Bioinformatics Report HW2 Group 5

Rheumatoid arthritis-related human-oral microbiome proteins

Giorgio Giannone[1], Livia Lombardi[1]

Abstract

The study of the biological network and its related complex dynamics is crucial to better understand human diseases. A disease represents not only a malfunction of a single gene but especially a consequence of a inter-cellular network distortion. Network science simplify the proteins interaction complexity to only two elements: the components/nodes and the interactions/edges. Using this model, network information can be gathered in order to examine the underlying disease mechanism. For this reason the study of network properties and characteristics became interesting as the reflection of biological system activity comprehension.[1]

I. Introduction

THE first step in the network analysis is to gather technical informations about the network structure. Network properties provides insights about the organization of the biological system, the partition of the molecules, the functional structure. In the next sections we will report descriptions about a Rheumatoid Arthritis related network.

II. Calculate the main network measures for SGI, I, U and LCC-I

In Table I are reported all the network's global features. In particular for the 'seed gene interactome' (SGI), the 'intersection' (I), the 'union' (U) and the 'largest connected component of the intersection' (LCC-I) we show respectively the following global measures:

- Number of nodes, most nodes have approximately the same number of links
- Number of edges
- Number of connected components, highly connected nodes rise especially in the union network
- Number of isolated nodes, results zero for all the measurements
- Average path length, represents the easiness of the proteins to communicate their reciprocal functions
- Average degree, reflects the network structure
- Average clustering coefficient, represents how much a graph tends to be splitted into clusters (0<coeff.<1)
- Network diameter
- Network radius
- Centralization, closer to 1 means more star-topology network so same connectivity in average

The 'largest connected component of the union' (LCC-U) presents only one component, so it shows the same features reported for the union one.

In Table II are reported all the network's local features. In particular, for the largest connected component of the intersection (LCC-I) and the largest connected component of the union (LCC-U) we show the following local measures:

- betweenness, shows how much nodes tends to be intermediaries between neighbors, protein with high betweenness plays an essential role for the communication in the network
- degree, represents the numbers of links connected to the nodes
- closeness, this value shows how much a node can communicate quickly with others nodes in the network so how much the information is spread along the system

- eigenvalues, since not all connection are important in the same way, the eigenvalue rank highlights nodes that are connected to important neighbors, so proteins with high eigenvalues interact with several important proteins in the network
- ratio, an high ratio means that nodes are connected with hubs instead of nodes with small connections

Table I
GLOBAL PROPERTIES

	nodes	edges	conn comps	isolated nodes	avg path	avg degree	avg cluster coeff	diameter	radius	centralization
SGI	54	80	1	0	3.70	1.03	0.2	9	5	0.004
U	7889	15800	1	0	3.65	0.077	0.136	7	4	0.000016
I	114	104	10	0	1.73 (avg)	0.18	0.0 (avg)	2.39 (avg)	1.4	0.0012
LCC-I	54	54	1	0	4.15	0.26	0.0	9	5	0.0054

Table II
LOCAL PROPERTIES LCC-I

	betweenness	degree	closeness	eigenvalues	ratio
STAT1	1.0	0.5294117647058824	1.0	0.13877500425817477	3.7414529914529915
SQSTM	0.8480868075385495	1.0	0.9225806451612903	1.0	1.6798642533936654
BRCA2	0.6362078812107368	0.7058823529411765	0.7258883248730964	0.02595184468199791	1.7852564102564101
FANCE	0.5939463163906339	0.11764705882352941	0.8461538461538461	0.03948076877350637	10.0
KS6B1	0.5345516847515706	0.11764705882352941	0.934640522875817	0.2729349399774816	9.0
MK01	0.3500856653340948	0.23529411764705882	0.7814207650273224	0.30428403573047963	2.9471153846153846
STAT3	0.34380354083380926	0.35294117647058826	0.8265895953757226	0.11482652142559402	1.9294871794871795
EP300	0.32495716733295266	0.17647058823529413	0.8362573099415205	0.0647264592227144	3.6474358974358974
IKBA	0.17133066818960593	0.11764705882352941	0.7566137566137566	0.03556922041890927	2.8846153846153846
SKP1	0.11764705882352942	0.17647058823529413	0.6033755274261603	0.009631568888323928	1.3205128205128205
MEF2D	0.059394631639063396	0.11764705882352941	0.6470588235294118	0.016458708755414178	1.0
BCCIP	0.059394631639063396	0.11764705882352941	0.5789473684210527	0.006599061008683257	1.0
MP2K4	0.059394631639063396	0.11764705882352941	0.6137339055793991	0.07737364878523269	1.0
ITAV	0.059394631639063396	0.11764705882352941	0.6137339055793991	0.07737364878523292	1.0
PGFRA	0.0	0.058823529411764705	0.7333333333333333	0.03326078225808384	0.0
KAT2B	0.0	0.058823529411764705	0.5742971887550201	0.00621998651542248	0.0
UBC	0.0	0.058823529411764705	0.6908212560386474	0.23967415771939773	0.0
HDAC5	0.0	0.058823529411764705	0.5238095238095238	0.00394472715810276	0.0
FUS	0.0	0.058823529411764705	0.6908212560386474	0.23967415771939773	0.0
CBL	0.0	0.058823529411764705	0.6355555555555555	0.02752094980652772	0.0

III. Apply clustering methods for disease modules discovery

Making hypothesis in the context of the network lets to discover relationship within it. Starting from the hypothesis that proteins involved in the same disease have a tendency to interact with each other, we can say that components associated with a specific disease have the tendency to cluster in the same network area. Consequentially a disease module represents an area in the network in which a group of nodes that suffered a perturbation can be related to a particular disease phenotype. In order to identify disease modules is useful to adopt some clustering methods. In this section we performed three different clustering methods: Simulated annealing, Markov clustering and Louvain methods.

Simulated annealing is a stochastic optimization algorithm that enables to find the global minimum when more local minima are present. An ideal network partition consists in many within-modules links and few between-modules links, so the optimization in this case is represented by the minimization of the between-modules links. The algorithm contains two types of 'moves', local moves and global moves. The local moves are essentially where a vertex is randomly shifted in a different cluster, the globals one are safer and consist in merging and split neighborhood. The algorithm outcome strongly depend on the parameter chosen for the analysis.

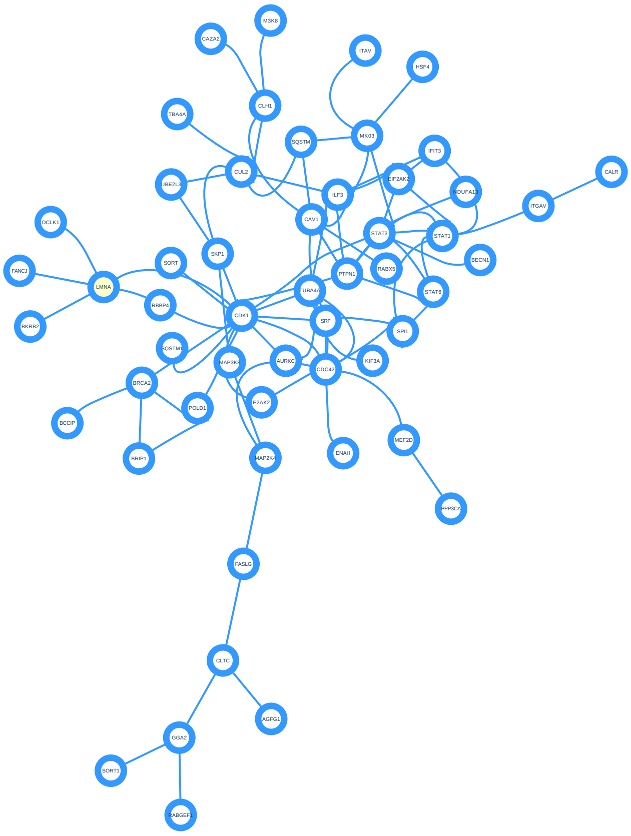

Figure 1. SGI visualization with cytoscape

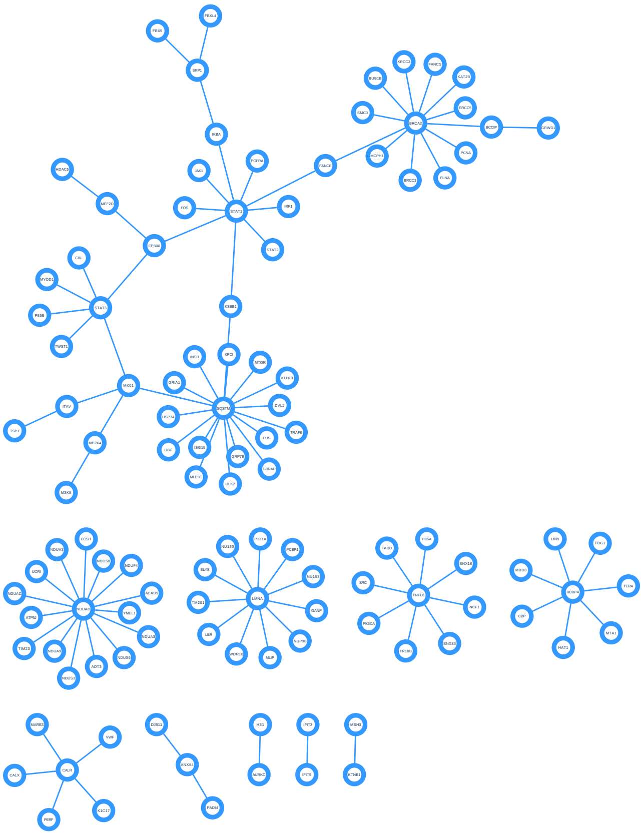

Figure 2. I visualization with cytoscape

Table III
LOCAL PROPERTIES LCC-U

	betweenness	degree	closeness	eigenvalues	ratio
CDK1	1.0	1.0	1.0	1.0	1.2467241122835793
CDC42	0.9007761900624165	0.9567839195979898	0.9147670961347869	0.7347220190108563	1.173744011493847
CUL2	0.6624471882223877	0.7427135678391958	0.8926499032882012	0.4498546761231611	1.1119884144221286
STAT3	0.5516198687215935	0.6331658291457286	0.9004878048780488	0.450207087058751	1.08615746380026
SKP1	0.480419365112232	0.6954773869346733	0.8784201760647157	0.5544181725174815	0.8612075931516204
RBBP4	0.4098609071818159	0.4974874371859296	0.8546296296296296	0.27916646864342165	1.0271284005811354
BDKRB2	0.4064779428428935	0.33065326633165826	0.7641361039821177	0.03458075982463057	1.5326201312808316
TUBA4A	0.37972078743591997	0.514572864321608	0.9127768987341773	0.45475102418549607	0.9200000514131068
ILF3	0.3284974832254715	0.36482412060301506	0.8619723571161749	0.16165961737340054	1.1225840344238502
PPP3CA	0.3144136781743872	0.42010050251256276	0.8094716071037052	0.23815604915514718	0.933079373786409
LMNA	0.31318202922849836	0.34271356783919593	0.856413825098585	0.19261792352947338	1.1392942212205404
CAV1	0.30042725663887665	0.33366834170854265	0.8716592690527907	0.19213594704423004	1.1225215521527114
STAT1	0.2986478728434332	0.46231155778894467	0.8517903285345146	0.3255602153492194	0.8053692318159305
MK03	0.28256052481796956	0.2834170854271357	0.8348030570252792	0.10525311336369177	1.2429561857188478
SQSTM1	0.2754375384770579	0.3366834170854271	0.8694423511680482	0.21815259718869603	1.019933276251186
CLTC	0.2732645358251815	0.378894472361809	0.815083009537266	0.1829905924744135	0.8991566536233625
RAP2A	0.23518892582387332	0.2844221105527638	0.7849306913853219	0.127709533663455	1.0309174072185994
CALR	0.2216192679516999	0.2723618090452261	0.8000693451220041	0.10713608752076802	1.0144523788801094
CAZA2	0.21824446568689881	0.22412060301507536	0.8267646005016124	0.07827404892752703	1.2140367020429665
POLD1	0.21436651300283863	0.4040201005025125	0.834463430069614	0.3239385875696569	0.6614915948349661

MCL Markov Clustering simulates the flow diffusion in a graph. Practically is based on the idea of preserving flows where the current is strong and debasing flows where the current is weak. In this way if natural clusters are present in the network by penalizing current across different groups borders, the cluster structure of the graph will automatically appear. Thanks to the transfer matrix can be obtained the probability description for a random walker to reach all elements in i'th rows and j'th columns in one step. This technique allows to identify vertexes which are likely to be in the same community.[3]
Louvain modularity idea is characterized by the comparison between the density of the connected the nodes within a community and the suitable density connection in a random graph. Louvain algorithm consists in two main steps: the first one is a local optimization modularity looking for small communities and the second one is made up of a nodes aggregation belonging to the same community.[2] In the Tables IV and V we show the performances of these three algorithms on the LCC-I and LCC-U. For each clustered partition we report information about modules which result having a p-value<0.5 in the hyper-geometric test (in other words the putative disease modules). We present respectively for each algorithm used: the module reference index, the number of seed in the module, the number of genes in the module, the ratio between the number of seed genes on the total genes in the module and the related p-value.

Table IV
PUTATIVE DISEASE MODULES LCC-I

algo	index	n seed	n genes	ratio	p-value
louvain	1	2	8	0.25	0.0358201916
markov	6	1	3	0.3333333333	0.0287783495
markov	8	1	2	0.5	0.0101010101
annealing	2	2	8	0.25	0.0358201916

IV. FIND ROLES OF I-LCC NODES (ACCORDING TO GUIMERA E AMARAL METHOD)

Using the Netcarto command line tool we clustered LCC-I and LCC-U in modules by maximizing modularity according to Guimera&Amaral method. This technique works associating to each node a z-score of within-module degree, so how much the node is well connected to the other nodes in the module.

Table V
PUTATIVE DISEASE MODULES LCC-U

algo	index	n seed	n genes	ratio	p-value
louvain	13	8	668	0.0119760479	0.0359347397
markov	0	1	36	0.0277777778	0.0249732551
markov	9	1	39	0.0256410256	0.0289939845
markov	13	1	27	0.037037037	0.0144685843
markov	37	1	42	0.0238095238	0.0332549471
markov	39	1	16	0.0625	0.005189896
markov	40	1	34	0.0294117647	0.0224318631
markov	47	1	11	0.0909090909	0.0024314038
markov	51	1	32	0.03125	0.0200057904
markov	54	1	34	0.0294117647	0.0224318631
markov	56	1	36	0.0277777778	0.0249732551
markov	57	1	30	0.0333333333	0.0176985836
markov	59	1	42	0.0238095238	0.0332549471
markov	60	1	49	0.0204081633	0.0440697897
markov	62	1	44	0.0227272727	0.036223563
markov	63	1	15	0.0666666667	0.0045610867
markov	64	1	46	0.0217391304	0.039290814
markov	65	1	27	0.037037037	0.0144685843
markov	66	1	30	0.0333333333	0.0176985836
annealing	0	8	534	0.0149812734	0.009591349
annealing	7	11	854	0.0128805621	0.011253429
annealing	16	6	421	0.0142517815	0.0237318864

Nodes are classified in 'non-hub nodes' and in 'hub nodes' depending on a z-score which is for the two classes respectively <2.5 and >2.5. Inside the two classes we can find another kind of classification which measures the participation (P) of the node in the module, e.g. non-hub nodes can be ultra-peripherals (R1), peripherals (R2), connectors (R3) or kinlesses (R4) and hub-nodes can be provincials (R5), connectors (R6) or kinlesses (R7) based on the participation coefficient score. In the Table VI we show the Netcarto modules classification and the related modules information for the LCC-I interactome, among which the participation roles so thanks to the labels in the table it comes easy to read the figure 3 nodes configuration.

V. Find putative disease proteins using the approach from Ghiassian et al.

Using Diamonds software, given the human interactome and given the original seed genes list we retrieved the first 40 relevant proteins related to the disease for both Apid and Biogrid databases - reported in the Table VII. Then we performed the intersection between them (Table VIII) to find the most important proteins and finally we carry out the enrichment analysis (Table IX and Table X). Diamond tool represents a different bioinformatic approach based on the idea that proteins associated to the disease don't lie within locally dense communities. Since the disease-associated proteins present distinct connectivity patterns in order to highlight them is useful evaluate the significance of the connections instead of the density of the modules, hence when the connections to the seed genes are more than the expected. These specific disease-modules within the interactome are far from the topological densely interconnected communities investigated before.

VI. Conclusion

We can clearly observe the presence of the RNA polymerase II from the Table V. RNA polymerase II is an enzyme to which the transcription of proteins is imputed. Given its consistent presence in the results it is linkable for the activation of cells responsible in the production of antibodies and therefore determinant in the disease through the same RNA polymerase II promoter.

Table VI
NETCARTO MODULARITY LCC-I

	module	connectivity	participation	role
DVL2	0	-0.2581988897	0	Ultra peripheral
FUS	0	-0.2581988897	0	Ultra peripheral
GBRAP	0	-0.2581988897	0	Ultra peripheral
GRIA1	0	-0.2581988897	0	Ultra peripheral
GRP78	0	-0.2581988897	0	Ultra peripheral
HSP74	0	-0.2581988897	0	Ultra peripheral
INSR	0	-0.2581988897	0	Ultra peripheral
ISG15	0	-0.2581988897	0	Ultra peripheral
KLHL3	0	-0.2581988897	0	Ultra peripheral
KPCI	0	-0.2581988897	0	Ultra peripheral
MLP3C	0	-0.2581988897	0	Ultra peripheral
MTOR	0	-0.2581988897	0	Ultra peripheral
TRAF6	0	-0.2581988897	0	Ultra peripheral
UBC	0	-0.2581988897	0	Ultra peripheral
ULK2	0	-0.2581988897	0	Ultra peripheral
SQSTM	0	3.8729833462	0.214532872	Connector Hub
M3K8	1	-1.2247448714	0	Ultra peripheral
TSP1	1	-1.2247448714	0	Ultra peripheral
ITAV	1	0.8164965809	0	Ultra peripheral
MP2K4	1	0.8164965809	0	Ultra peripheral
MK01	1	0.8164965809	0.625	Connector
FOS	2	-0.377964473	0	Ultra peripheral
IRF1	2	-0.377964473	0	Ultra peripheral
JAK1	2	-0.377964473	0	Ultra peripheral
PGFRA	2	-0.377964473	0	Ultra peripheral
STAT2	2	-0.377964473	0	Ultra peripheral
FANCE	2	-0.377964473	0.5	Peripheral
KS6B1	2	-0.377964473	0.5	Peripheral
STAT1	2	2.6457513111	0.3703703704	Connector Hub
FBX5	3	-0.5773502692	0	Ultra peripheral
FBXL4	3	-0.5773502692	0	Ultra peripheral
IKBA	3	-0.5773502692	0.5	Peripheral
SKP1	3	1.7320508076	0	Ultra peripheral
BRCC3	4	-0.3186064455	0	Ultra peripheral
BUB1B	4	-0.3186064455	0	Ultra peripheral
ERCC5	4	-0.3186064455	0	Ultra peripheral
FANCG	4	-0.3186064455	0	Ultra peripheral
FLNA	4	-0.3186064455	0	Ultra peripheral
GRWD1	4	-0.3186064455	0	Ultra peripheral
KAT2B	4	-0.3186064455	0	Ultra peripheral
MCPH1	4	-0.3186064455	0	Ultra peripheral
PCNA	4	-0.3186064455	0	Ultra peripheral
SMC3	4	-0.3186064455	0	Ultra peripheral
XRCC3	4	-0.3186064455	0	Ultra peripheral
BCCIP	4	0.0579284446	0	Ultra peripheral
BRCA2	4	3.4467424559	0.1527777778	Connector Hub
CBL	5	-0.5773502692	0	Ultra peripheral
HDAC5	5	-0.5773502692	0	Ultra peripheral
MYOD1	5	-0.5773502692	0	Ultra peripheral
P85B	5	-0.5773502692	0	Ultra peripheral
TWST1	5	-0.5773502692	0	Ultra peripheral
MEF2D	5	0.1924500897	0	Ultra peripheral
EP300	5	0.1924500897	0.4444444444	Peripheral
STAT3	5	2.5018511665	0.2777777778	Connector Hub

REFERENCES

[1] Tuba Sevimoglu, Kazim YalcinArga. "The role of protein interaction networks in systems biomedicine." Vol. 11, pp. 22-27, 2014, Istanbul.

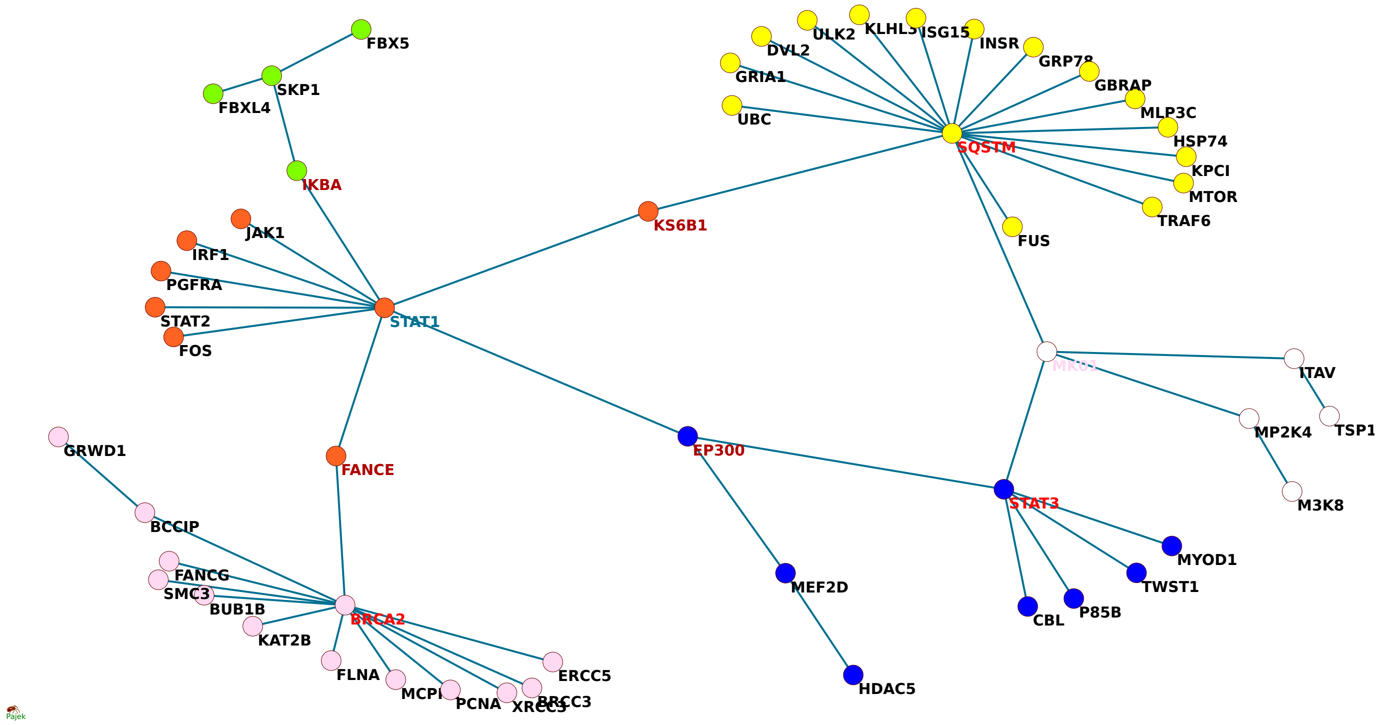

Figure 3. netcarto modularity LCC-I

[2] https://perso.uclouvain.be/vincent.blondel/research/louvain.html

[3] Van Dongen, Stijn Marinus. Graph clustering by flow simulation. Diss. 2001.

Table VII
FIRST 40 DIAMOND MODULARITY

	apid	biogrid
0	PTPN2	B5B2P4
1	PML	RFA1
2	MAPK1	TAT
3	BRCA1	A0A0U3FYV6
4	STAT5A	A8K503
5	BCL2	ATF3
6	CEBPB	A0A0S2Z4Z9
7	TP53	SIN3A
8	NR3C1	AKT1
9	JUN	F1D8Q5
10	CREBBP	MYOD1
11	RELA	UHRF2
12	SP1	B3KRS7
13	EP300	TOP2A
14	NCOA1	DNM3L
15	PPARG	B3KU66
16	NFKB1	AP2A
17	NCOR2	B4DPW8
18	RXRA	A0A024R1S7
19	NCOR1	E9PJK2
20	NCOA3	OBSL1
21	VDR	CTBP1
22	ETS1	B3KNL2
23	AR	CEBPA
24	HDAC3	HMGA1
25	RARA	E2F1
26	NCOA2	PIAS1
27	NCOA6	A0A087WVR4
28	ESR1	DNMT1
29	FOS	HDAC2
30	BCL3	A0A024R2F2
31	SRC	TYY1
32	PPARD	CDC5L
33	HIF1A	Q9BVT2
34	SMARCA4	SKI
35	SMAD3	B3KYA8
36	RUNX2	B4DT73
37	RUNX1	B7Z855
38	HDAC1	B7Z3N9
39	RB1	A0A024R4A0

Table VIII
INTERSECTION LIST APID BIOGRID

	Symbol	Name
0	PML	promyelocytic leukemia
1	ATF3	activating transcription factor 3
2	CEBPB	CCAAT/enhancer binding protein beta
3	SIN3A	SIN3 transcription regulator family member A
4	AKT1	AKT serine/threonine kinase 1
6	MYOD1	myogenic differentiation 1
7	UHRF2	ubiquitin like with PHD and ring finger domains 2
8	SP1	Sp1 transcription factor
9	TOP2A	DNA topoisomerase II alpha
10	EP300	E1A binding protein p300
11	NCOA1	nuclear receptor coactivator 1
12	NCOA3	nuclear receptor coactivator 3
13	CTBP1	C-terminal binding protein 1
14	CEBPA	CCAAT/enhancer binding protein alpha
15	HMGA1	high mobility group AT-hook 1
16	HDAC3	histone deacetylase 3
17	E2F1	E2F transcription factor 1
18	PIAS1	protein inhibitor of activated STAT 1
19	DNMT1	DNA methyltransferase 1
20	HDAC2	histone deacetylase 2
22	SKI	SKI proto-oncogene
23	RUNX1	runt related transcription factor 1
24	HDAC1	histone deacetylase 1
25	CCND1	cyclin D1
26	FHL2	four and a half LIM domains 2
27	KLF5	Kruppel like factor 5
28	MBD3	methyl-CpG binding domain protein 3
29	FOXO1	forkhead box O1
30	HDAC4	histone deacetylase 4
31	DDX5	DEAD-box helicase 5
32	CHD4	chromodomain helicase DNA binding protein 4
33	KAT2B	lysine acetyltransferase 2B
34	CDK8	cyclin dependent kinase 8
35	CARM1	coactivator associated arginine methyltransferase 1
37	KAT2A	lysine acetyltransferase 2A
38	NR0B2	nuclear receptor subfamily 0 group B member 2
39	XRCC5	X-ray repair cross complementing 5
41	GATA1	GATA binding protein 1
42	MED1	mediator complex subunit 1
43	GSK3B	glycogen synthase kinase 3 beta
44	HDAC5	histone deacetylase 5
45	KMT2A	lysine methyltransferase 2A
46	SNAI1	snail family transcriptional repressor 1
47	KDM1A	lysine demethylase 1A
48	MTA1	metastasis associated 1

Table IX
TABLE GO INTERSECTION AND SEED GENES

	name	p-value	Z-score	Combined scores
0	positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	6.015181379119896e-33	-7.770452474596514	576.4978071154287
1	positive regulation of transcription from RNA polymerase II promoter in response to nitrogen starvation (GO:0036278)	5.856161765384095e-31	-6.608655127804344	460.04595331174755
2	positive regulation of transcription from RNA polymerase II promoter involved in meiotic cell cycle (GO:0010673)	7.914569613030066e-30	-6.494855311741902	435.2127689507595
3	positive regulation of sulfate assimilation by positive regulation of transcription from RNA polymerase II promoter (GO:1900478)	7.914569613030066e-30	-6.494046125695954	435.1585463264284
4	positive regulation of ethanol catabolic process by positive regulation of transcription from RNA polymerase II promoter (GO:0061425)	7.914569613030066e-30	-6.490769881465212	434.93900897648274
5	positive regulation of filamentous growth of a population of unicellular organisms in response to starvation by positive regulation of transcription from RNA polymerase II promoter (GO:1904741)	7.914569613030066e-30	-6.479550972234497	434.1872427990162
6	regulation of glycolytic process by positive regulation of transcription from RNA polymerase II promoter (GO:0072363)	7.914569613030066e-30	-6.476627945634206	433.99137415539457
7	positive regulation of transcription from RNA polymerase II promoter involved in heart development (GO:1901228)	7.914569613030066e-30	-6.474686292504587	433.86126621698435
8	positive regulation of SREBP signaling pathway (GO:2000640)	1.0275793243930177e-29	-6.496760578781519	433.6442277859492
9	positive regulation of oligopeptide transport by positive regulation of transcription from RNA polymerase II promoter (GO:0035951)	7.914569613030066e-30	-6.470116999270874	433.55508314362214

Table X
TABLE PATHWAY INTERSECTION INTERACTOME GENES

	name	p-value	Z-score	Combined scores
	name	P-value	Z-score	Combined score
0	Pathways in cancer_Homo sapiens_hsa05200	2.9824677694473457e-19	-2.0830857455542535	88.85686734371791
1	Thyroid hormone signaling pathway_Homo sapiens_hsa04919	2.9054730525778293e-20	-1.8616235246819735	83.74533092348008
2	Viral carcinogenesis_Homo sapiens_hsa05203	4.1291973396933465e-16	-1.9314667609813165	68.41888488838426
3	HTLV-I infection_Homo sapiens_hsa05166	3.047776881352479e-13	-1.8680592277152335	53.83596086751211
4	Hepatitis B_Homo sapiens_hsa05161	1.2550002527982697e-11	-1.8344093172621592	46.04605905200994
5	Epstein-Barr virus infection_Homo sapiens_hsa05169	3.7276374733322716e-11	-1.8245370579339306	43.81199053852064
6	Acute myeloid leukemia_Homo sapiens_hsa05221	5.036063846326759e-10	-1.7530670629208276	37.53180918139089
7	Transcriptional misregulation in cancer_Homo sapiens_hsa05202	1.4664015127643367e-10	-1.6084084039944122	36.41925499286356
8	Influenza A_Homo sapiens_hsa05164	1.6684304708719847e-09	-1.7621834139591441	35.6161629974912
9	Pancreatic cancer_Homo sapiens_hsa05212	1.6850719086926446e-09	-1.6782809246982724	33.90372093827867