Graph Theoretic and Pearson Correlation Based Discovery of Network Biomarkers for Cancer

Raihanul Bari Tanvir
Tasmia Aqila
Mona Maharjan
Abdullah Al Mamun

Background

- Existence of pattern in disease progression (human, mouse)

Article | OPEN | Published: 10 December 2012

Identifying critical transitions and their leading biomolecular networks in complex diseases

Rui Liu, Meiyi Li, Zhi-Ping Liu, Jiarui Wu, Luonan Chen & Kazuyuki Aihara

Background (Cont'd)

Genome Res. 2008 Apr; 18(4): 644–652.
doi: 10.1101/gr.071852.107
PMCID: PMC3863981
PMID: 18381899

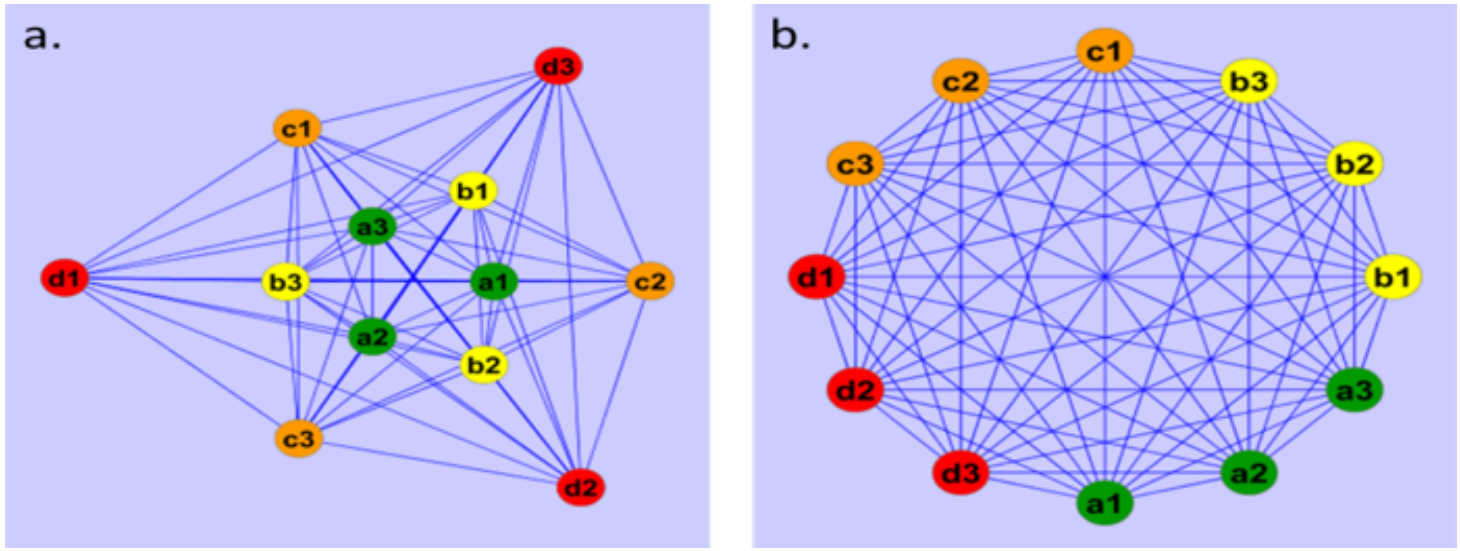
Protein networks in disease

Trey Ideker[1] and Roded Sharan[2,3]

- Applications of gene interaction networks to diseases:
  - ✓ Identifying new disease genes
  - ✓ Studying the network properties of disease genes
  - ✓ Classifying diseases based on gene network
  - ✓ Identifying disease-related subnetworks

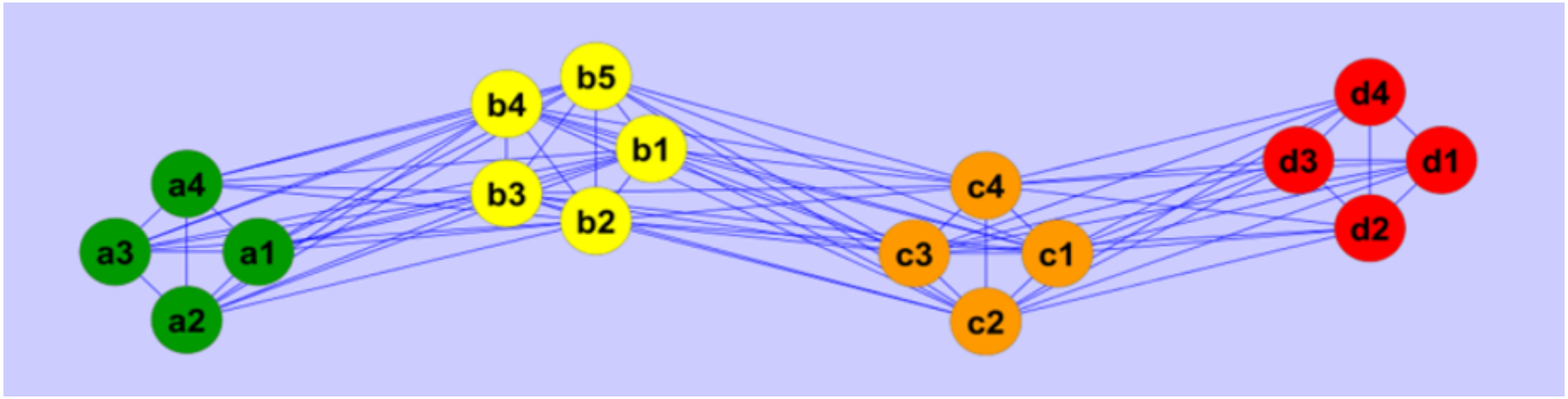
Hypothesis Development

- ***Hypothesis-1: Disease Progression as Cliques***



Hypothesis Development (Cont'd)

- ***Hypothesis-2: Disease Progression as Bipartite Graphs***



Data Collection

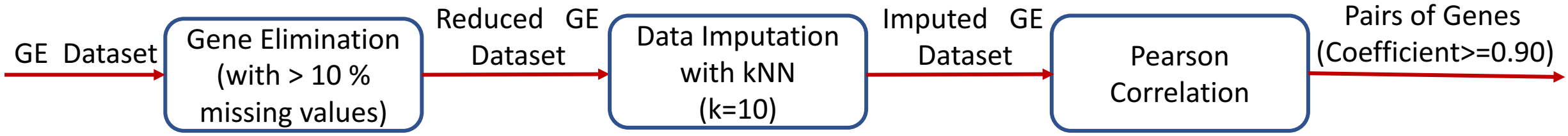
Gene expression data obtained from ***LinkedOmics.***

- We took three cancers :
  - Breast invasive carcinoma (BRCA)
  - Glioblastoma multiforme (GBM)
  - Colorectal adenocarcinoma (COAD)

- All samples are cancer patients.

Data Cleaning

attrib_name	TCGA.02.0047	TCGA.02.0055	TCGA.02.2483	TCGA.02.2485	TCGA.02.2486	TCGA.06.0125	TCGA.06.0129	TCGA.06.0130	TCGA.06.0132	TCGA.06.0138
A1BG	6.9774	8.6177	8.092	6.4084	6.7716	7.663	7.0741	7.2658	7.0304	7.9765
A1CF	0	0	0	0	0	0	0.4977	0	0	0
A2BP1	7.9403	7.1122	6.8077	8.012	2.3973	5.743	4.2578	6.5118	9.9285	6.9862
A2M	15.0538	15.3879	14.3622	12.9292	15.3224	13.4481	14.4261	14.4349	14.9153	13.8007
AACS	8.9155	8.3143	8.9486	8.254	8.6187	8.8189	8.0617	8.4043	9.7849	8.2917
AADACL2	0	0	0	0	0	0	0	0	0	0
AADACL3	0.6912	0	0	0	0	0	0	0	0	0
AADACL4	0	0	0	0	0	0	0	0	0	0



Cancer	No of Genes	No of Samples	Reduced no of genes
Breast Invasive Carcinoma (BRCA)	20155	1093	16011
Colorectal Adenocarcinoma (COAD)	19828	379	15769
Glioblastoma Multiforme (GBM)	19660	153	16186

Clique Identification

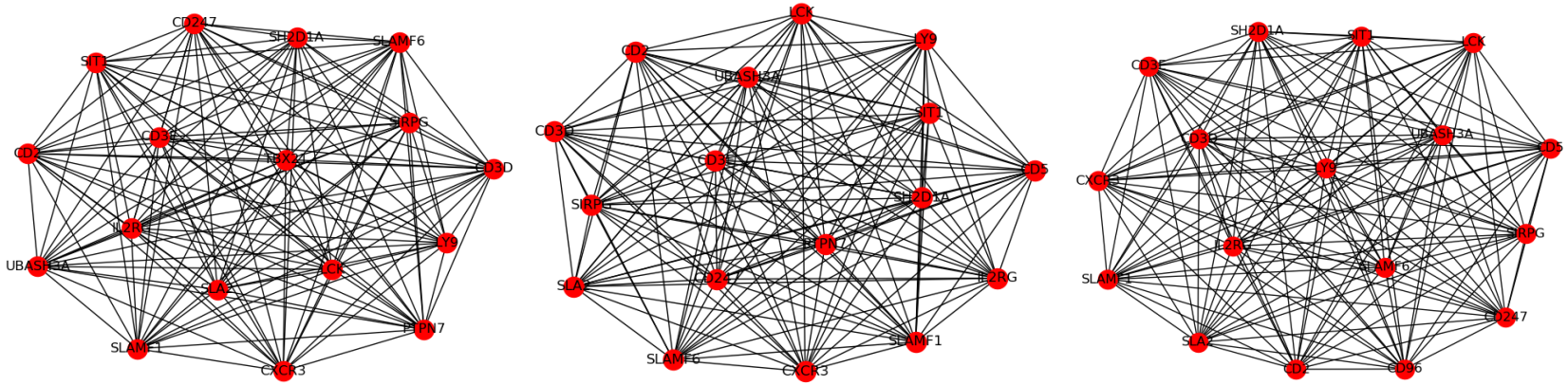
- We have the network for three cancers generated by Pearson's correlation network with PCC=>0.9
- Now we construct the network in Pythons NetworkX package.
- we get all the cliques from the networks and sort them based on their number of nodes, with NetworkX.

Size and Frequency of cliques

- We got 209, 1535 and 322 cliques total in Networks of BRCA, COAD and GBM.

BRCA		COAD		GBM	
Size	Frequency	Size	Frequency	Size	Frequency
3	48	3	81	3	68
4	43	4	95	4	76
5	35	5	82	5	71
6	14	6	129	6	18
7	13	7	115	7	29
8	4	8	113	8	20
9	2	9	139	9	20
10	2	10	133	10	24
11	6	11	101	11	6
12	13	12	119		
13	15	13	103		
14	7	14	97		
15	3	15	70		
16	1	16	80		
17	3	17	27		
		18	41		
		19	10		

Largest Cliques from BRCA



clique_1	CD3E	CD2	SH2D1A	SLAMF6	CD247	CD3D	IL2RG	SIRPG	SLAMF1	SLA2	LCK	SIT1	LY9	UBASH3A	CXCR3	TBX21	PTPN7
clique_2	CD3E	CD2	SH2D1A	SLAMF6	CD247	CD3D	IL2RG	SIRPG	SLAMF1	SLA2	LCK	SIT1	LY9	UBASH3A	CXCR3	CD5	PTPN7
clique_3	CD3E	CD2	SH2D1A	SLAMF6	CD247	CD3D	IL2RG	SIRPG	SLAMF1	SLA2	LCK	SIT1	LY9	UBASH3A	CXCR3	CD5	CD96

- From here we can see that 15 genes are common among all three. 16 out of 17 are common between any two.

Largest Clique in COAD

- These are the largest cliques in COAD

clique_1	CD53	CD4	LILRB1	CYBB	LAPTM5	LAIR1	ITGB2	SPI1	HCK	CD86	C3AR1	LRRC25	SLAMF8	CLEC7A	C1QC	SIGLEC7	CSF1R	DOK2	CYTH4
clique_2	CD53	CD4	LILRB1	C1QB	C1QC	LAIR1	C3AR1	LRRC25	SLAMF8	ITGB2	HCK	CLEC7A	SIGLEC7	SPI1	CD86	LAPTM5	CSF1R	DOK2	CYTH4
clique_3	CD53	HAVCR2	CD86	LAPTM5	LAIR1	HCK	C3AR1	LRRC25	ITGB2	SPI1	LILRB4	CLEC7A	SIGLEC7	C1QC	CYBB	SLAMF8	TFEC	FPR3	MS4A4A
clique_4	CD53	HAVCR2	CD86	LAPTM5	LAIR1	HCK	C3AR1	LRRC25	ITGB2	SPI1	LILRB4	CLEC7A	SIGLEC7	C1QC	C1QB	SLAMF8	TFEC	FPR3	MS4A4A
clique_5	CD53	HAVCR2	CD86	LAPTM5	LAIR1	HCK	C3AR1	PDCD1LG2	LILRB4	TFEC	CYBB	FPR3	SLAMF8	ITGB2	CLEC7A	SIGLEC7	SPI1	MS4A4A	C1QC
clique_6	FCER1G	LAIR1	ITGB2	C3AR1	LAPTM5	SPI1	HAVCR2	CD86	LRRC25	C1QC	MS4A4A	TYROBP	CD300A	SIGLEC9	LILRB4	FPR3	HCK	CLEC7A	SIGLEC7
clique_7	FCER1G	LAIR1	ITGB2	C3AR1	LAPTM5	SPI1	HAVCR2	CD86	LRRC25	C1QC	MS4A4A	TYROBP	CD300A	C1QB	HCK	LILRB4	FPR3	CLEC7A	SIGLEC7
clique_8	FCER1G	LAIR1	ITGB2	C3AR1	LAPTM5	SPI1	HAVCR2	CD86	LRRC25	C1QC	MS4A4A	TYROBP	SLAMF8	SIGLEC9	LILRB4	FPR3	HCK	CLEC7A	SIGLEC7
clique_9	FCER1G	LAIR1	ITGB2	C3AR1	LAPTM5	SPI1	HAVCR2	CD86	LRRC25	C1QC	MS4A4A	TYROBP	SLAMF8	C1QB	HCK	LILRB4	FPR3	CLEC7A	SIGLEC7
clique_10	FCER1G	LAIR1	ITGB2	C3AR1	LAPTM5	SPI1	HAVCR2	CD86	LRRC25	C1QC	MS4A4A	TFEC	LILRB4	C1QB	FPR3	SLAMF8	HCK	CLEC7A	SIGLEC7

- 11 genes are common in all of them.

Largest Clique in GBM

- These are the largest cliques in GBM

clique_1	ITGB2	NCKAP1L	VAV1	FERMT3	HCK	LAIR1	SPI1	PTPN6	SASH3	WAS	STXBP2
clique_2	ITGB2	NCKAP1L	VAV1	FERMT3	HCK	LAIR1	SPI1	PTPN6	SASH3	WAS	LAPTM5
clique_3	ITGB2	NCKAP1L	VAV1	CD4	LAIR1	HCK	STXBP2	PTPN6	SPI1	SASH3	WAS
clique_4	ITGB2	NCKAP1L	VAV1	CD4	LAIR1	HCK	LAPTM5	SASH3	PTPN6	SPI1	WAS
clique_5	ALOX5	ITGB2	NCKAP1L	VAV1	HCK	SASH3	LAIR1	STXBP2	PTPN6	SPI1	WAS
clique_6	ALOX5	ITGB2	NCKAP1L	VAV1	HCK	SASH3	LAIR1	LAPTM5	SPI1	PTPN6	WAS

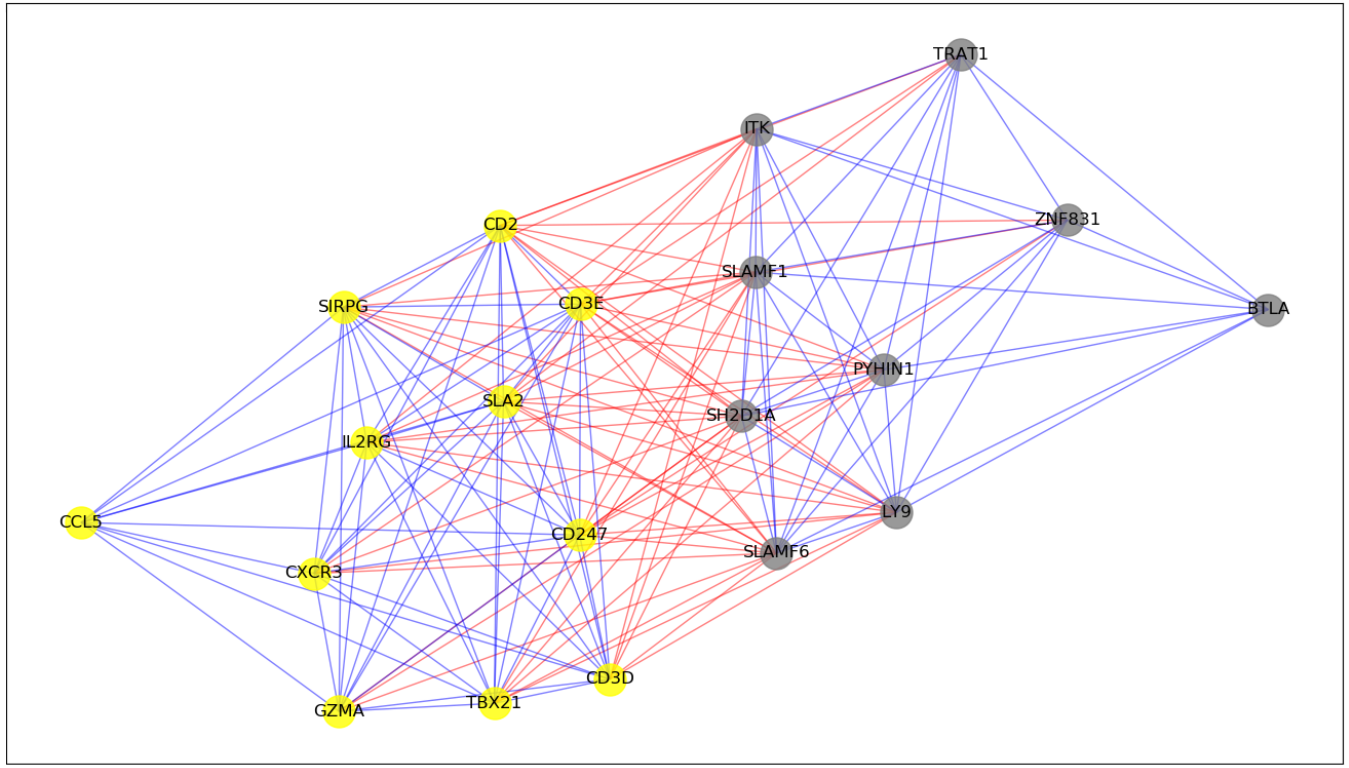
- 9 genes are common in all of them.

2-Clique-1-Bipartite Module

- We discovered the connection between cliques.
- Condition: Clique 1 and Clique 2 must not have any common genes.
- We found 4273, 172595, 11256 such modules for in network of BRCA, COAD and GBM respectively.

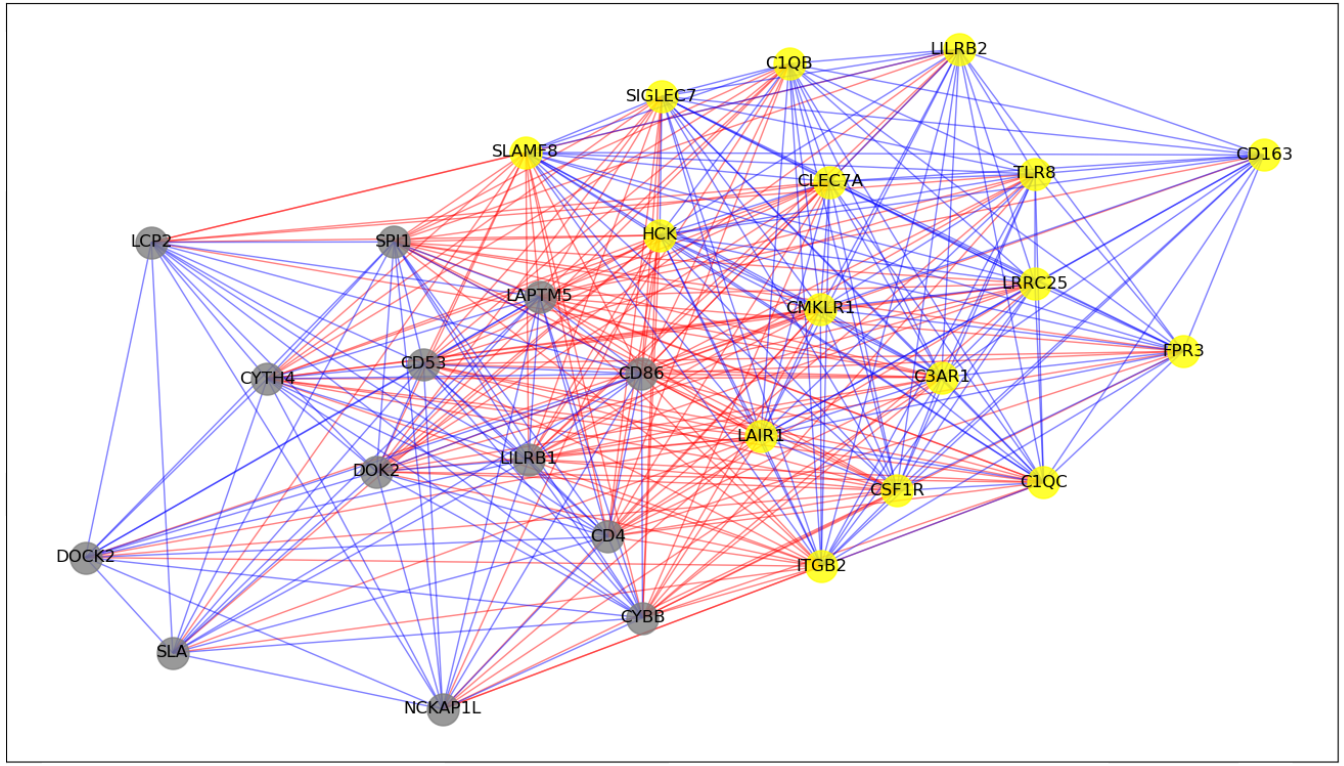
2-Clique-1-Bipartite: BRCA

- 11 genes in clique #1
- 9 genes in clique #2
- 59 edges are connecting these two cliques



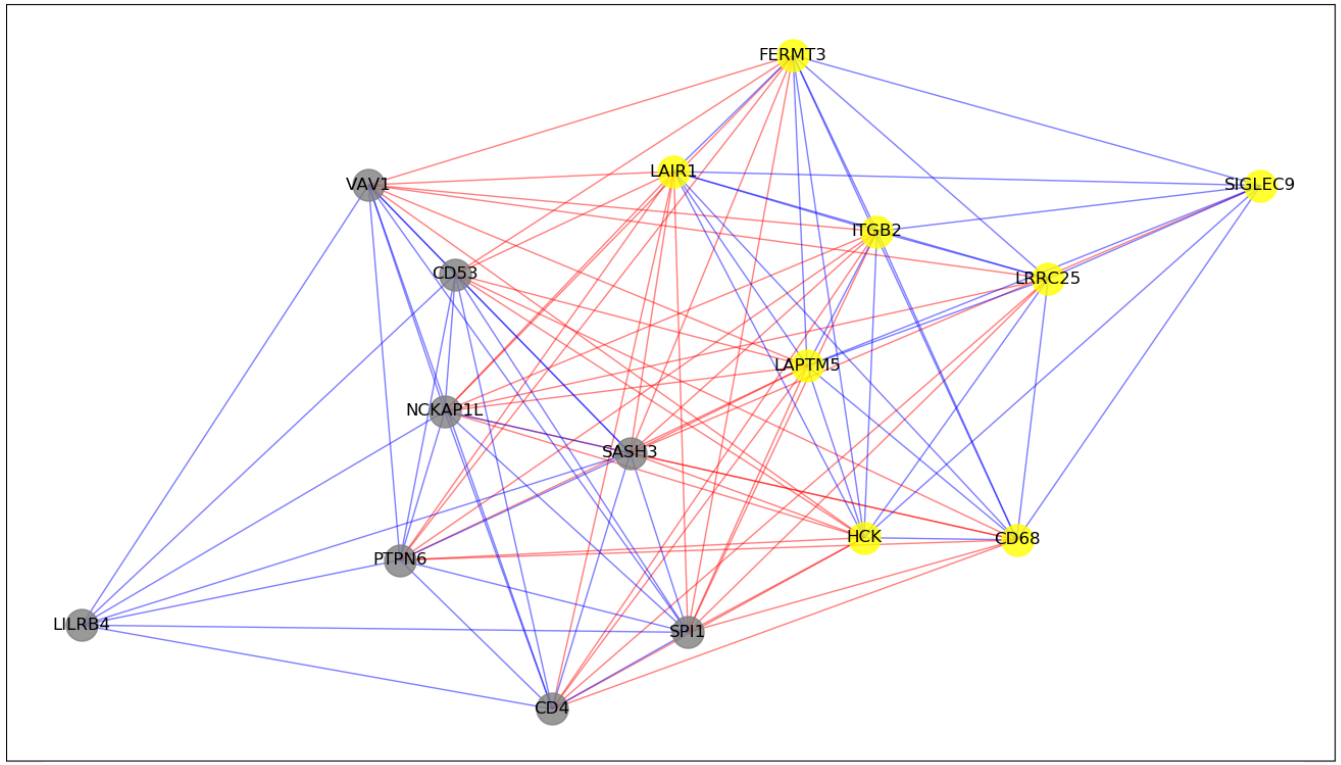
2-Clique-1-Bipartite: COAD

- 16 genes in clique #1
- 13 genes in clique #2
- 145 edges are connecting these two cliques



2-Clique-1-Bipartite: GBM

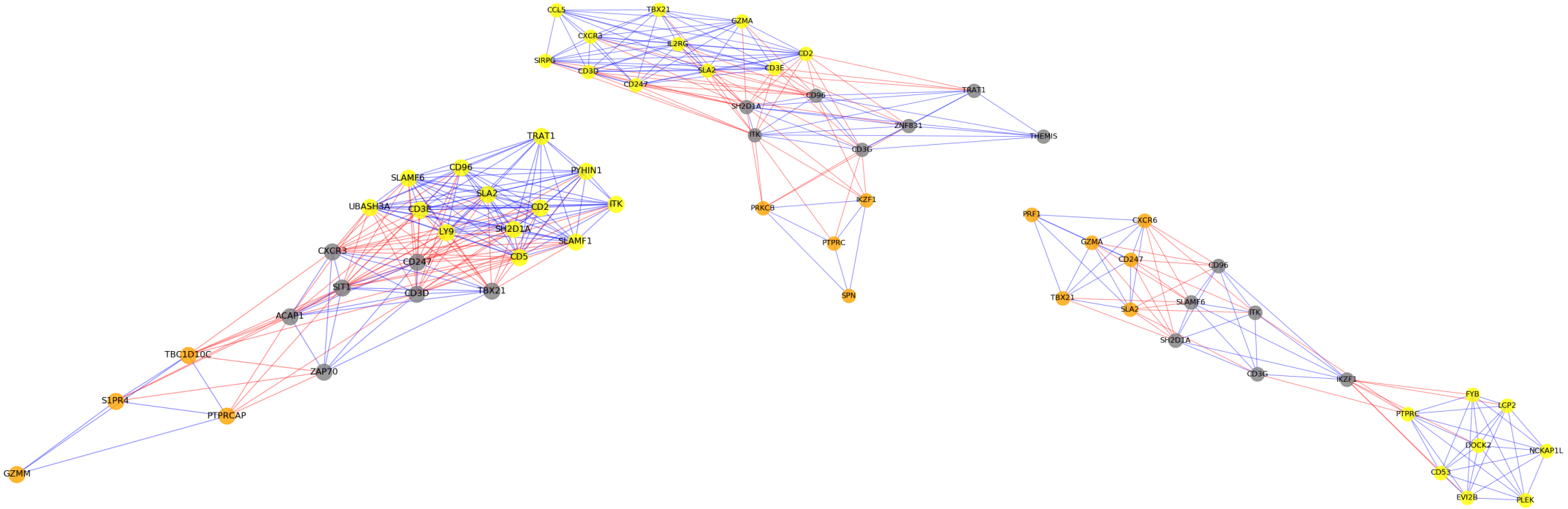
- 8 genes in clique #1
- 8 genes in clique #2
- 44 edges are connecting these two cliques



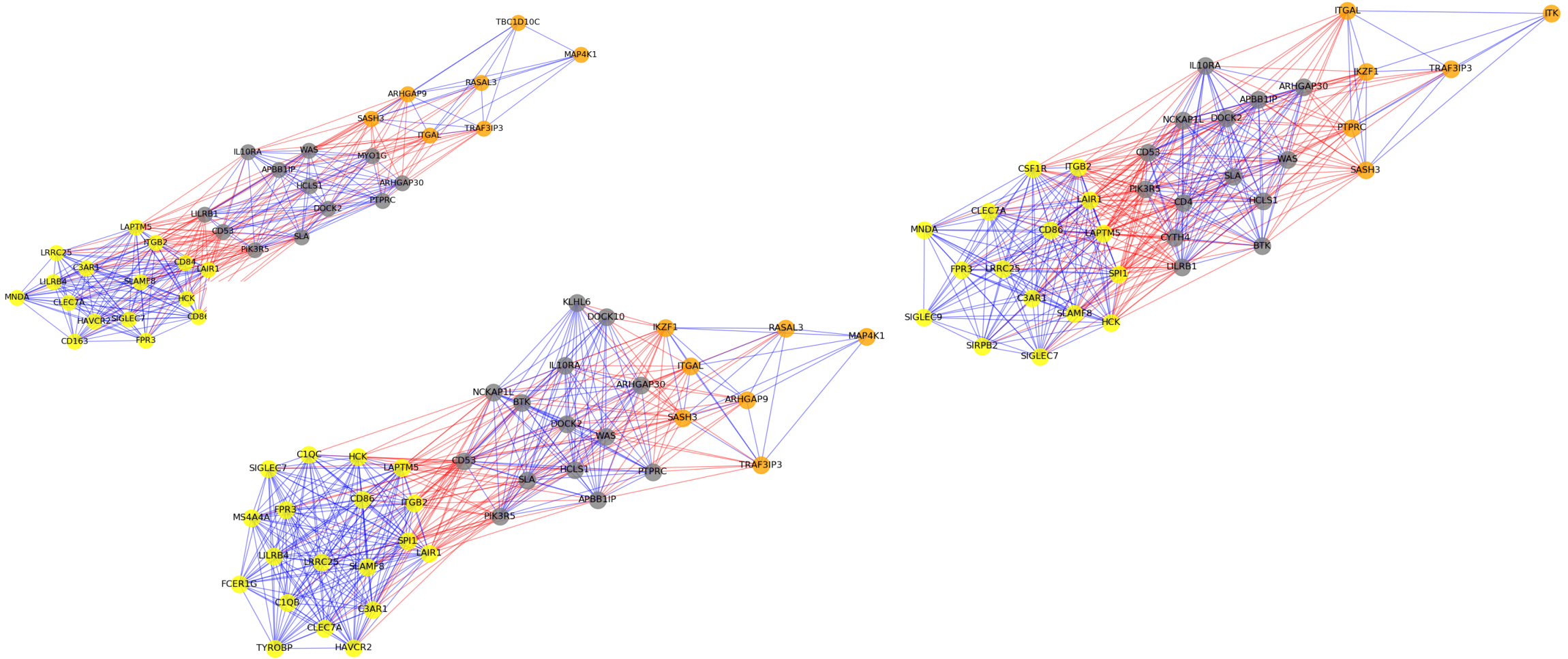
3-Clique-2-Bipartite Module

- We discovered connections between three cliques such that:
  - If there are 3 cliques A,B and C
  - A is connected to B and B is connected to C
  - None of them have any common genes
  - A is not connected to C
- We found that BRCA, COAD and GBM has 32451, 1995768 and 42338 such modules in their base network.

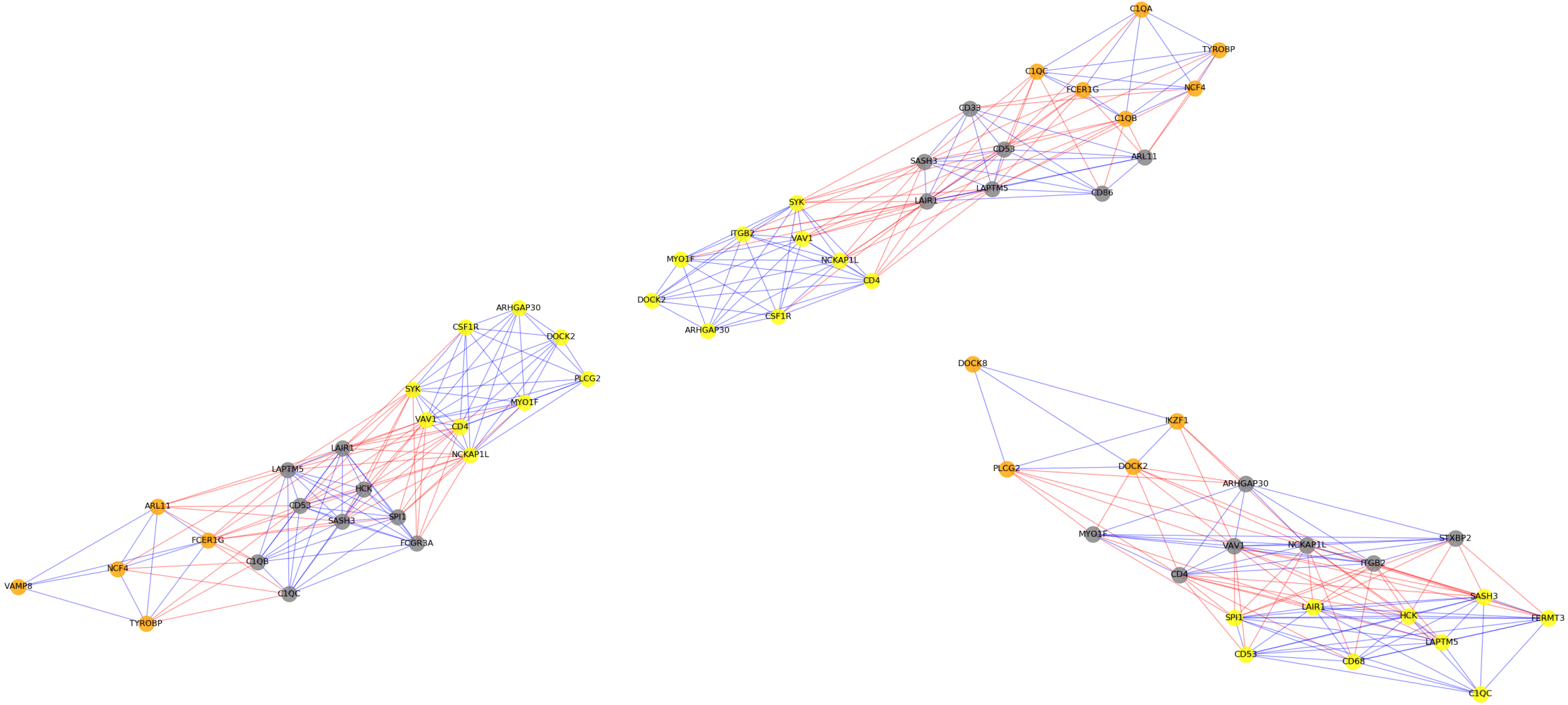
3-Clique-2-Bipartite Module: BRCA



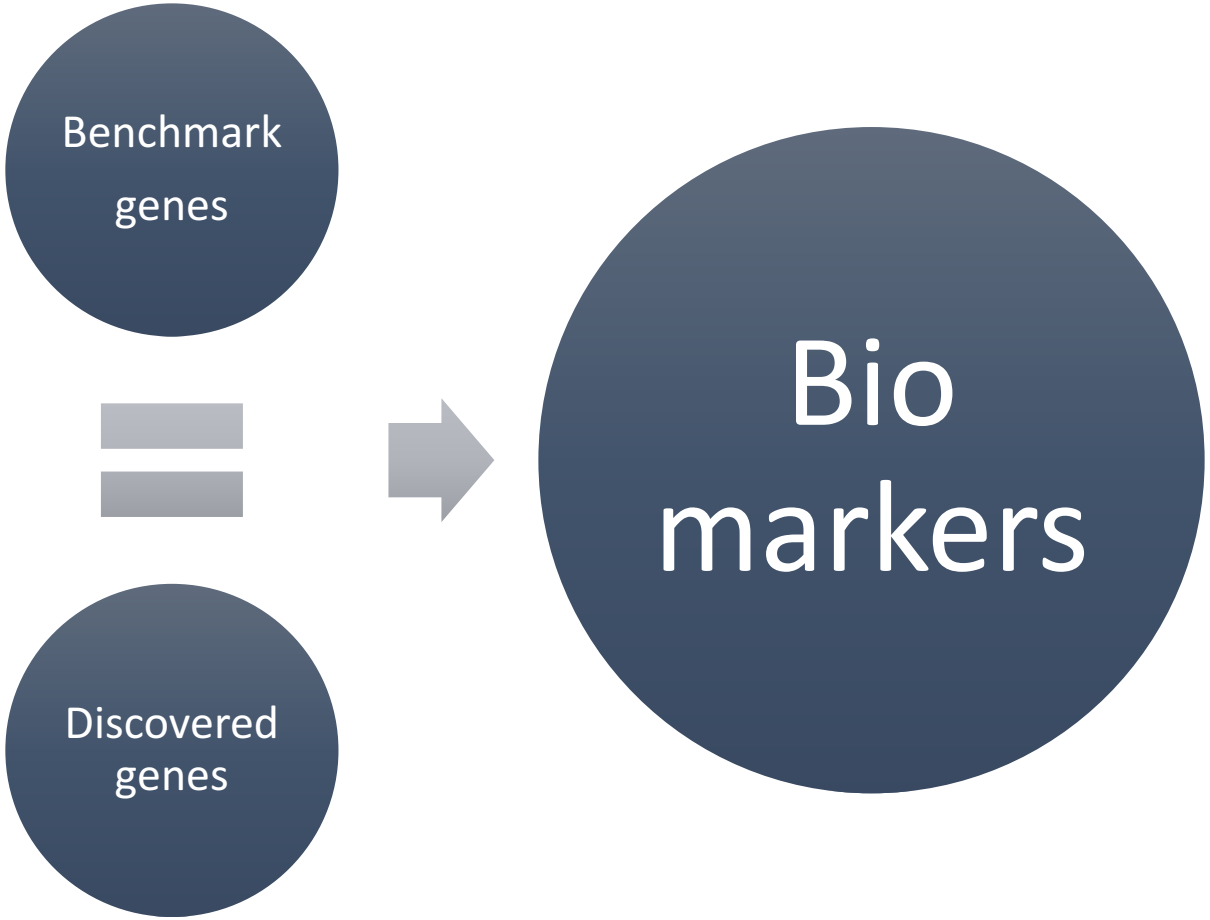
3-Clique-2-Bipartite Module: COAD



3-Clique-2-Bipartite Module: GBM



Validation



How metrics are created?

Chin *et al. BMC Systems Biology* 2014, **8**(Suppl 4):S11
http://www.biomedcentral.com/1752-0509/8/S4/S11

BMC Systems Biology

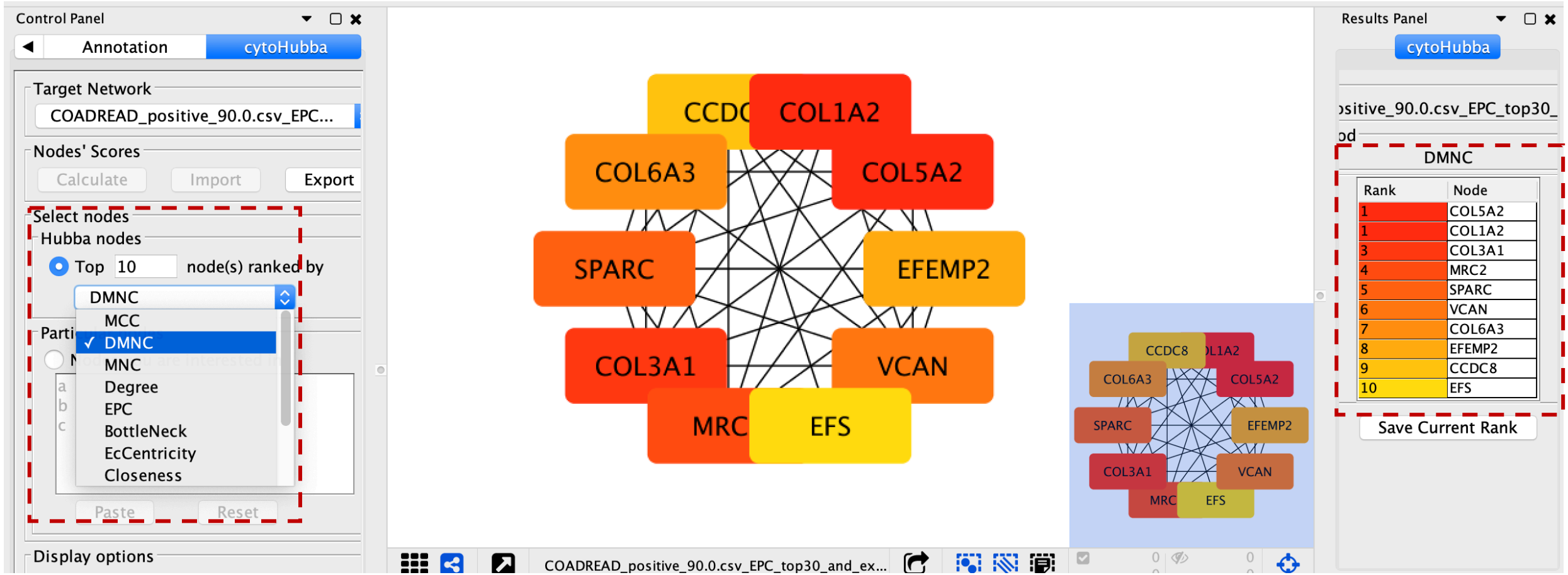
RESEARCH Open Access

cytoHubba: identifying hub objects and sub-networks from complex interactome

Chia-Hao Chin[1†], Shu-Hwa Chen[2†], Hsin-Hung Wu[6], Chin-Wen Ho[5], Ming-Tat Ko[2,6*], Chung-Yen Lin[2,3,4*]

From Asia Pacific Bioinformatics Network (APBioNet) Thirteenth International Conference on Bioinformatics (InCoB2014)
Sydney, Australia. 31 July - 2 August 2014

cytoHubba Interface



Scoring Methods

Graph Topological analysis methods

1. Betweenness
2. BottleNeck
3. ..... (Closeness, Clustering Coefficient (CC), **Degree**, Density of Maximum Neighborhood Component (DMNC), EcCentricity (EcC), Edge Percolated Component (EPC), Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Radiality, and Stress)

$$\text{Degree (v)} = |N(v)|$$

Top-10 genes in BRCA

Betweenness	PTPN7	SASH3	CD53	CD27	CD79A	HLA-DMB	CD48	WAS	IKZF1	FCRL5
Bottleneck	PTPN7	SASH3	TPX2	IKZF1	WAS	COL6A3	CD48	SIRPG	NCKAP1L	ASXL2
Closeness	PTPN7	CD3E	SH2D1A	SLAMF6	SASH3	CD247	CD2	CD3D	LY9	CD53
Degree	CD3E	SH2D1A	SLAMF6	CD247	CD2	CD3D	LY9	ITK	PTPN7	SASH3
EcC	ACAP1	CD247	CD27	CD3D	CD3E	CD5	CXCR3	IL2RG	LCK	LY9
EPC	CD247	SLAMF6	CD2	CD3E	SH2D1A	CD5	CD3D	IL2RG	UBASH3A	SLA2
MCC	CD3E	CD2	SH2D1A	SLAMF6	CD247	IL2RG	SLAMF1	SIRPG	UBASH3A	LY9
MNC	CD3E	SH2D1A	SLAMF6	CD247	CD2	CD3D	LY9	ITK	CD5	PTPN7
Radiability	PTPN7	SASH3	CD48	SH2D1A	SLAMF6	CD53	CD3E	IKZF1	CD247	CD3D
Stress	CD27	CD79A	FCRL5	IRF4	CD53	PTPN7	HLA-DMB	CD48	SASH3	SLAMF7
CC	CD6	BTK	CD40LG	THEMIS	BTLA	GPR171	SPARC	GZMK	NKG7	KLRK1
DMNC	CD27	LCK	SAMD3	SLAMF1	GZMK	IL2RG	SIRPG	UBASH3A	PYHIN1	CD5

Metrics

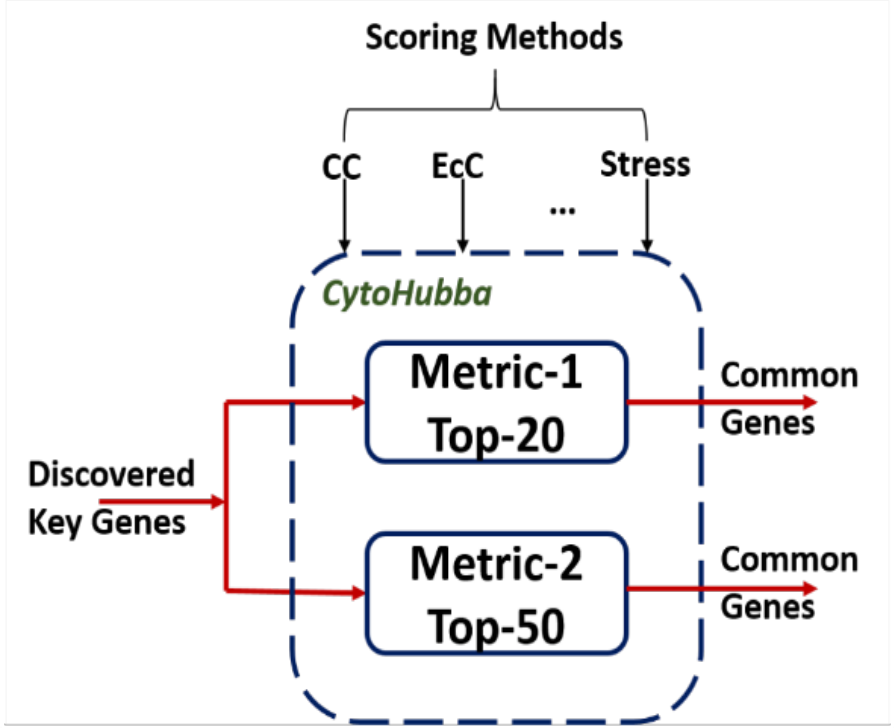
Top 20 and 50 genes from each methods

Gene appears >= 2 methods

Metric Name	BRCA	COAD	GBM
Metric-1 (Top 20)	41	53	42
Metric-2 (Top 50)	92	130	99

BRCA	COAD	GBM
47	61	38

Results



Metric Name	BRCA	COAD	GBM
Metric-1 (Top 20)	55%	37%	65%
Metric-2 (Top 50)	96%	70%	95%

Conclusion

This validation supports the proposed hypotheses that there exist clique-like and clique-bipartite-like structures in gene networks of diseases

Pathway Analysis

- A pathway has a set of genes related to a specific biological function and describes the relationship between the genes.
- Some of the databases with biological pathways are KEGG, Reactome, Pathguide, PANTHER.
- Identification of genes and proteins associated with a specific disease.

Pathway Analysis

BRCA	COAD	GBM
TCR signaling in -ve CD4+ T cells(N)	Neutrophil degranulation(R)	Neutrophil degranulation(R)
TCR signaling in -ve CD8+ T cells(N)	Osteoclast differentiation(K)	Osteoclast differentiation(K)
TCR signaling(R)	TCR signaling in -ve CD4+ T cells(N)	Natural killer cell mediated cytotoxicity(K)
IL12-mediated signaling events(N)	Staphylococcus aureus infection(K)	Fc gamma R-mediated phagocytosis(K)
Downstream signaling in -ve CD8+ T cells(N)	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell(R)	GPVI-mediated activation cascade(R)
T cell receptor signaling pathway(K)	Fc-epsilon receptor I signaling in mast cells(N)	Fcgamma receptor (FCGR) dependent phagocytosis(R)
T cell activation(P)	Natural killer cell mediated cytotoxicity(K)	Staphylococcus aureus infection(K)
Th1 and Th2 cell differentiation(K)	Cell adhesion molecules (CAMs)(K)	TCR signaling in -ve CD4+ T cells(N)
IL12 signaling mediated by STAT4(N)	Fc gamma R-mediated phagocytosis(K)	Fc epsilon RI signaling pathway(K)
Th17 cell differentiation(K)	T cell receptor signaling pathway(K)	Platelet activation(K)

Gene Ontology (GO)

- An ontology represents the knowledge within a given domain.
- GO is set of associations from biological phrases to specific genes
- GO terms are either chosen by trained curators or generated automatically.
- These are organized hierarchically – higher level terms are more general and are assigned to more genes.

Gene Ontology (GO)

- There are three GO categories:
  - Biological Process
    - Biological events accomplished by larger multiple molecular activities.
  - Cellular Component
    - Locations relative to cellular structures in which gene product performs a function.
  - Molecular Function
    - Molecular-level activity performed by gene products.

GO Biological Process

BRCA	COAD	GBM
TCR signaling in -ve CD4+ T cells(N)	Neutrophil degranulation(R)	Neutrophil degranulation(R)
TCR signaling in -ve CD8+ T cells(N)	Osteoclast differentiation(K)	Osteoclast differentiation(K)
TCR signaling(R)	TCR signaling in -ve CD4+ T cells(N)	Natural killer cell mediated cytotoxicity(K)
IL12-mediated signaling events(N)	Staphylococcus aureus infection(K)	Fc gamma R-mediated phagocytosis(K)
Downstream signaling in -ve CD8+ T cells(N)	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell(R)	GPVI-mediated activation cascade(R)
T cell receptor signaling pathway(K)	Fc-epsilon receptor I signaling in mast cells(N)	Fcgamma receptor (FCGR) dependent phagocytosis(R)
T cell activation(P)	Natural killer cell mediated cytotoxicity(K)	Staphylococcus aureus infection(K)
Th1 and Th2 cell differentiation(K)	Cell adhesion molecules (CAMs)(K)	TCR signaling in -ve CD4+ T cells(N)
IL12 signaling mediated by STAT4(N)	Fc gamma R-mediated phagocytosis(K)	Fc epsilon RI signaling pathway(K)
Th17 cell differentiation(K)	T cell receptor signaling pathway(K)	Platelet activation(K)

Pathway and GO Enrichment Analysis

- Reduces data dimensionality by arranging the genes into pathways and ontologies.
- Helps interpret the data in the context of biological process, pathways and networks.

Thank you!