Bioinformatics@Data Science A.Y. 2019-2020

Network Medicine project

Simone Faricelli, Lorenzo Germano

Group no. 09

Abstract

In this project we analyzed the genes relatives to the human disease Diabetes Mellitus. We analyzed the structure of the network generated by their connections and extracted graph and information from two different databases. We also made use of tools called Erichr for the Gene Ontlogies and Pathways extraction. We clustered our Largest Connected Component in order to obtain Putative Disease Modules and applied hypergeometric test to calculate their pvalue. We finally used the tool DIAMOnD to compare our results.

Basic introduction about the disease/process

Diabetes mellitus: More commonly referred to as "diabetes", a chronic disease associated with abnormally high levels of the sugar glucose in the blood.
Diabetes is due to one of two mechanisms: Inadequate production of insulin (which is made by the pancreas and lowers blood glucose), or Inadequate sensitivity of cells to the action of insulin.
The two main types of diabetes correspond to these two mechanisms and are called insulin dependent (type 1) and non-insulin dependent (type 2) diabetes. In type 1 diabetes there is no insulin or not enough of it. In type 2 diabetes, there is generally enough insulin but the cells upon which it should act are not normally sensitive to its action. Both are caused by a combination of genetic and environmental risk factors.
The signs and symptoms of both types of diabetes include increased urine output and decreased appetite as well as fatigue. Diabetes is diagnosed by blood glucose testing, the glucose tolerance test, and testing of the level of glycosylated hemoglobin (glycohemoglobin or hemoglobin A1C). The mode of treatment depends on the type of the diabetes.
The major complications of diabetes include dangerously elevated blood sugar, abnormally low blood sugar due to diabetes medications, and disease of the blood vessels which can damage the eyes, kidneys, nerves, and heart.

Seed genes

We started from the DisGeNet dataset, looking for every gene involved in the disease of Diabetes mellitus regarding the Homo Sapiens. We manually downloaded the whole dataset from the site and stored in the following table all the information about the above-mentioned genes using Matlab and selecting only the genes with the corresponding pathology code of Diabetes ('C0011853').

Besides the gene symbols, we stored also the corresponding Gene Entrez ID, UniProt ID, Protein Name and (only inside the "**Data Table Gene Seed.xls**" file) also related gene description, by requesting the additional information with a GET call to the "Hugo Gene Nomenclature Committee" site.

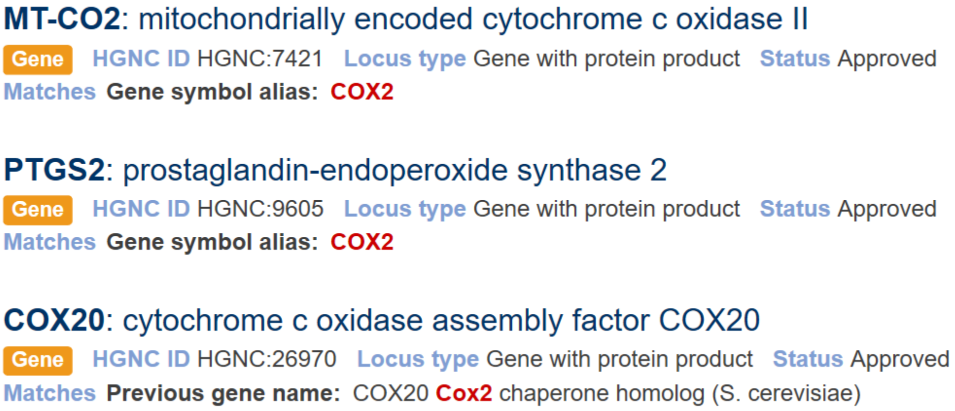
GENE_SYMBOL	GENE_ID	UNIPROT_ID	PROTEIN_NAME
ACOX1	51	Q15067	acyl-CoA oxidase 1
ADRA1A	148	P35348	adrenoceptor alpha 1A
ADRB3	155	P13945	adrenoceptor beta 3
AGT	183	P01019	angiotensinogen
FAS	355	P25445	Fas cell surface death receptor
STS	412	P08842	steroid sulfatase
ATF3	467	P18847	activating transcription factor 3
ATP2A2	488	P16615	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 2
ATP2A3	489	Q93084	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 3
BAX	581	Q07812	BCL2 associated X, apoptosis regulator
BCL2	596	P10415	BCL2 apoptosis regulator
BCL2L1	598	Q07817	BCL2 like 1
BDKRB1	623	P46663	bradykinin receptor B1
CASP3	836	P42574	caspase 3
CAT	847	P04040	catalase
CAV1	857	Q03135	caveolin 1
CAV3	859	P56539	caveolin 3
CD68	968	P34810	CD68 molecule
CHRM2	1129	P08172	cholinergic receptor muscarinic 2
CPT1A	1374	P50416	carnitine palmitoyltransferase 1A
CPT1B	1375	Q92523	carnitine palmitoyltransferase 1B
CYBA	1535	P13498	cytochrome b-245 alpha chain
CYBB	1536	P04839	cytochrome b-245 beta chain
CYP1A1	1543	P04798	cytochrome P450 family 1 subfamily A member 1
ACE	1636	P12821	angiotensin I converting enzyme
NQO1	1728	P15559	NAD(P)H quinone dehydrogenase 1
EDN1	1906	P05305	endothelin 1
ESRRA	2101	P11474	estrogen related receptor alpha
ACSL1	2180	P33121	acyl-CoA synthetase long chain family member 1
FOXO3	2309	O43524	forkhead box O3
GCK	2645	P35557	glucokinase
GPD2	2820	P43304	glycerol-3-phosphate dehydrogenase 2
GPX1	2876	P07203	glutathione peroxidase 1
GSR	2936	P00390	glutathione-disulfide reductase

HK1	3098	P19367	hexokinase 1
HMOX1	3162	P09601	heme oxygenase 1
HSD11B1	3290	P28845	hydroxysteroid 11-beta dehydrogenase 1
IAPP	3375	P10997	islet amyloid polypeptide
ICAM1	3383	P05362	intercellular adhesion molecule 1
ID1	3397	P41134	inhibitor of DNA binding 1, HLH protein
IFNG	3458	P01579	interferon gamma
IGF1	3479	P05019	insulin like growth factor 1
IL1B	3553	P01584	interleukin 1 beta
IL6	3569	P05231	interleukin 6
INSR	3643	P06213	insulin receptor
PDX1	3651	P52945	pancreatic and duodenal homeobox 1
IRS1	3667	P35568	insulin receptor substrate 1
KCNJ11	3767	Q14654	potassium inwardly rectifying channel subfamily J member 11
LEP	3952	P41159	leptin
LEPR	3953	P48357	leptin receptor
MAP3K5	4217	Q99683	mitogen-activated protein kinase kinase kinase 5
MFGE8	4240	Q08431	milk fat globule EGF and factor V/VIII domain containing
MMP2	4313	P08253	matrix metallopeptidase 2
MMP9	4318	P14780	matrix metallopeptidase 9
MPO	4353	P05164	myeloperoxidase
COX2	4513	null	null
ND1	4535	null	null
NEUROD1	4760	Q13562	neuronal differentiation 1
NKX6-1	4825	P78426	NK6 homeobox 1
NOS2	4843	P35228	nitric oxide synthase 2
NOS3	4846	P29474	nitric oxide synthase 3
SERPINE1	5054	P05121	serpin family E member 1
PAX6	5080	P26367	paired box 6
PCK1	5105	P35558	phosphoenolpyruvate carboxykinase 1
PCSK2	5126	P16519	proprotein convertase subtilisin/kexin type 2
PDK4	5166	Q16654	pyruvate dehydrogenase kinase 4
PFKM	5213	P08237	phosphofructokinase, muscle
PKLR	5313	P30613	pyruvate kinase L/R
PPARA	5465	Q07869	peroxisome proliferator activated receptor alpha
PPARG	5468	P37231	peroxisome proliferator activated receptor gamma
PRKCA	5578	P17252	protein kinase C alpha
PRKCD	5580	Q05655	protein kinase C delta
PRKCE	5581	Q02156	protein kinase C epsilon
PTGS2	5743	P35354	prostaglandin-endoperoxide synthase 2

RELA	5970	Q04206	RELA proto-oncogene, NF-kB subunit
REN	5972	P00797	renin
S100A6	6277	P06703	S100 calcium binding protein A6
CCL20	6364	P78556	C-C motif chemokine ligand 20
CX3CL1	6376	P78423	C-X3-C motif chemokine ligand 1
SLC2A2	6514	P11168	solute carrier family 2 member 2
SLC2A4	6517	P14672	solute carrier family 2 member 4
SLC9A1	6548	P19634	solute carrier family 9 member A1
SLC9A3	6550	P48764	solute carrier family 9 member A3
SNAP25	6616	P60880	synaptosome associated protein 25
SOD1	6647	P00441	superoxide dismutase 1
SOD2	6648	P04179	superoxide dismutase 2
SREBF1	6720	P36956	sterol regulatory element binding transcription factor 1
TGFB1	7040	P01137	transforming growth factor beta 1
TIMP1	7076	P01033	TIMP metallopeptidase inhibitor 1
TIMP2	7077	P16035	TIMP metallopeptidase inhibitor 2
TNF	7124	P01375	tumor necrosis factor
TNFRSF1A	7132	P19438	TNF receptor superfamily member 1A
TP53	7157	P04637	tumor protein p53
UCP2	7351	P55851	uncoupling protein 2
VEGFA	7422	P15692	vascular endothelial growth factor A
YWHAH	7533	Q04917	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta
AOC3	8639	Q16853	amine oxidase copper containing 3
IRS2	8660	Q9Y4H2	insulin receptor substrate 2
S1PR4	8698	O95977	sphingosine-1-phosphate receptor 4
AIFM1	9131	O95831	apoptosis inducing factor mitochondria associated 1
S1PR2	9294	O95136	sphingosine-1-phosphate receptor 2
PPARGC1A	10891	Q9UBK2	PPARG coactivator 1 alpha
SIRT1	23411	Q96EB6	sirtuin 1
FGF21	26291	Q9NSA1	fibroblast growth factor 21
S1PR5	53637	Q9H228	sphingosine-1-phosphate receptor 5
GPAM	57678	Q9HCL2	glycerol-3-phosphate acyltransferase, mitochondrial
ACOT1	641371	Q86TX2	acyl-CoA thioesterase 1
NCF1	653361	P14598	neutrophil cytosolic factor 1

Tab.1: Seed Genes Information

Notes: As noticeable in the above table, the genes named “COX2” and “ND1” found no match in the HUGO dataset. So, we manually check for the correspondence found in the HUGO Dataset and we show the results in the following images.

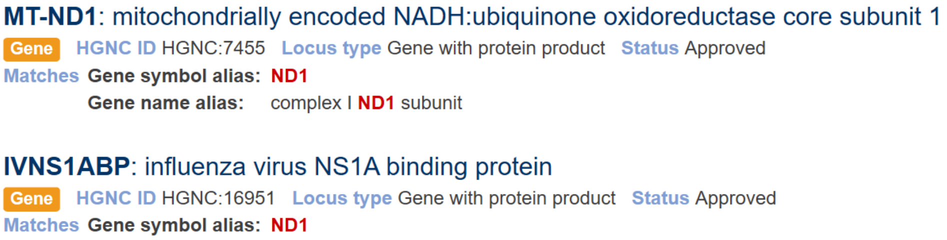


MT-CO2: mitochondrially encoded cytochrome c oxidase II
Gene HGNC ID HGNC:7421 Locus type Gene with protein product Status Approved
Matches **Gene symbol alias:** **COX2**

PTGS2: prostaglandin-endoperoxide synthase 2
Gene HGNC ID HGNC:9605 Locus type Gene with protein product Status Approved
Matches **Gene symbol alias:** **COX2**

COX20: cytochrome c oxidase assembly factor COX20
Gene HGNC ID HGNC:26970 Locus type Gene with protein product Status Approved
Matches **Previous gene name:** COX20 **Cox2** chaperone homolog (S. cerevisiae)

Fig. 1: COX2 results from HUGO



MT-ND1: mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 1
Gene HGNC ID HGNC:7455 Locus type Gene with protein product Status Approved
Matches **Gene symbol alias:** **ND1**
Gene name alias: complex I **ND1** subunit

IVNS1ABP: influenza virus NS1A binding protein
Gene HGNC ID HGNC:16951 Locus type Gene with protein product Status Approved
Matches **Gene symbol alias:** **ND1**

Fig. 2: ND1 results from HUGO

Summary on interaction data

In order to collect always updated information about the interactions between seed genes and non-seed genes connected to them we chose to use their REST site service to make in-code API call, requesting to their database all the interaction needed. Even if just one request, with the whole list of gene seeds, would have fulfilled the demand of point 1.2.a in the project requests, we decided to make multiple requests, asking for the connection of just one seed per-call and being careful to consider only Homo Sapiens genes interactions (Tax ID code: 9606). That’s because the GET request to the site can only answer with a json file of maximum 10.000 interactions. So, we obtained all the interaction for each gene and their corresponding interactome and put everything together in an excel file called “**Biogrid Interaction Table_1.2.xls**”. We ended up with a total number of 29461 interactions. The same process was performed with the second dataset, the one called Integrated Interactions Database, and like the previous one, all the interactions are stored in the excel file called

“**IID Integrated Interactions Database_1.2**”, in which can be found a total number of 115558 interactions, due to excel tabs limitations, placed into 2 columns.

The following table shows the summary of the interactions for the two datasets.

Database	*Genes_Found_in_DB*	*Proteins_Interacting*	*Interactions*
BioGrid Human	106	4012	29460
Integrated Interactions Database	106	4314	115558

Tab. 2: Summary Interactions

Interactomes data

For a matter of dimensions we started from the IID interactions data, which contains more PPI than the other one, storing in a .mat file a table variable which contains 5 columns with respectively “*interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC, interactor B Uniprot AC*” and “*database source*” which actually contains only “IID” string. After that we opened the set of interactions obtained from Biogrid Human and, interaction by interaction, we checked if the couple Symbol_A-Symbol_B or Symbol_B-Symbol_A already appears in the above-mentioned table. When the answer was positive, we just added the string “Biogrid” to the last column, otherwise we added a new row with the interaction and its corresponding database source.
We applied this method of analysis on interactions between only seed genes and store the results inside the file “**Seed Gene Interactome.mat**”, and also on interactions between genes in which at least one of them is a seed gene, considering only Intersection between the two databases (“**Intersection Interactome.mat**”) and considering only the union of the two (“**Union Interactome.mat**”).

Enrichment analysis

Due to the lack of useful API call to the Enrichr server site, we manually copied the list of seed genes and the list of union interactome genes, carried-out by the previous step. The first 10 results are showed in the graphs below, sorted by pvalue ranking, which is nominally computed from the Fisher exact test, which is a proportion test that assumes a binomial distribution and independence for probability of any gene belonging to any set.

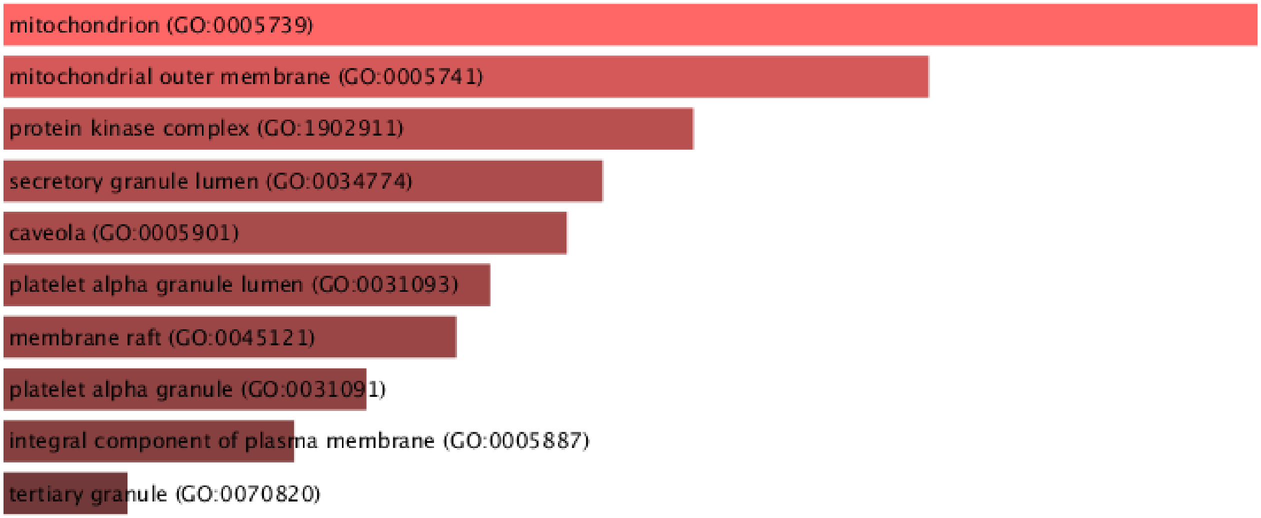


Fig. 3: The GO Biological Process graph of seed genes

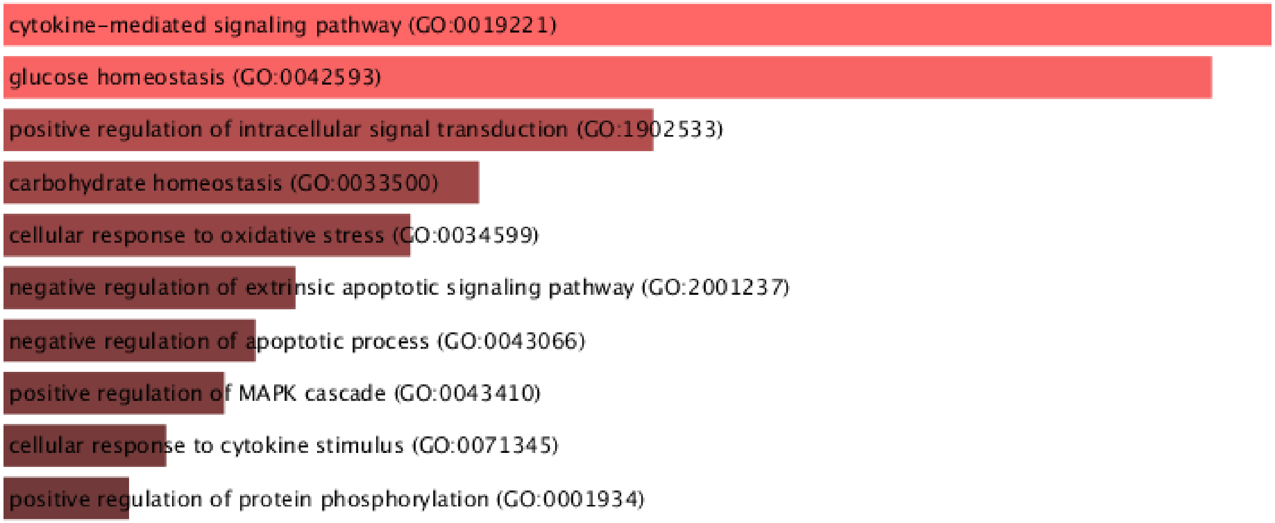


Fig. 4: The GO Cellular Component graph of seed genes

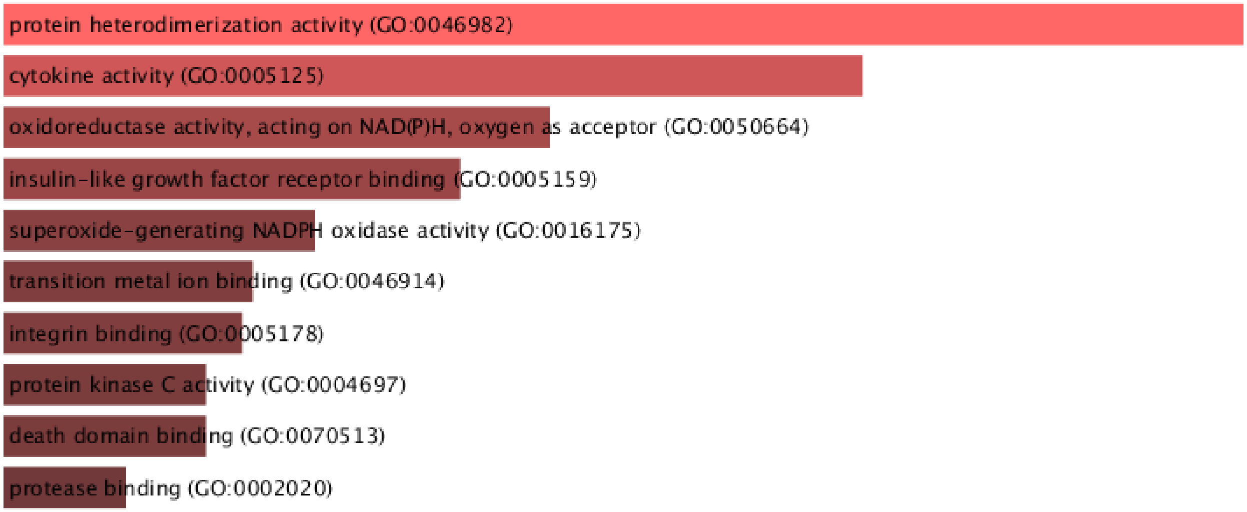


Fig. 5: The GO Molecular Function graph of seed genes

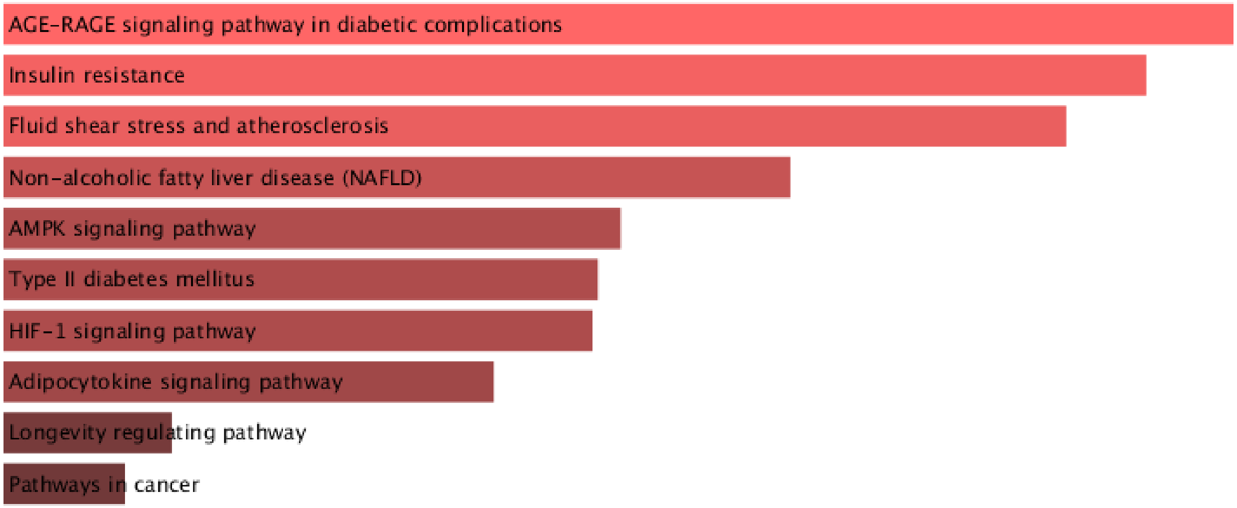


Fig. 6: The overrepresented pathways from KEGG 2019 Human graph of seed genes

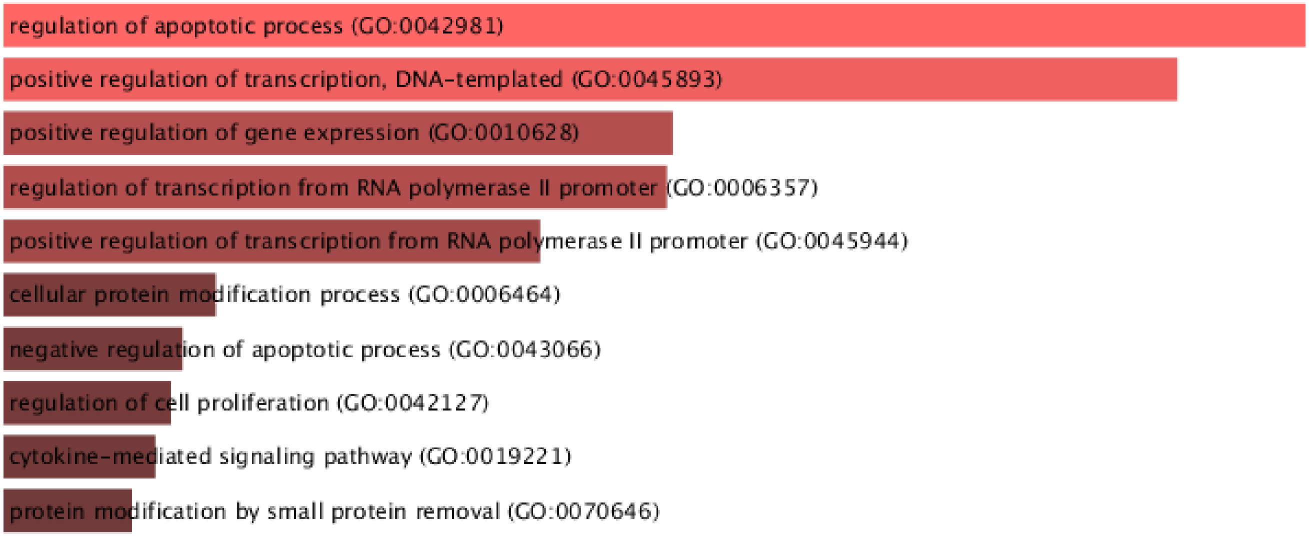


Fig. 7: The GO Biological Process graph of union interactome genes

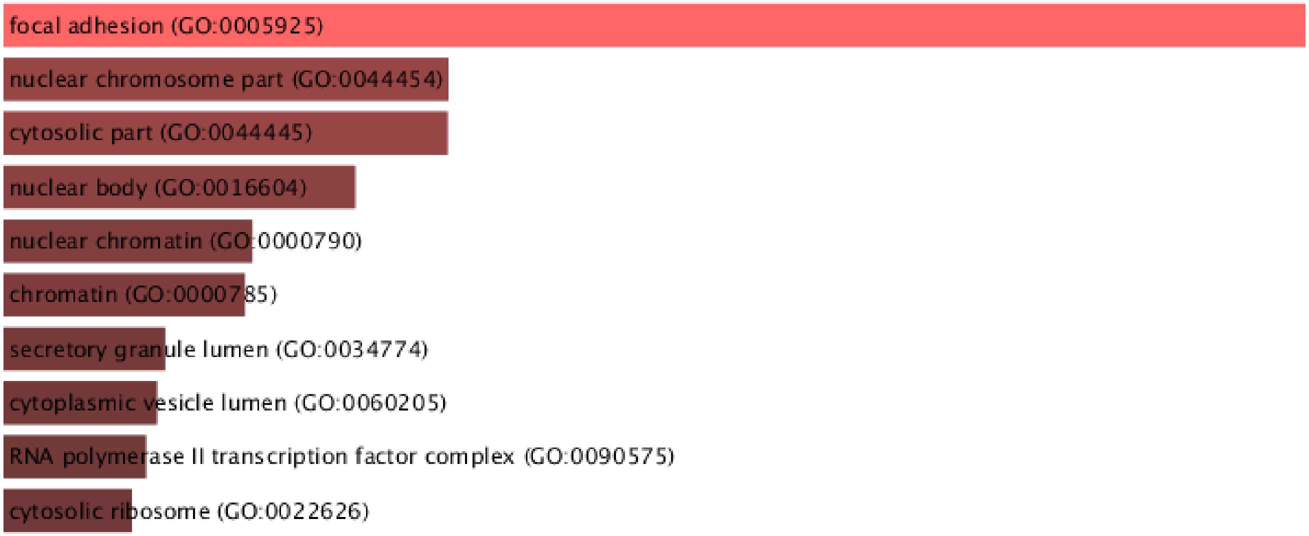


Fig. 8: The GO Cellular Component graph of union interactome genes

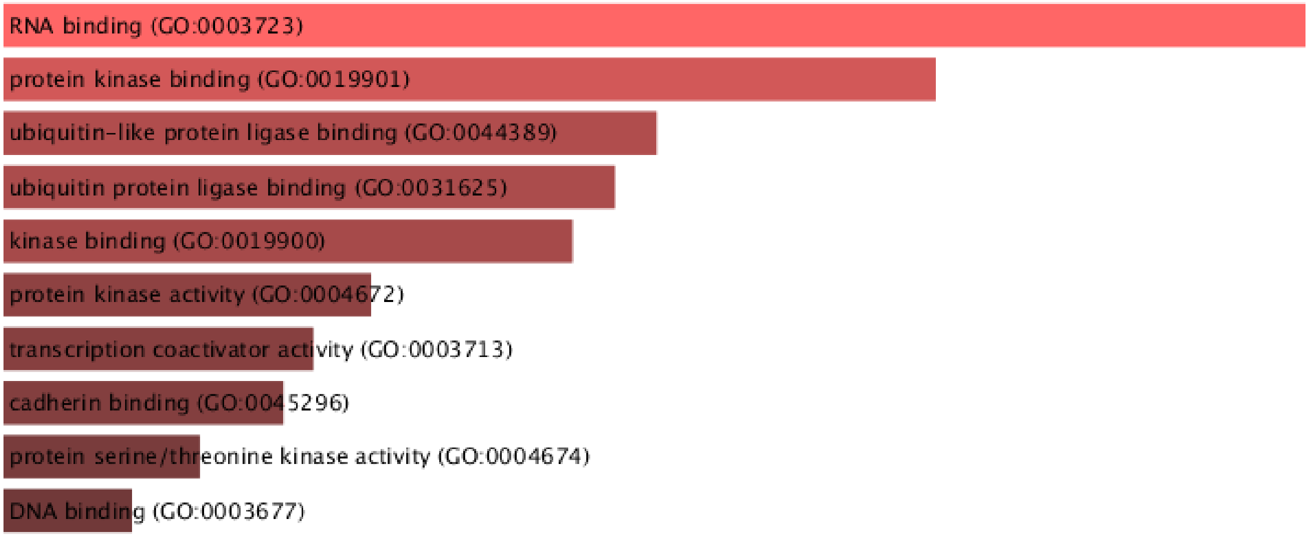


Fig. 9: The GO Molecular Function graph of union interactome genes

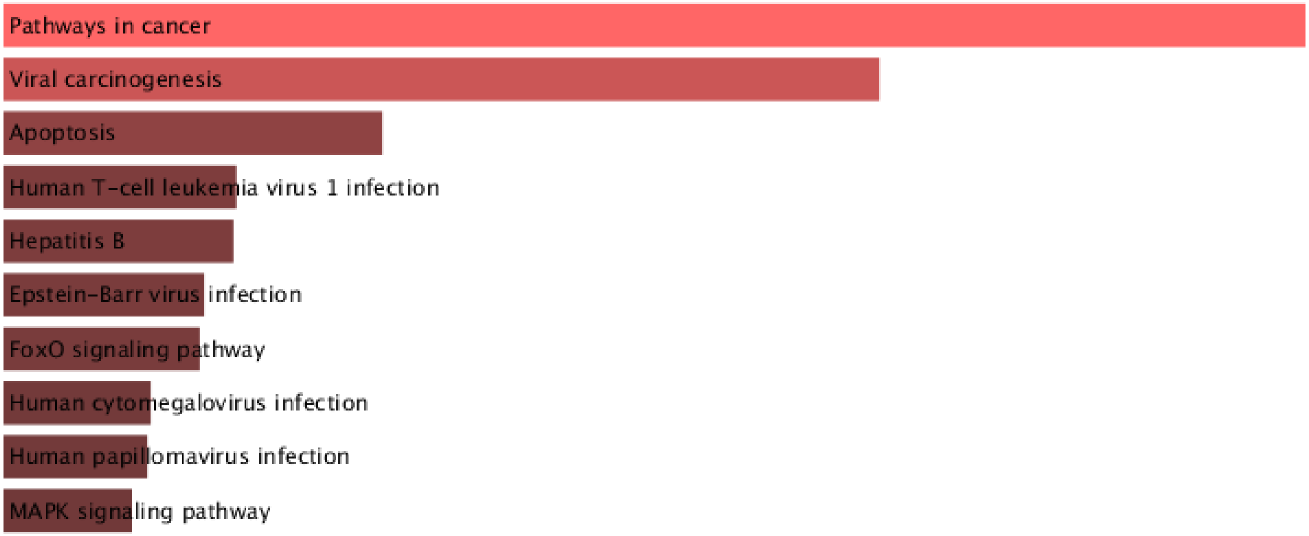


Fig. 10: The overrepresented pathways from KEGG 2019 Human graph of union interactome genes

Network analysis

After generating the tables of the interactions mentioned in the previous step, we can now build each Graph with the help of the so-called built-in MATLAB class and compute the main network-measures, which results are showed in table below, through the use of methods and properties provided by above-mentioned MATLAB class. Listed below, we also show the results of an adjustments of the graph, meaning that, through the use of the method *conncomp()*, we extracted and analyzed the Largest Connected Component (which luckily was also the only one present in the graph after discarding connected components with a number of nodes less than 20).

NETWORK	*No. of nodes*	*No. of links*	*No. of connected components*	*No. of isolated nodes*	*Average path length*	*Average degree*	*Average clustering coefficient*	*Network diameter*	*Network radius*	*Centralization*
Seed Genes Interactome (SGI)	88	219	1	12	3,2211	5,6389	1,0628	8	4	0,2489
Union (U)	4314	8378	1	0	3,5853	3,8841	1,1811	8	4	0,2542
Intersection (I)	2919	4382	1	28	4,0548	2,9338	0,8649	8	5	0,1027
Intersection largest connected component (I-LCC)	2891	4362	None	None	4,0548	2,9338	0,8649	8	5	0,1027
Union largest connected component (U-LCC)	4314	8378	None	None	3,5853	3,8841	1,1811	8	4	0,2542

Tab. 3: Global Measures of Networks

About local measures, we stored the arrays containing "*Node degree, Betweenness centrality, Eigenvector centrality, Closeness centrality, ratio Betweenness/Node degree*", respectively in the file named **x_node_degree.mat**, **x_betweenness.mat**, **x_eigenvector.mat**, **x_closeness.mat** and **x_ratio.mat** where x has to be replaced by "i" for the intersection or "u" for the union. All the measures have been obtained with the method *centrality('method')* provided by the class *Graph*.

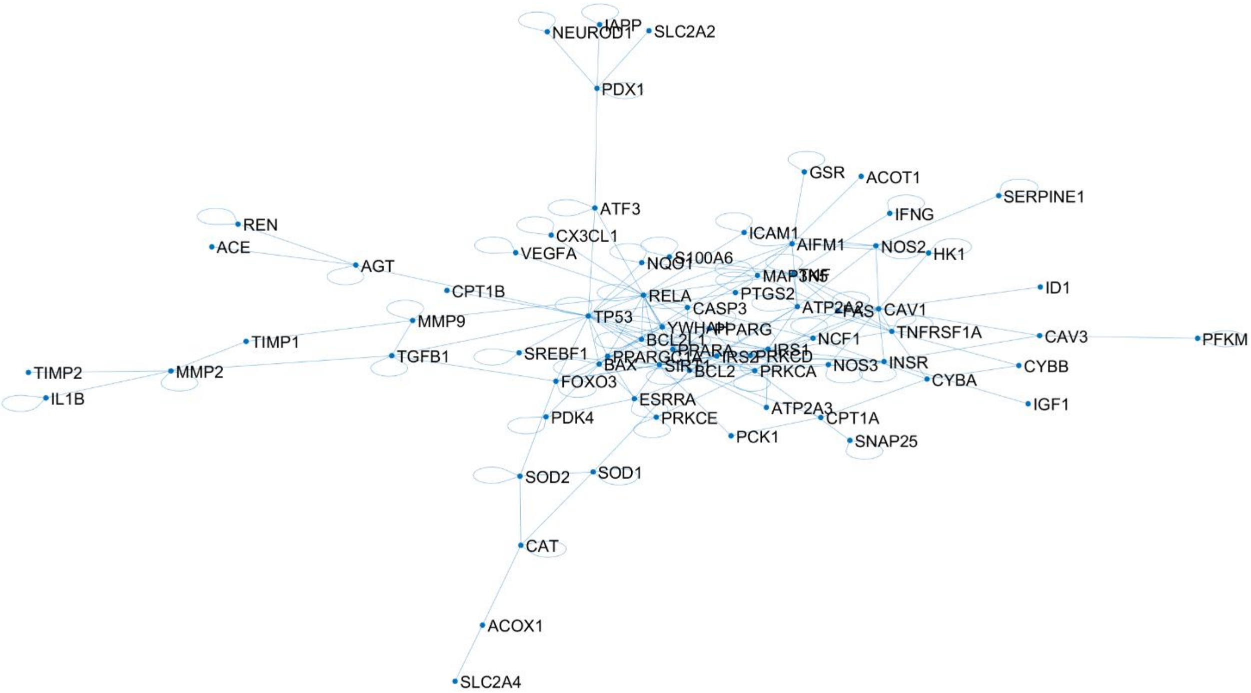


Fig. 11: SGI Network

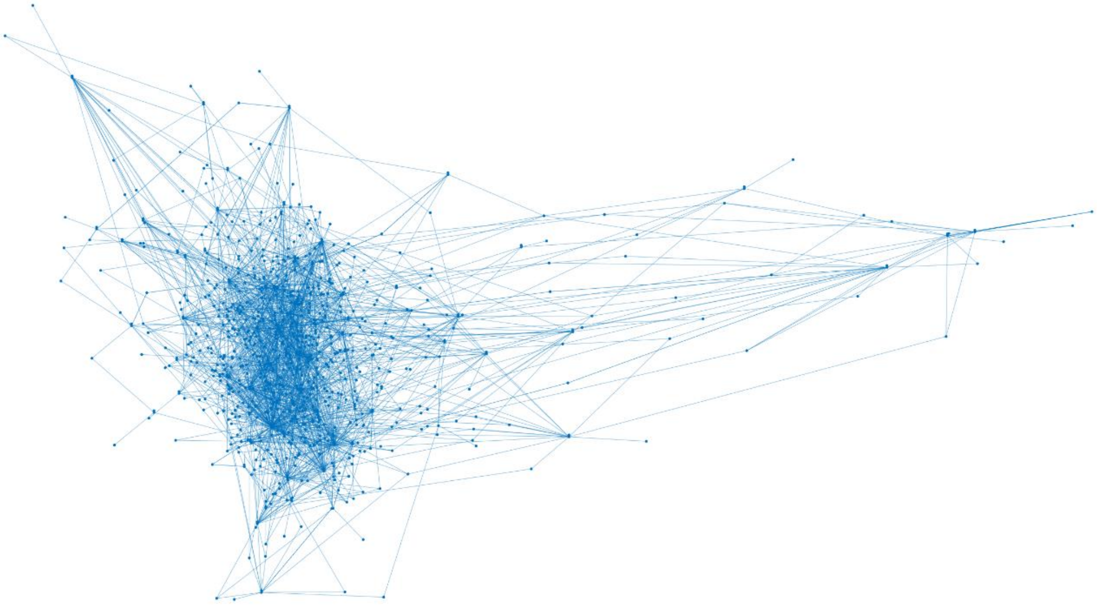
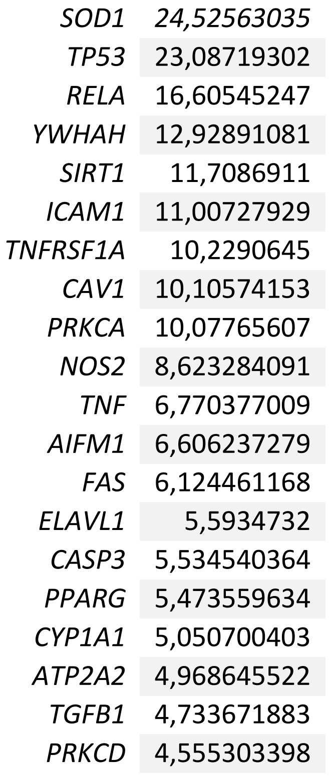
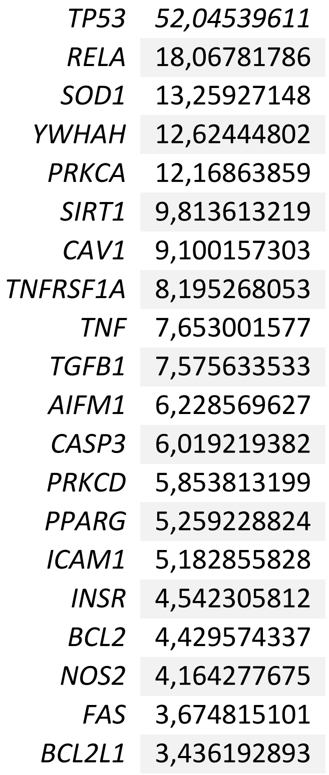


Fig. 12: I-LCC Network



SOD1	*24,52563035*
TP53	23,08719302
RELA	16,60545247
YWHAH	12,92891081
SIRT1	11,7086911
ICAM1	11,00727929
TNFRSF1A	10,2290645
CAV1	10,10574153
PRKCA	10,07765607
NOS2	8,623284091
TNF	6,770377009
AIFM1	6,606237279
FAS	6,124461168
ELAVL1	5,5934732
CASP3	5,534540364
PPARG	5,473559634
CYP1A1	5,050700403
ATP2A2	4,968645522
TGFB1	4,733671883
PRKCD	4,555303398

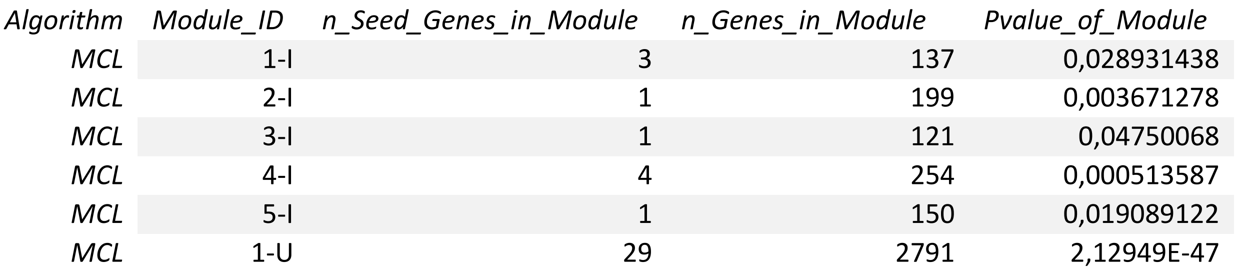
Tab. 4: First 20 highest ranking genes for betweenness for I-LCC



TP53	*52,04539611*
RELA	18,06781786
SOD1	13,25927148
YWHAH	12,62444802
PRKCA	12,16863859
SIRT1	9,813613219
CAV1	9,100157303
TNFRSF1A	8,195268053
TNF	7,653001577
TGFB1	7,575633533
AIFM1	6,228569627
CASP3	6,019219382
PRKCD	5,853813199
PPARG	5,259228824
ICAM1	5,182855828
INSR	4,542305812
BCL2	4,429574337
NOS2	4,164277675
FAS	3,674815101
BCL2L1	3,436192893

Tab. 5: First 20 highest ranking genes for betweenness for U-LCC

We then divide the LCC into sub-clusters for disease modules discovery, using Markov Clustering Alghoritm, through the use of a third-party library for MATLAB which exploits Dijkstra Alghoritms to calculate them. We finally calculate the pvalue of each sub-cluster using the hypergeometric test thanks to the built-in MATLAB function *hygepdf().*



Algorithm	*Module_ID*	*n_Seed_Genes_in_Module*	*n_Genes_in_Module*	*Pvalue_of_Module*
MCL	1-I	3	137	0,028931438
MCL	2-I	1	199	0,003671278
MCL	3-I	1	121	0,04750068
MCL	4-I	4	254	0,000513587
MCL	5-I	1	150	0,019089122
MCL	1-U	29	2791	2,12949E-47

Tab. 6: Putative Disease Modules Summary

The rest of information needed, as the list of genes mentioned in the table above, are stored in the matlab table called “**putative_disease_modules_table.mat**”.

The Enrichr analysis for each putative disease module is stored in the folder called “**Part2.3**” as .png images.

DIAMOnD Tool

The following list contains the first 30 genes identified by DIAMOnD Tool from the Putative Disease Proteins using as reference the whole Biogrid Human interactome dataset, already used to collect PPIs.

'COL2A1'
'GNA12'
'ECT2'
'COL12A1'
'MOAP1'
'MAFA'
'CFLAR'
'TRAF3'
'ACTBL2'
'LIG4'
'UFL1'
'MTDH'
'COL6A1'
'RNF123'
'THBS1'
'COL5A1'
'COL4A1'
'COL4A2'
'FYN'
'NEDD4'
'JAK2'
'CSK'
'AHSG'
'MAD2L1'
'EPHB2'
'UBQLN1'
'HNRNPL'
'GRB10'
'ARRB1'
'IDE'

The Enrichr analyses for the 200 newly found genes resulting from DIAMOnD Tool usage, are showed below.

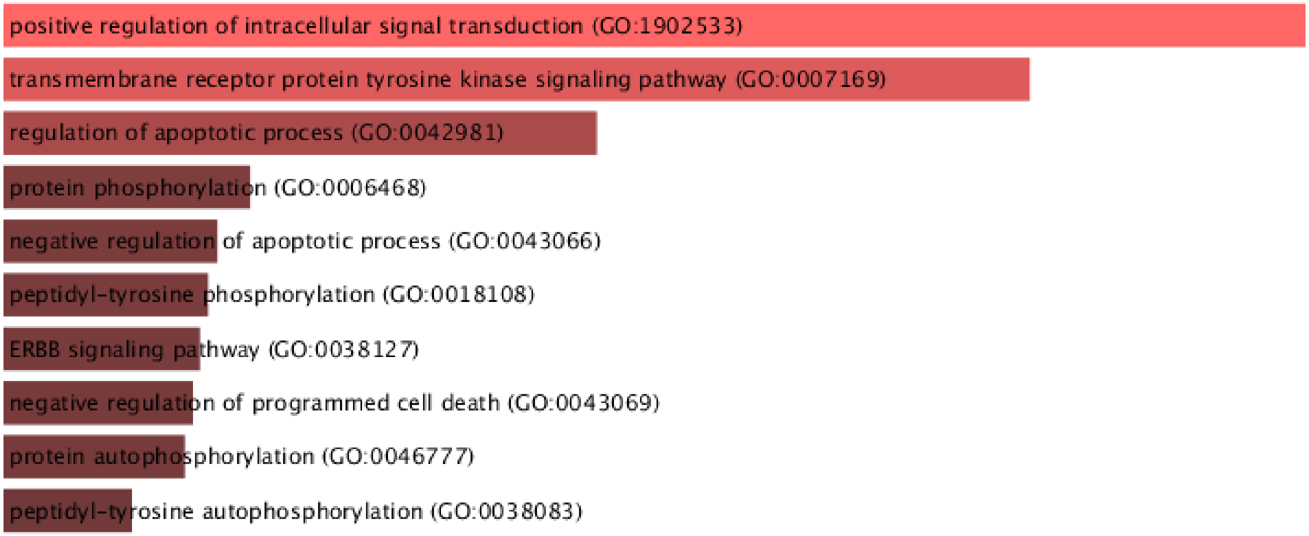


Fig. 13: The GO Biological Process graph of 200 newly genes

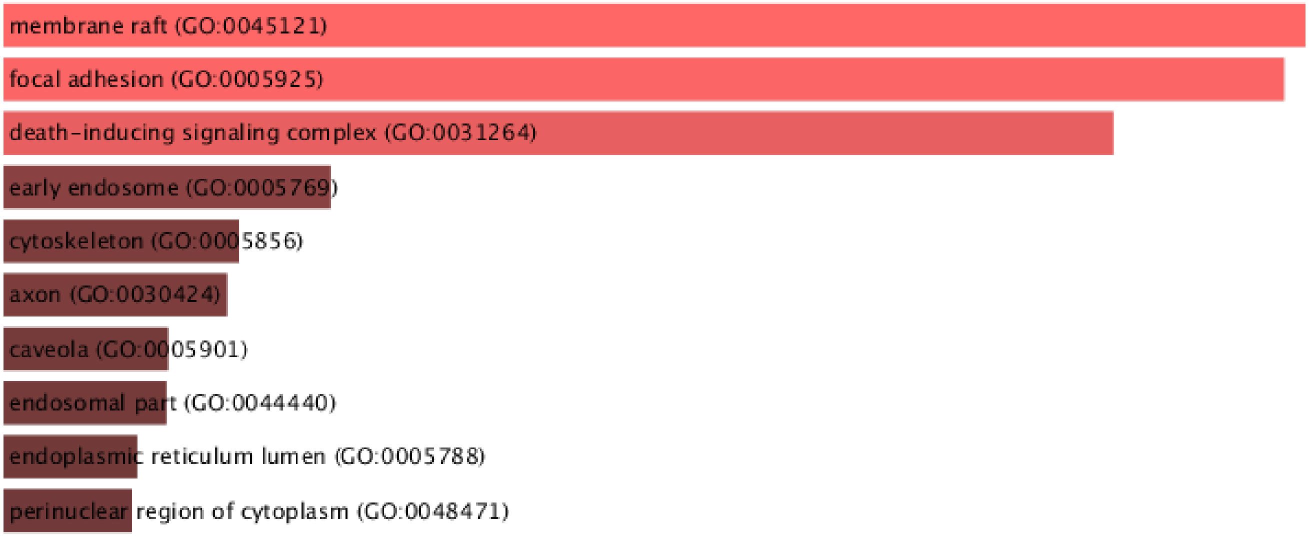


Fig. 14: The GO Cellular Component graph of 200 newly genes

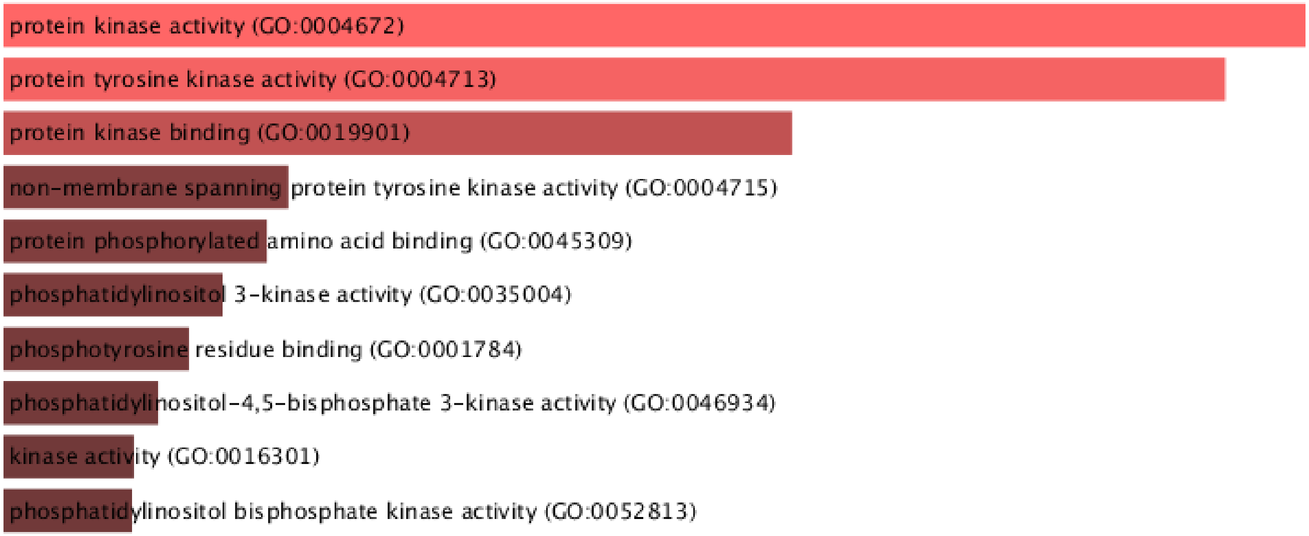


Fig. 15: The GO Molecular Function graph of 200 newly genes

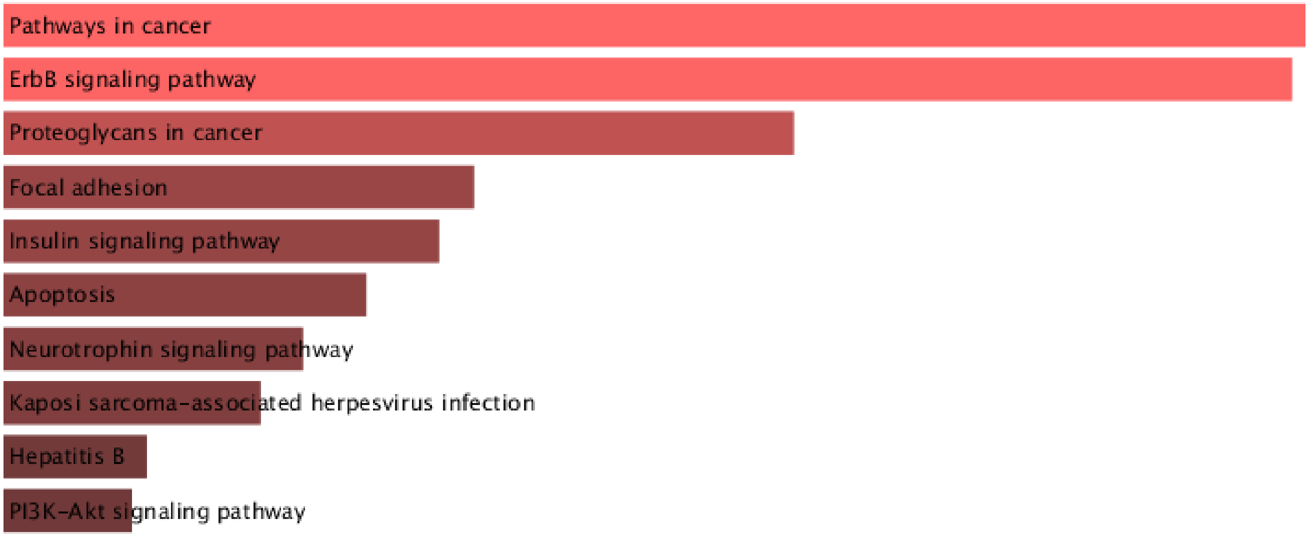


Fig. 16: The overrepresented pathways from KEGG 2019 Human graph of 200 newly genes