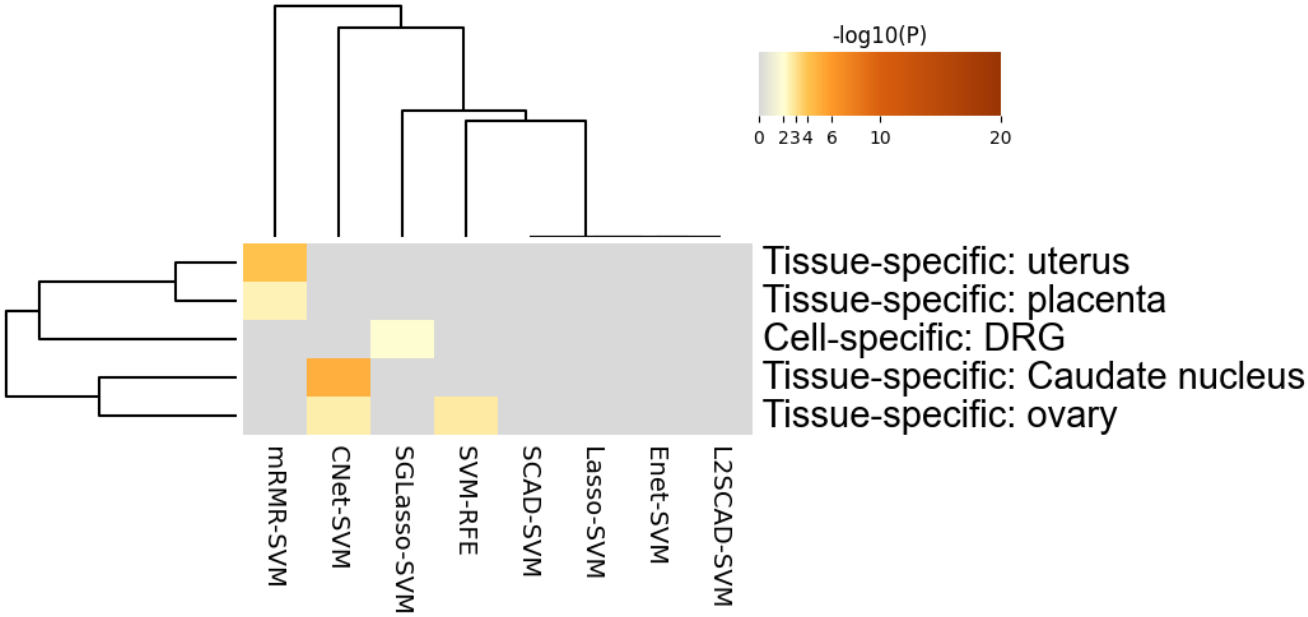
Figure 10. Summary of enrichment analysis in DisGeNET[15].



(./Enrichment_QC/HeatmapSelectedGO_DisGeNET.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	C1257915 (http://www.disgenet.org/browser/0/1/2/C1257915/)	Intestinal Polyposis	5	17.00	-8.90	-5.10
	C0040100 (http://www.disgenet.org/browser/0/1/2/C0040100/)	Thymoma	7	23.00	-7.90	-4.20
	C4551686 (http://www.disgenet.org/browser/0/1/2/C4551686/)	Malignant neoplasm of soft tissue	9	30.00	-7.80	-4.20
	C0262584 (http://www.disgenet.org/browser/0/1/2/C0262584/)	Carcinoma, Small Cell	6	20.00	-7.60	-4.10
	C1319315 (http://www.disgenet.org/browser/0/1/2/C1319315/)	Adenocarcinoma of large intestine	8	27.00	-7.40	-4.00
	C0042133 (http://www.disgenet.org/browser/0/1/2/C0042133/)	Uterine Fibroids	8	27.00	-7.20	-3.80
	C0153690 (http://www.disgenet.org/browser/0/1/2/C0153690/)	Secondary malignant neoplasm of bone	7	35.00	-6.90	-3.60
	C0334299 (http://www.disgenet.org/browser/0/1/2/C0334299/)	Carcinoid tumor no ICD-O subtype	5	17.00	-6.90	-3.60
	C0345967 (http://www.disgenet.org/browser/0/1/2/C0345967/)	Malignant mesothelioma	7	23.00	-6.80	-3.60
	C3811653 (http://www.disgenet.org/browser/0/1/2/C3811653/)	Experimental Organism Basal Cell Carcinoma	7	23.00	-6.70	-3.50
	C0037579 (http://www.disgenet.org/browser/0/1/2/C0037579/)	Soft Tissue Neoplasms	5	17.00	-6.50	-3.40
	C0032927 (http://www.disgenet.org/browser/0/1/2/C0032927/)	Precancerous Conditions	7	23.00	-6.50	-3.40
	C0555198 (http://www.disgenet.org/browser/0/1/2/C0555198/)	Malignant Glioma	8	27.00	-6.40	-3.30
	C0152171 (http://www.disgenet.org/browser/0/1/2/C0152171/)	Idiopathic pulmonary hypertension	5	17.00	-6.30	-3.20
	C0027831 (http://www.disgenet.org/browser/0/1/2/C0027831/)	Neurofibromatosis 1	6	20.00	-6.30	-3.20
	C0346191 (http://www.disgenet.org/browser/0/1/2/C0346191/)	Carcinoma in situ of endometrium	3	10.00	-6.10	-3.10
	C0017185 (http://www.disgenet.org/browser/0/1/2/C0017185/)	Gastrointestinal Neoplasms	5	17.00	-6.00	-3.00
	C0007112 (http://www.disgenet.org/browser/0/1/2/C0007112/)	Adenocarcinoma of prostate	6	20.00	-6.00	-3.00
	C4721806 (http://www.disgenet.org/browser/0/1/2/C4721806/)	Carcinoma, Basal Cell	7	23.00	-6.00	-3.00
	C0334108 (http://www.disgenet.org/browser/0/1/2/C0334108/)	Multiple polyps	4	13.00	-6.00	-3.00

Figure 11. Summary of enrichment analysis in PaGenBase[16].



(./Enrichment_QC/HeatmapSelectedGO_PaGenBase.pdf)

PATTERN	GO	Description	Count	%	Log10(P)	Log10(q)
	PGB:00079	Tissue-specific: Caudate nucleus	3	9.40	-5.00	-2.30
	PGB:00072	Tissue-specific: uterus	3	10.00	-4.10	-1.80
	PGB:00057	Tissue-specific: ovary	3	10.00	-2.70	-0.92
	PGB:00045	Tissue-specific: placenta	3	10.00	-2.50	-0.75
	PGB:00014	Cell-specific: DRG	3	11.00	-2.10	-0.44

Reference

1. Zhou et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications (2019) 10(1):1523.
2. Saldanha AJ. Java Treeview - extensible visualization of microarray data. Bioinformatics (2004) 20:3246-3248
3. Krzywinski M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645
4. Zar, J.H. Biostatistical Analysis 1999 4th edn., NJ Prentice Hall, pp. 523
5. Hochberg Y., Benjamini Y. More powerful procedures for multiple significance testing. Statistics in Medicine (1990) 9:811-818.
6. Cohen, J. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. (1960) 20:27-46.
7. Shannon P. et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res (2003) 11:2498-2504.
8. Szklarczyk D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. (2019) 47:D607-613.
9. Stark C. et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. (2006) 34:D535-539.
10. Turei D. et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods. (2016) 13:966-967.
11. Li T. et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods. (2017) 14:61-64.
12. Bader, G.D. et al. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics (2003) 4:2.
13. https://metascape.org/COVID.
14. Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545-15550 (2005).
15. Pinero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research 45, D833-D839 (2017).
16. Pan JB, et al. PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. PLoS One 8, e80747 (2013).