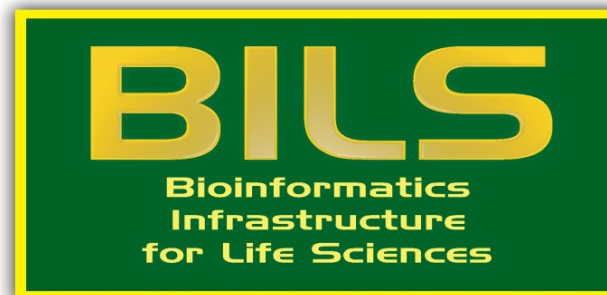
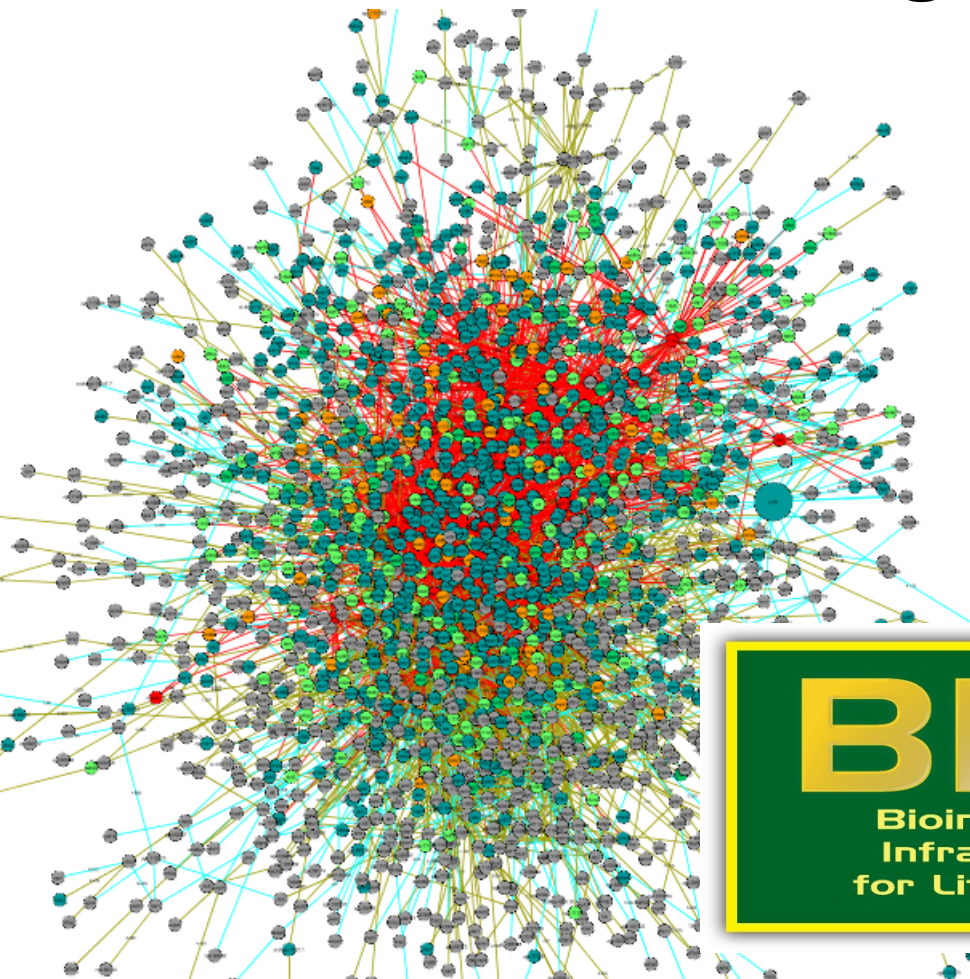


HyperSet: functional interpretation of biological data in gene/protein interaction networks

Andrey Alexeyenko



**Karolinska
Institutet**

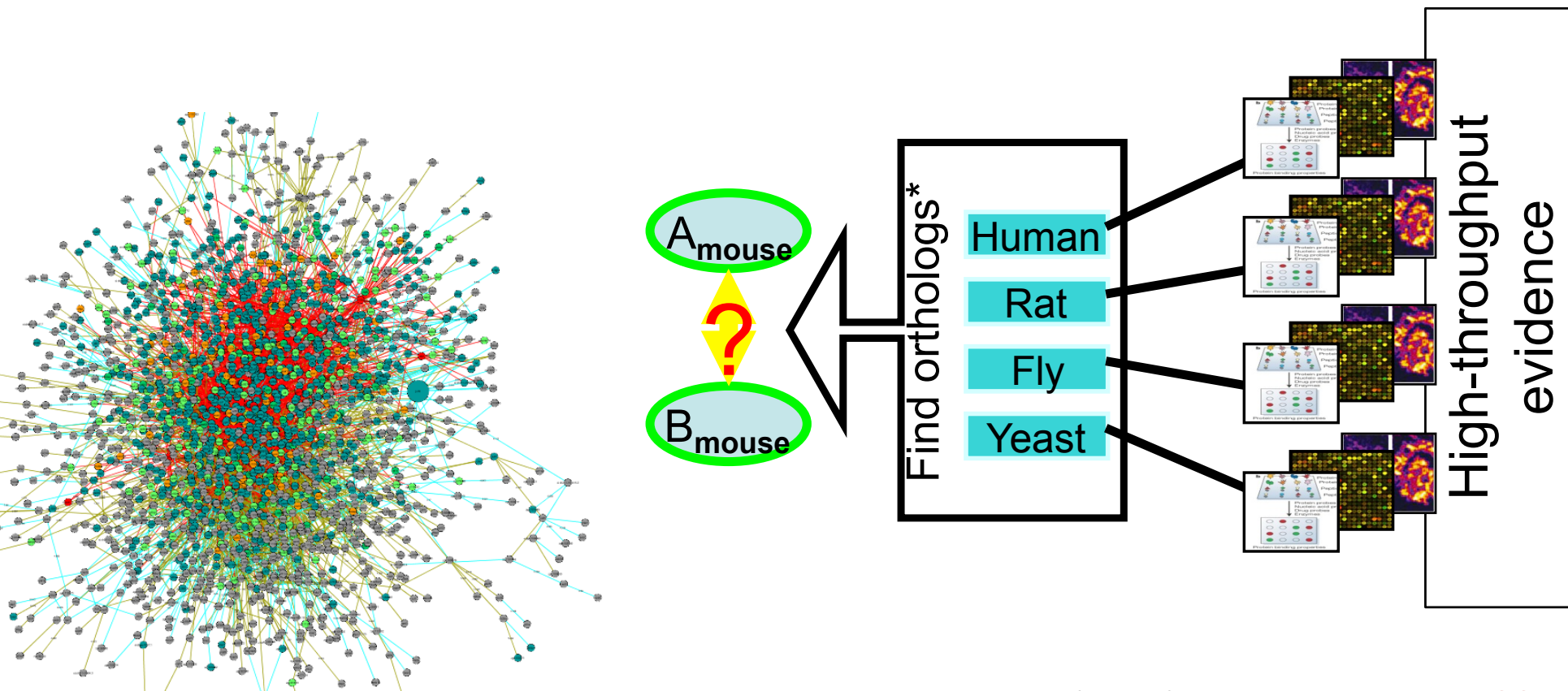
SciLifeLab

“Data is not information,
information is not **knowledge**,
knowledge is not wisdom,
wisdom is not truth,”

—*Robert Royar (1994)*



FunCoup is a data integration framework to discover functional coupling in eukaryotic proteomes with data from model organisms



Andrey Alexeyenko and Erik L.L. Sonnhammer (2009) **Global networks of functional coupling in eukaryotes from comprehensive data integration.** *Genome Research*.

State-of-the-art method to beat: Frequency analysis of somatic mutations

Do gene networks tell any story?

COSMIC

CGP Home
COSMIC Home
Help

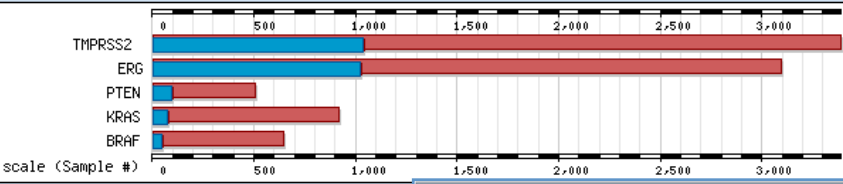
Current Selection
Tissues Chosen
Prostate
Sub Tissues Chosen
NS

Website Search
People Search
Library Services
Site Map
Feedback / Help

Tissue Overview for Prostate

The genes displayed below are associated with samples that have the tissue type you have selected.

» **Top 5 Genes with Sample Data**

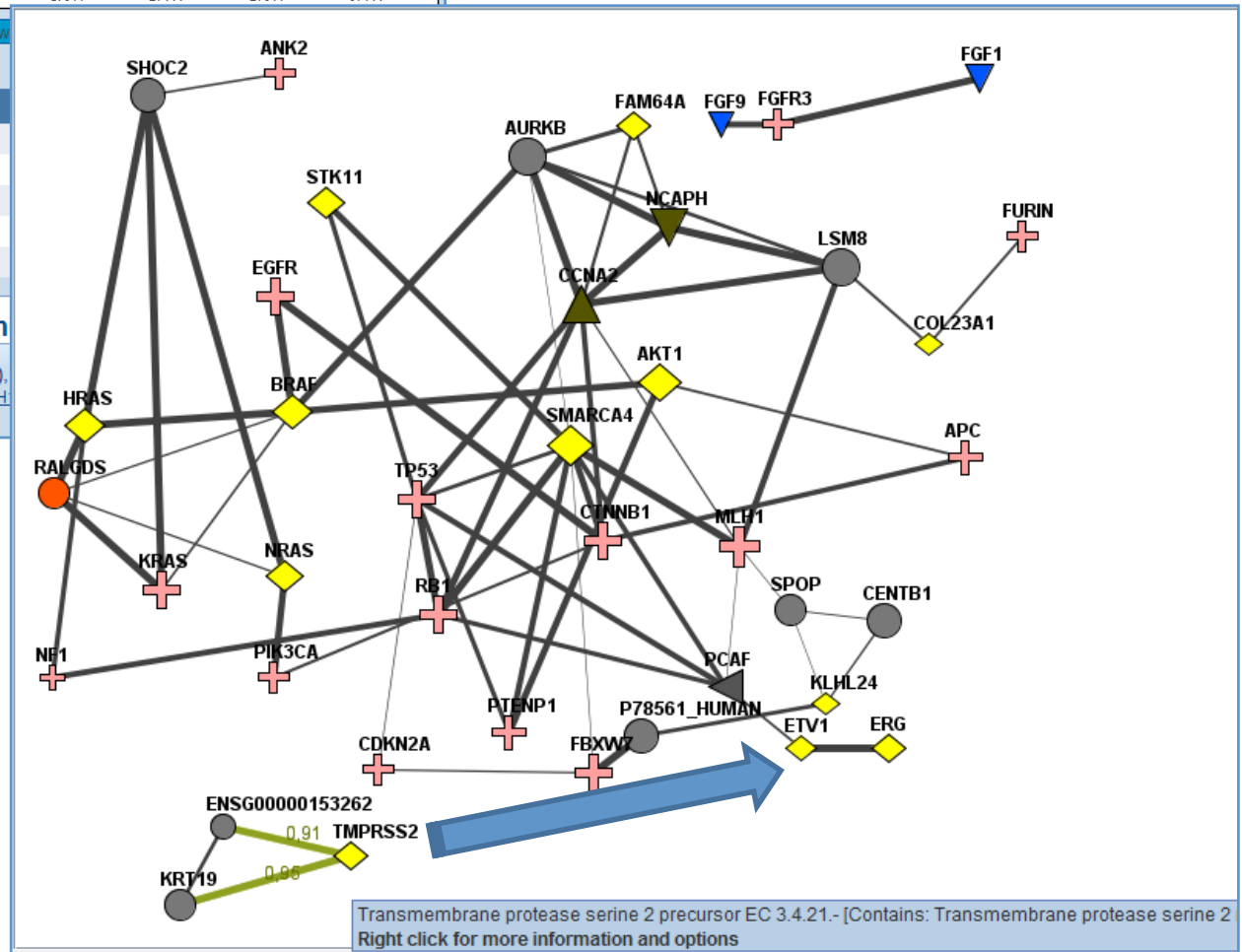


Gene Name	Sample Number
TMPRSS2	3384
ERG	3096
PTEN	506
KRAS	920
BRAF	643

» **Other genes with mutations for th**

[HRAS](#) (29), [CTNNB1](#) (22), [ETV1](#) (11), [CDKN2A](#) (11), [APC](#) (3), [SMARCA4](#) (3), [FBXW7](#) (2), [PIK3CA](#) (2), [IDH](#) (1)

- **Yellow diamonds:** somatic mutations in prostate cancer
- **Pink crosses:** also mutated in glioblastome (TCGA)



Network analysis

State-of-the-art:

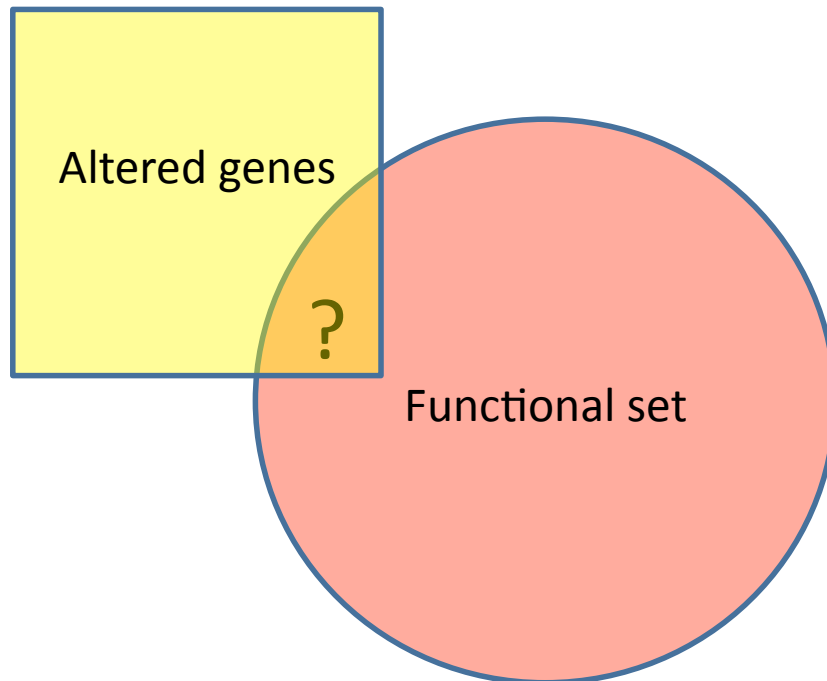
- The global gene/protein networks are widely available
- They do contain biological signal
- It is possible to use them in biological research
- Network visualization is just great!

Used for:

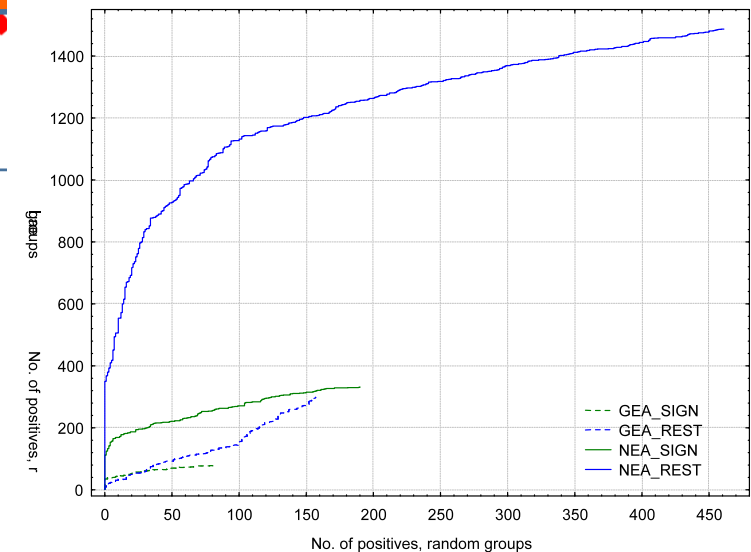
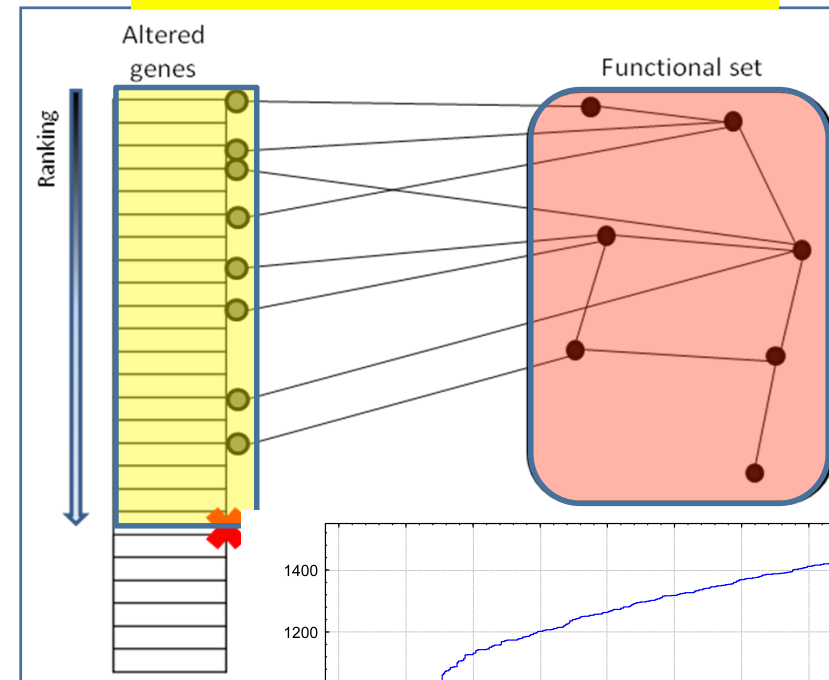
- Exploratory analysis
- Functional annotation
- Hypothesis formulation.
- Hypothesis testing, significance evaluation
- Network-based data transformation and processing

Functional characterization of novel gene sets

State-of-the-art method to beat:
Gene set enrichment analysis



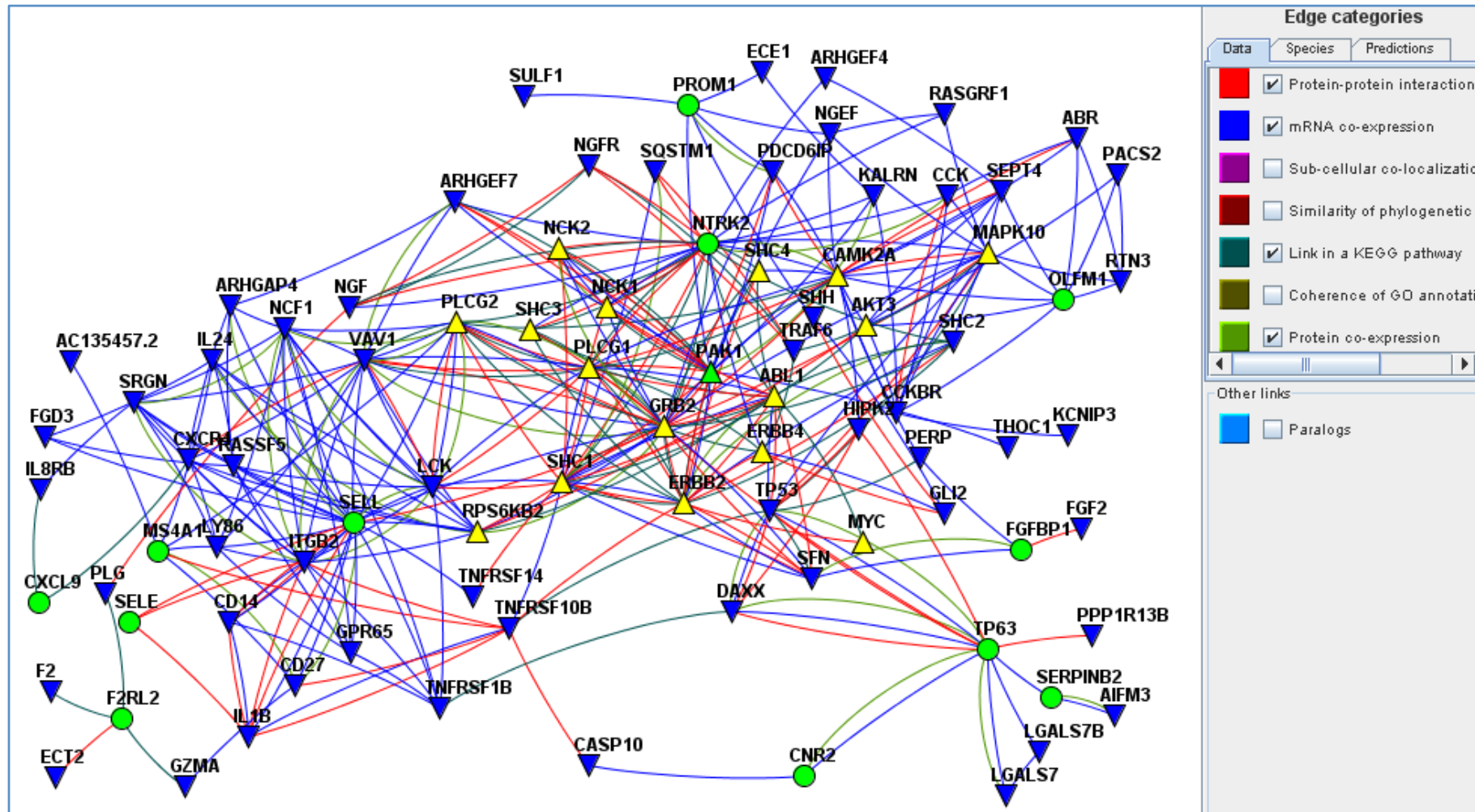
Our alternative:
Network enrichment analysis



Alexeyenko A, Lee W, ... Pawitan Y. **Network enrichment analysis**: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, 2012

Network analysis

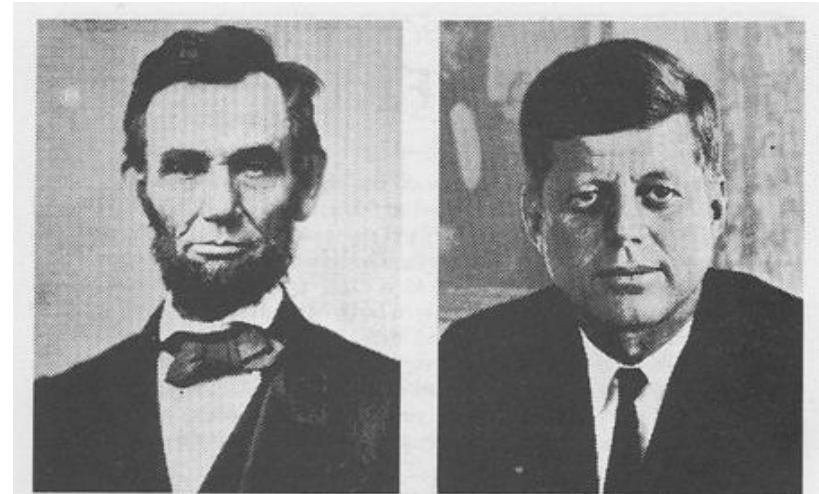
Question: How to prove that multiple altered genes have a common feature?



Altered genes (green) from one individual lung cancer are enriched in network connections to members of "ErbB pathway" (yellow) and "apoptosis" (blue).

Apophenia: the human propensity to see meaningful patterns in random data

(Brugger 2001; Fyfe et al. 2008)



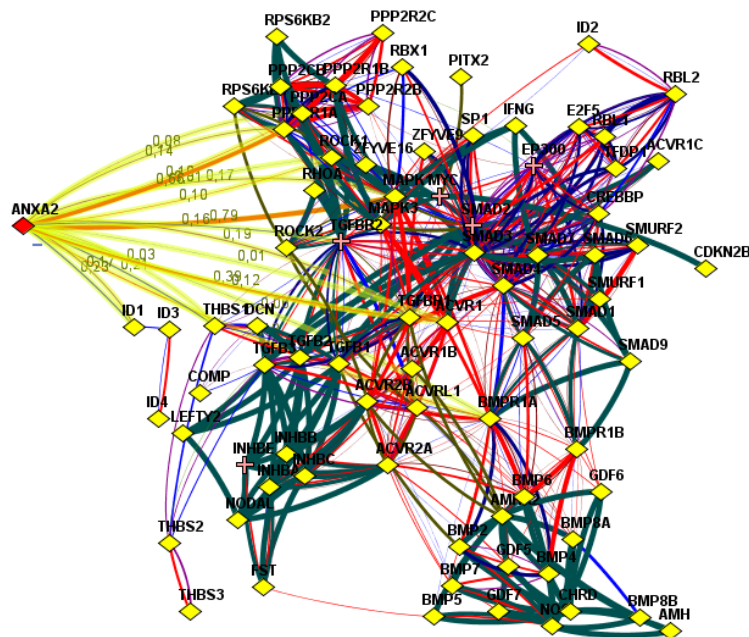
Coincidence or Not?

- Both the names "Kennedy" and "Lincoln" have 7 letters.
- Both presidents were assassinated by men with 15 letters in their names: Lee Harvey Oswald and John Wilkes Booth.
- Lincoln was shot in a theater and his assassin ran to a warehouse; Kennedy was shot from a warehouse and his assassin ran to a theater.
- Lincoln's secretary was named Kennedy; Kennedy's secretary was named Lincoln.
- Both men were succeeded in office by men named "Johnson"; the names "Andrew Johnson" and "Lyndon Johnson" both have 13 letters.
- Lincoln was a Republican; Kennedy was a Demmocratt. Both parties have 10 letters in their names.
- Both men left widows who eventually married Greek shipping magnates; both men had children named Tod.
- Mrs. Kennedy liked bananas; Mrs. Lincoln went bananas.
- Both men feared nuclear war with the Soviet Union.
- Kennedy's middle name was "Gettysburg"; Lincoln's middle name was "Bay of Pigs."
- Kennedy was a Catholic; Lincoln was a Satanist. Both religions have 8 letters in their names.
- Both men had affairs with Marilyn Monroe; both men were played by William Devane in television movies.
- Lincoln fought a war against "The South"; Kennedy fought a war against "Vietnam." Both countries have sneaky, devious citizens.
- Both men were "President" at the time they were shot; the word "President" has 9 letters.

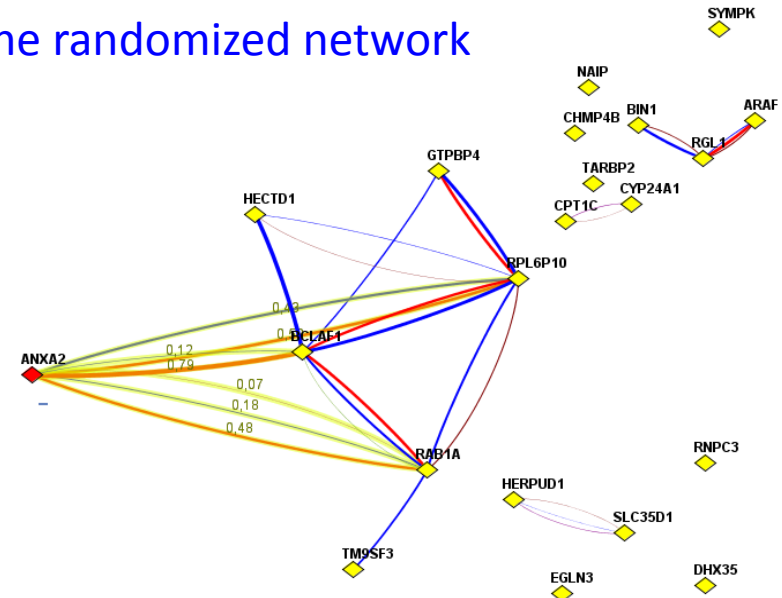
Network enrichment analysis

How to prove significance?
compare to a reference and quantify

In the actual network



In the randomized network



Question:

Is ANXA2 related to TGFbeta signaling?

Quantification:

$N \text{ links}_{real} = 12$

$N \text{ links}_{expected} = 4.65$

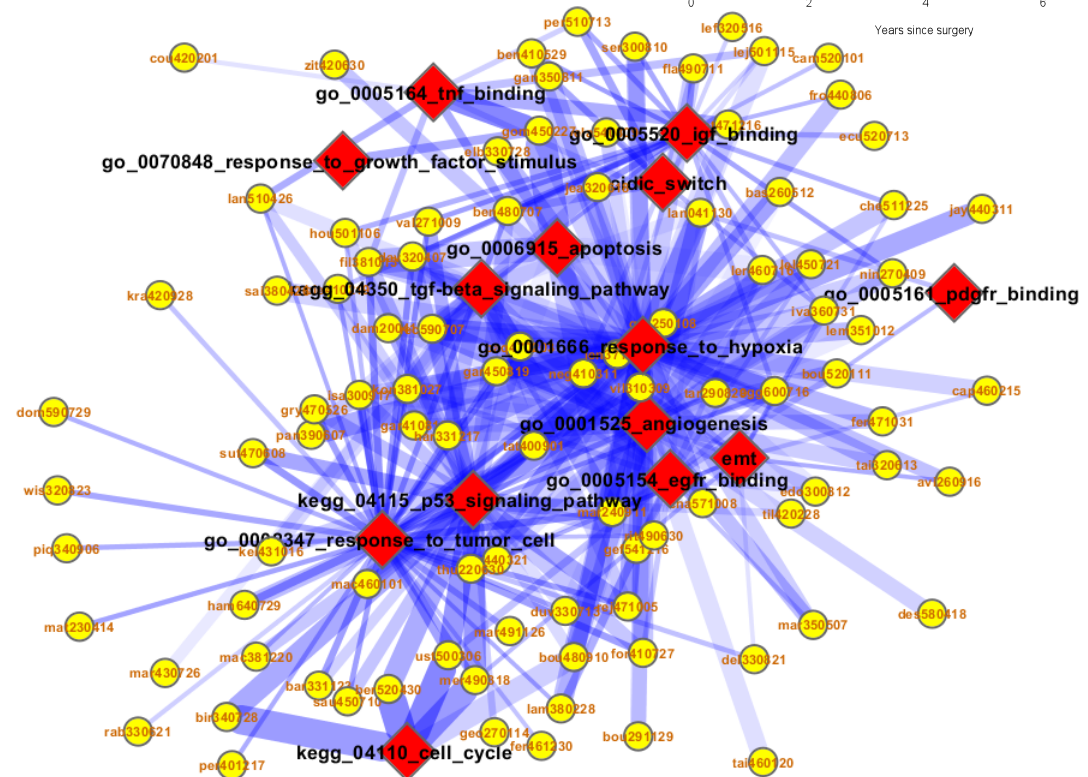
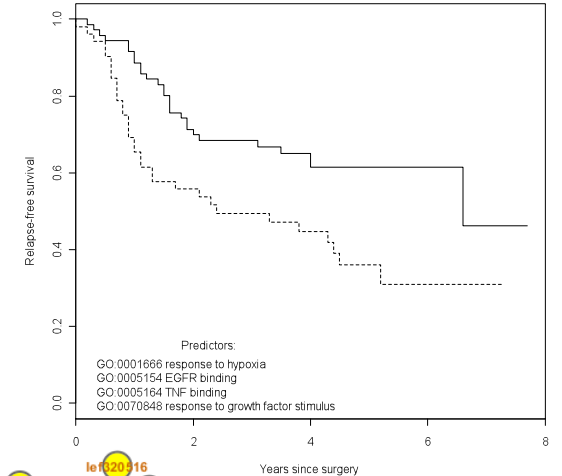
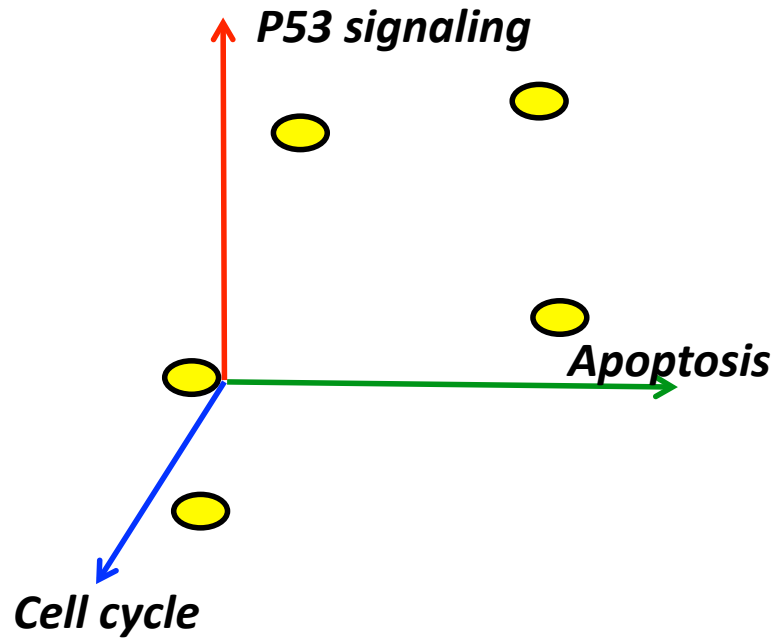
Standard deviation = 1.84

$Z = (N \text{ links}_{observed} - N \text{ links}_{expected}) / SD = 3.98$

P-value = 0.0000344

FDR < 0.1

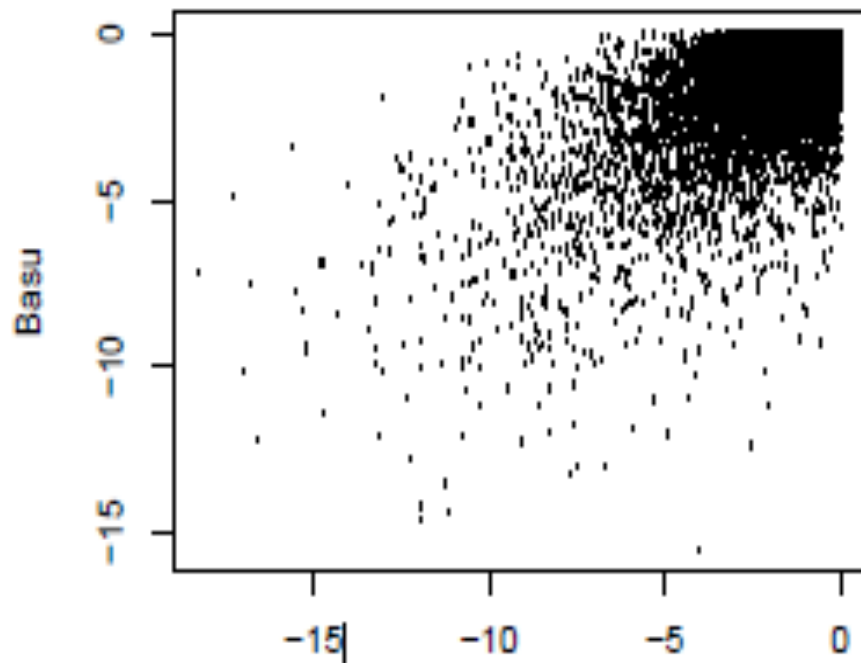
Molecular phenotypes in pathway space



P-values of correlation between cell line sensitivity to anti-cancer drug *Atra* and the cell line transcriptome

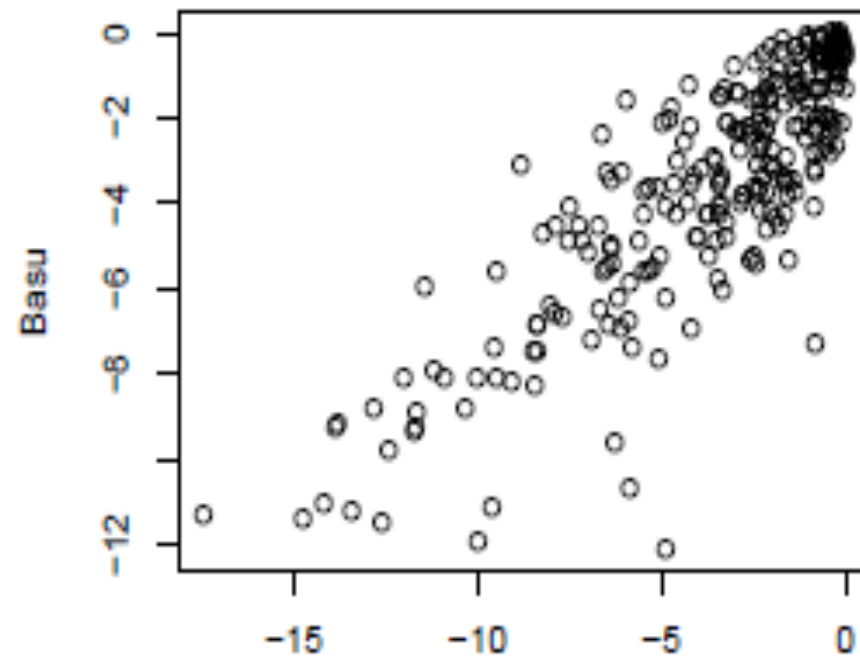
Log₁₀(p) values are plotted for the two drug screens, identified by the first author

Gene expression profiles



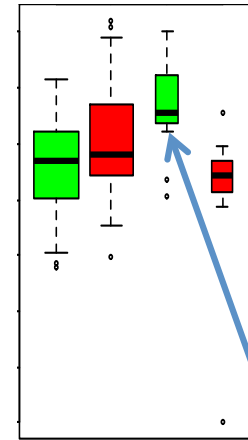
Garnett
Spearman correlation = 0.44421

NEA pathway scores for cell line-specific gene lists



Garnett
Spearman correlation = 0.82368

Resistance to vinorelbine in lung cancer



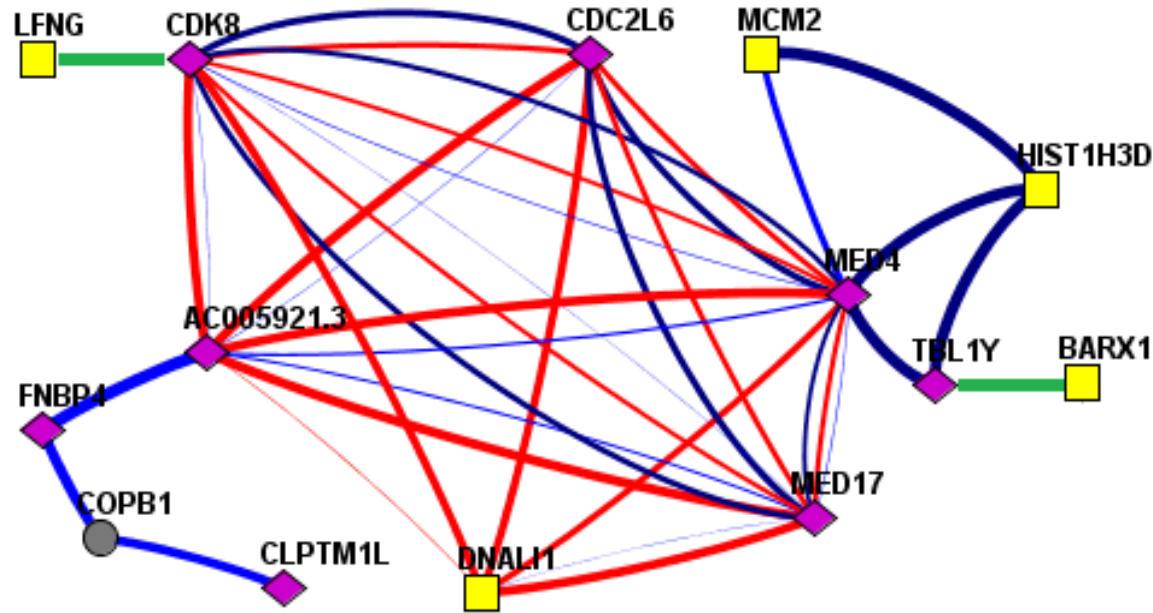
Question: What predicts tumor resistance to chemotherapy?

Answer: Depletion of differential transcriptome towards few specific genes

		Relapse	
		+	-
Treatment	+	A	B
	-	C	D

Blue arrows point from the 'Relapse +' column to the green box in the box plot above, and from the 'Treatment +' row to the red box in the box plot above.

Resistance to vinorelbine in lung cancer



Functionally coherent genes associated with vinorelbine resistance.

- A. Network representation of the group. Magenta: genes associated with resistance in NEA and likely producing a protein complex (ranked 1, 3, 5, 6, 11, and 18) plus one more gene CLPTM1L (beyond the ranking but also significantly associated, was previously reported as related to cisplatin resistance); yellow: non-commonly expressed genes linked with the NEA genes and less represented in the susceptible tumors, hence most contributing to the association (see Results for more explanation).
- B. Box-plots of NEA scores for the genes colored magenta in A.

Science 29 March 2013:

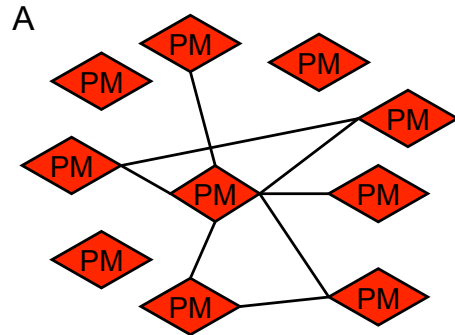
Cancer Genome Landscapes

[Bert Vogelstein](#), [Nickolas Papadopoulos](#), [Victor E. Velculescu](#),
[Shibin Zhou](#), [Luis A. Diaz Jr.](#), [Kenneth W. Kinzler*](#)

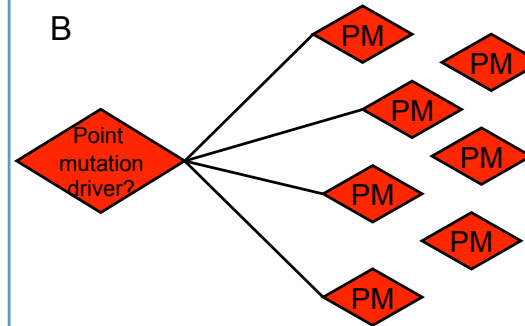
"Though all 20,000 protein-coding genes have been evaluated in the genome-wide sequencing studies of 3284 tumors, with a total of 294,881 mutations reported, only 125 Mut-driver genes, as defined by the 20/20 rule, have been discovered to date (table S2A). Of these, 71 are tumor suppressor genes and 54 are oncogenes. An important but relatively small fraction (29%) of these genes was discovered to be mutated through unbiased genome-wide sequencing; most of these genes had already been identified by previous, more directed investigations. "

"At best, methods based on mutation frequency can only prioritize genes for further analysis but cannot unambiguously identify driver genes that are mutated at relatively low frequencies"

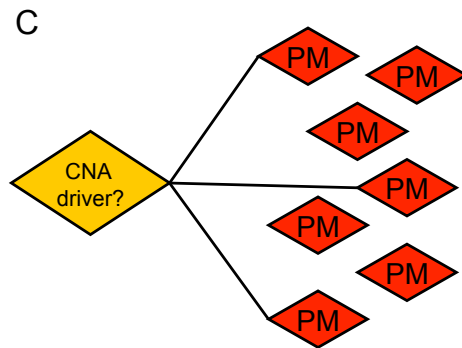
Could the emerged mutations interact with each other?



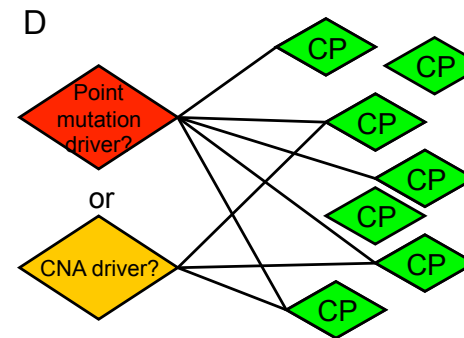
Could a given mutation interact with the others?



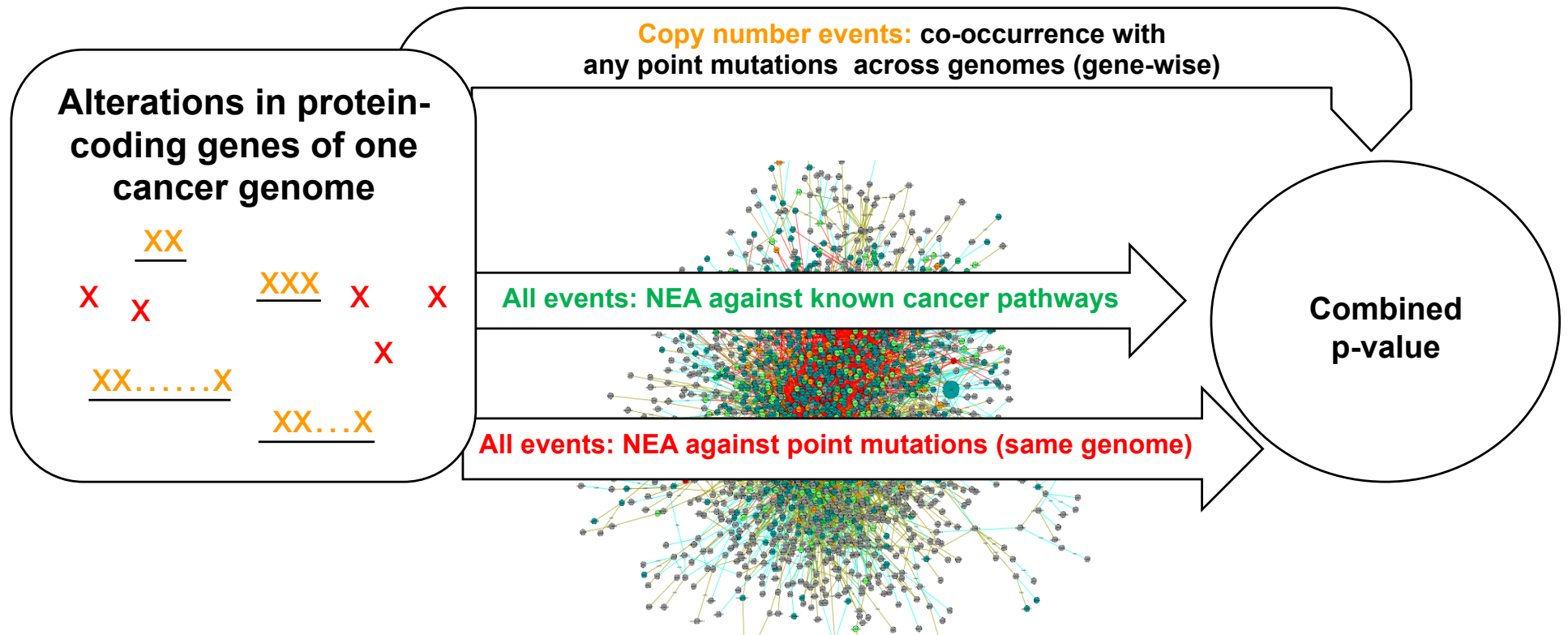
Could the emerged gene copy number change interact with the point mutations?



Could the emerged mutations (either point or copy number change) interact with known cancer pathways?

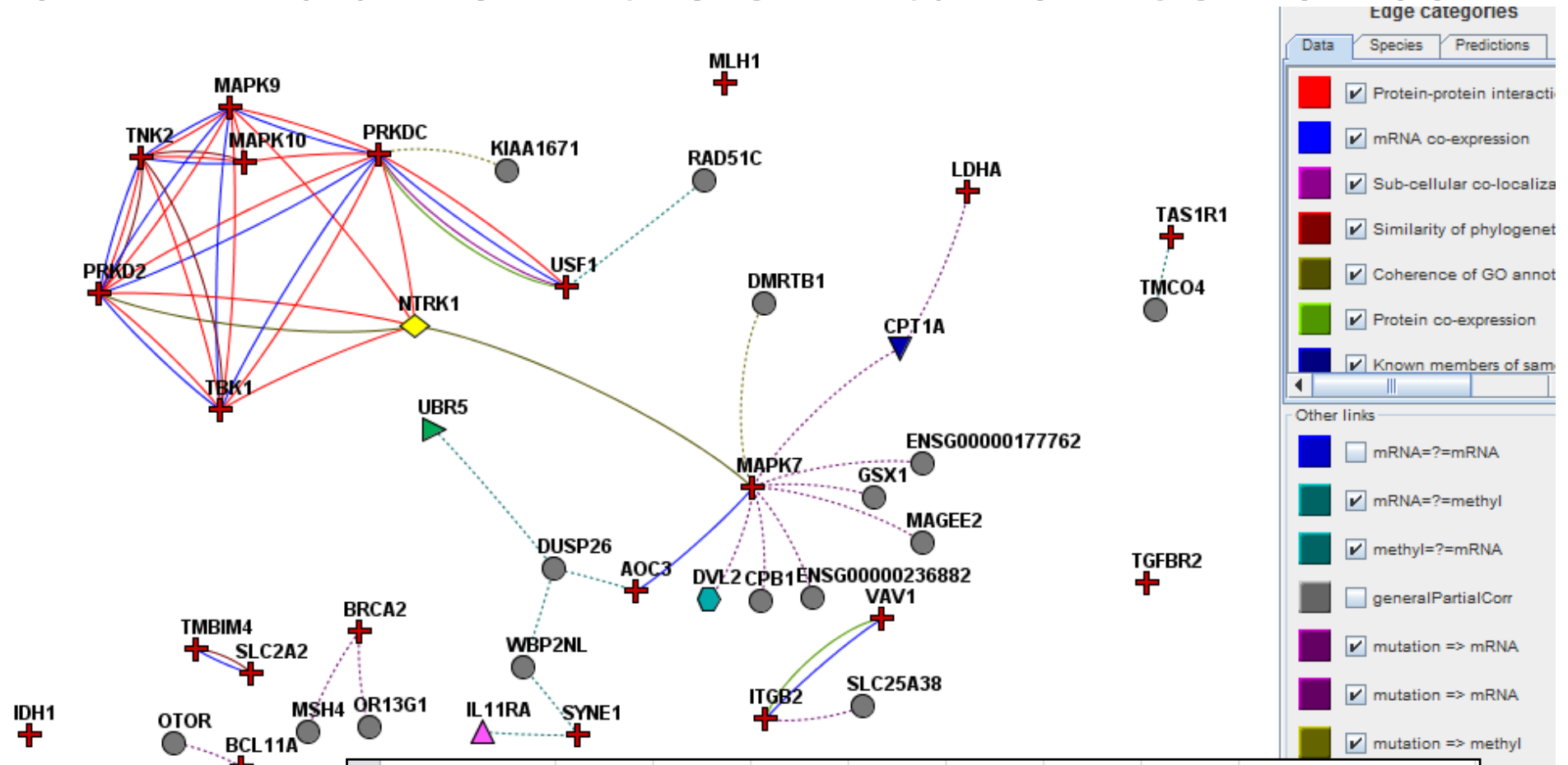


Simon Kebede Merid, Daria Goranskaya, Andrey Alexeyenko **Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis** *BMC Bioinformatics*, in press



Simon Kebede Merid, Daria Goranskaya, Andrey Alexeyenko
Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis
BMC Bioinformatics, in press.

Is NTRK1 a driver in the GBM tumor TCGA-02-0014?



	Tumor	N mutations	N links total	Mutation	N links observed	N links expected	SD	Z	FDR
1	tcga-02-0014-10	35	5495	brca2	5	0.16	0.374	12.9354	1.86909279270073e-33
2	tcga-02-0014-10	35	5495	mapk7	5	0.36	0.569	8.1600	1.01768292406201e-12
3	tcga-02-0014-10	35	5495	mapk9	6	0.76	0.779	6.7275	2.459600385499e-08
4	tcga-02-0014-10	35	5495	ntrk1	5	0.48	0.586	7.7140	3.63383661524901e-11
5	tcga-02-0014-10	35	5495	prkd2	5	0.64	0.638	6.8370	1.15540921257264e-08
6	tcga-02-0014-10	35	5495	slc2a2	4	0.24	0.436	8.6260	1.9404085486166e-14
7	tcga-02-0014-10	35	5495	stk36	3	0.12	0.332	8.6835	1.18130731254268e-14
8	tcga-02-0014-10	35	5495	tbk1	6	0.80	0.764	6.8084	1.40823249351197e-08

HyperSet: network analysis made practical

Project ID

Organism:

NEA & GSEA

Driver mutation analysis

Venn space

Does it look like a pathway?

How transcriptome alterations relate to "non-transcriptional" pathways?

How different high-throughput platform data relate to each other?

Chip-Seq

Prioritize candidate genes

Network-constrained gene signatures

Candidate disease genes

Network

Functional gene sets

Check and submit

Results

Project ID HYPERSET.ne

Organism: human

Altered gene sets

Network

Functional gene sets

Check and submit

Results

Project ID HYPERSET.ne

Organism: human

Altered gene sets

Network

Functional gene sets

Check and submit

Results

Select color Project ID HYPERSET.ne

Organism: human

Altered gene sets

Network

Functional gene sets

Check and submit

Results

- B
- C
- G
- G
- G
- K
- K
- K
- K
- M
- R
- T
- g
- W

CHECKLIST:

Selected AGS: fl2rd_fdr_0.141_top100 & fl2v_fdr_0.002_top100 & rd2v_fdr_0.004_top100

Selected network: merged_high_quality

Selected FGS: KEGG pathways, signaling

Max FDR 0.00001

Show genes behind enrichment

Filter AGS (your groups) by mask:

Filter FGS (pathways) by mask: kegg_04

Employ statistic Chi-squared (fast)

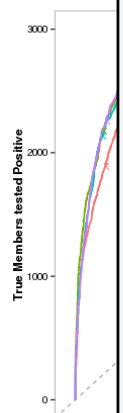
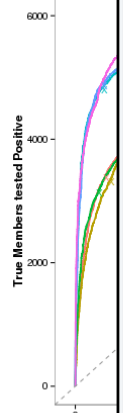
Enable table view

Enable network view using arbor network layout

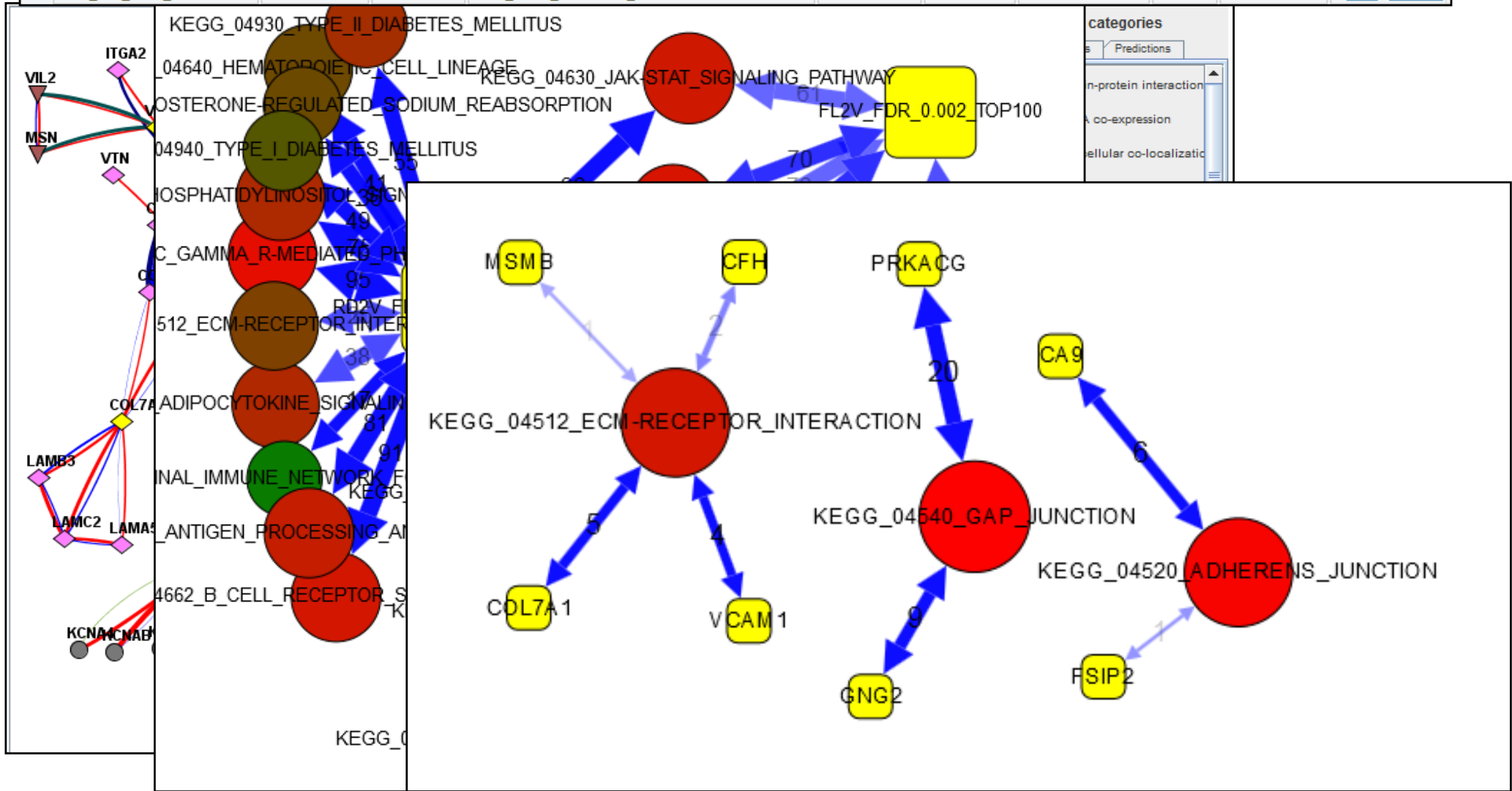
Analyze the AGS genes/proteins individually

Analyze the FGS genes/proteins individually

Submit Do not calculate, just restore the last analysis



AGS	#genes AGS	#links AGS	FGS	#genes FGS	#links FGS	#linksAGS2FGS	Score	FDR	Shared genes
RD2V_FDR_0.004_TOP100	98	3143	KEGG_04512_ECM-RECEPTOR_INTERACTION	84	5719	25	26.99	4.608675E-07	0 FClim FC3small
RD2V_FDR_0.004_TOP100	98	3143	KEGG_04540_GAP_JUNCTION	99	23898	63	15.52	1.411460E-04	1 FClim FC3small
RD2V_FDR_0.004_TOP100	98	3143	KEGG_04520_ADHERENS_JUNCTION	79	20313	49	8.05	7.309125E-03	0 FClim FC3small



Choose CSV File

Choose File DE12cmp.v2.c...0lines2.txt

Upload complete

 Header

Separator:

 Comma Semicolon Tab

Quote:

 None Double Quote Single Quote

Number of Comparisons:

 2 3 4

Comparison 1:

Padj: FC:

X5.Wt_d3.5S_X9.Wt_d5.5S

0.05

1

Comparison 2:

Padj: FC:

X6.Wt_SB_d3.5S_X10.Wt_SB_d5

0.05

1

Comparison 3:

Padj: FC:

X7.Nkx2.2_d3.5S_X11.Nkx2.2_d5

0.05

1

VennDiagram generation based upon the input fields from left menu

Click on the numbers to see the Genes

Filtered table to view and download the lists

Venn Diagram

Filtered Table



	Genes
1	Dazap2
2	Kat2b
3	Dbh
4	Araf
5	Spa17
6	Slc39a13
7	Rab11b-ps2
8	Ikzf4
9	Grin2d

HyperSet: network analysis made practical

Project ID

Organism:

NEA & GSEA

Driver mutation analysis

Venn space

Does it look like a pathway?

How transcriptome alterations relate to "non-transcriptional" pathways?

How different high-throughput platform data relate to each other?

Chip-Seq

Prioritize candidate genes

Network-constrained gene signatures


Candidate disease genes

Network



Functional gene sets

Check and submit

Results

 ANDREY ALEXEYENKO HYPERSET

andrey alexeyenko hyperset

ILS infrast...  Latest Headlines  KTH VPN Service (SSL)

Network analysis: how to succeed?

- Analyze ***prioritized*** candidates (from genotyping, DE, GWAS...) rather than ***any*** genes.
- Do not lean on single “interesting” network links. Employ statistics!
i.e.
“concrete questions” => “testable hypotheses” => “concrete answers”

The amount of information in known gene networks is
enormous.

Let's just use it!

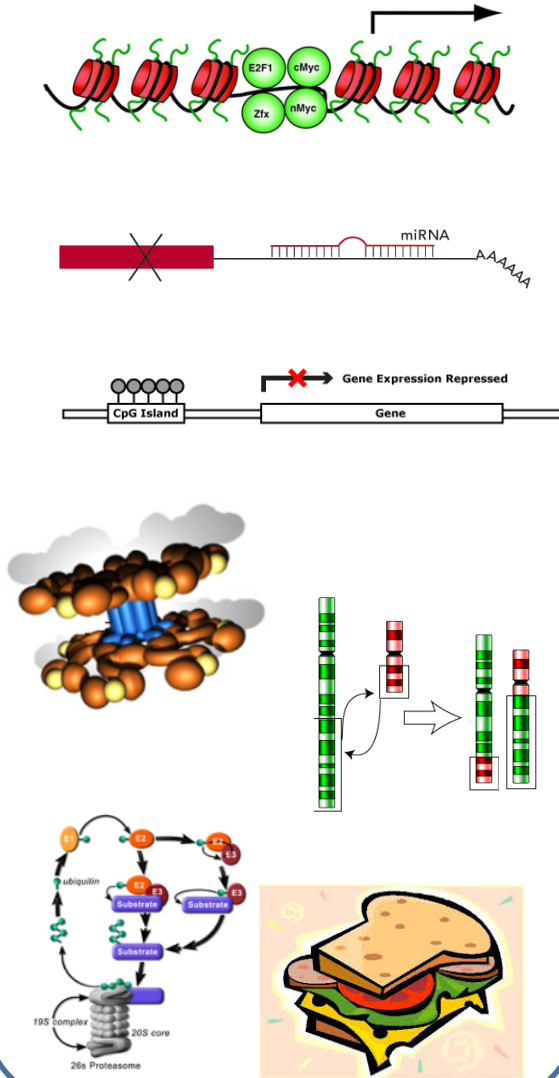
- Alexeyenko A, Wassenberg DM, Lobenhofer EK, Yen J, Linney E, Sonnhammer EL, Meyer JN. **Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity.** *PLoS One*. 2010 May 5;5(5):e10465.
- Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. **Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.** *BMC Bioinformatics*. 2012 Sep 11;13:226.
- McCormack T, Frings O, Alexeyenko A, Sonnhammer EL. **Statistical assessment of crosstalk enrichment between gene groups in biological networks.** *PLoS One*. 2013;8(1):e54945.
- Frings O, Alexeyenko A, Sonnhammer EL. **MGclus: network clustering employing shared neighbors.** *Mol Biosyst*. 2013 Jul;9(7):1670-5

Acknowledgements

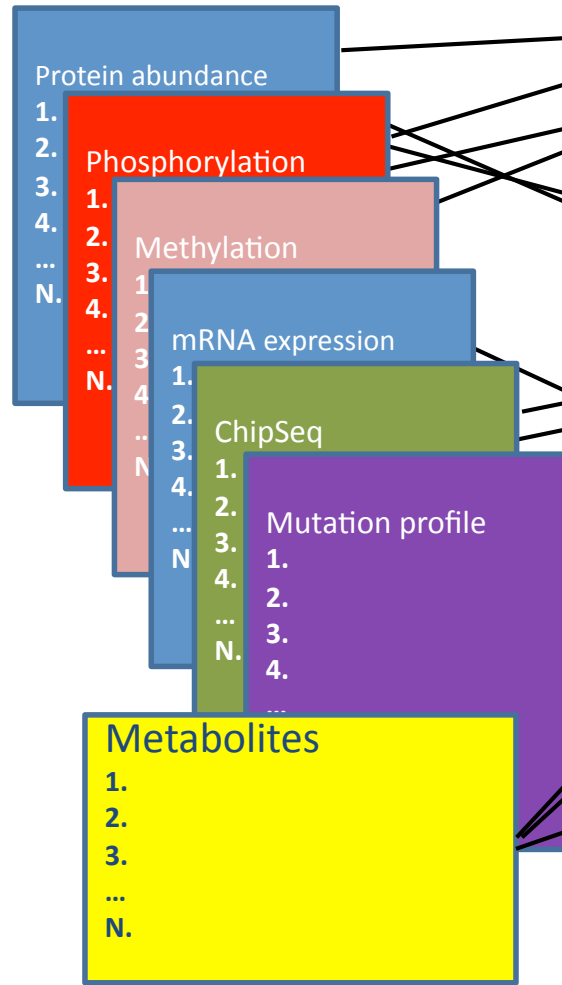
- Simon Merid
- Ashwini Jeggari
- Darya Goranskaya
- Pan Lu
- Erik Sonnhammer
 - Martin Klammer
 - Sanjit Roopra
 - Ted McCormack
 - Oliver Frings
- Yudi Pawitan
 - Setia Pramana
 - WooJoo Lee
- Joakim Lundeberg
- Pelin Akan
- Ingemar Ernberg
- Jonathan Prince
- Joel Meyer

Pathway analysis: why needed, what it is?

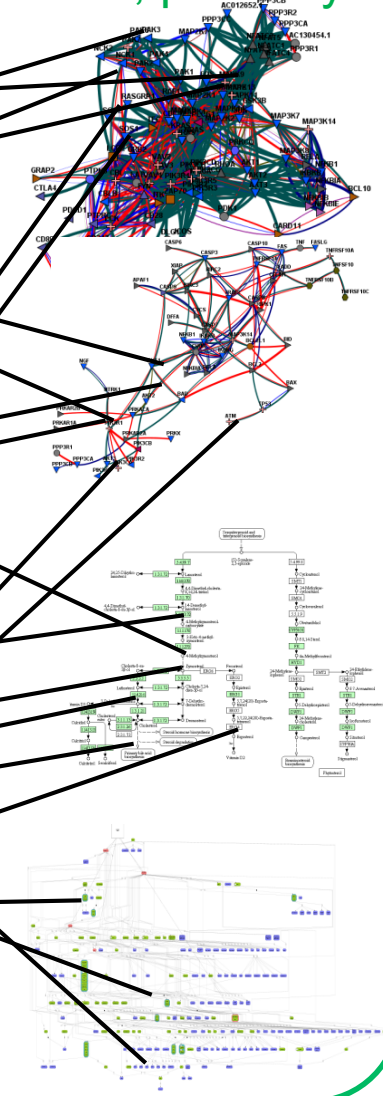
Primary molecular processes



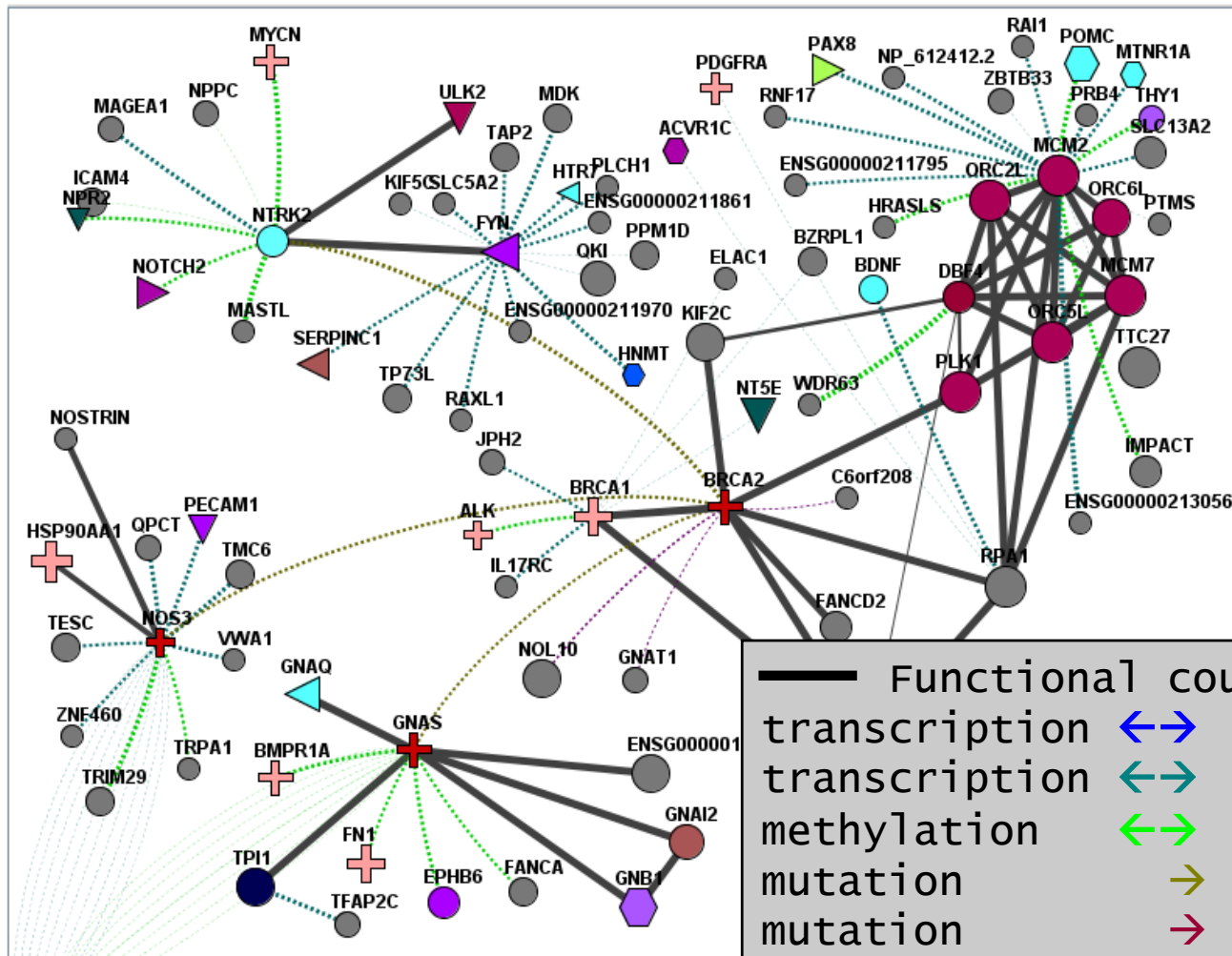
Response, recorded with high-throughput platforms



Known biological units: processes, pathways



Cancer-specific networks: links inferred from expression, methylation, mutations



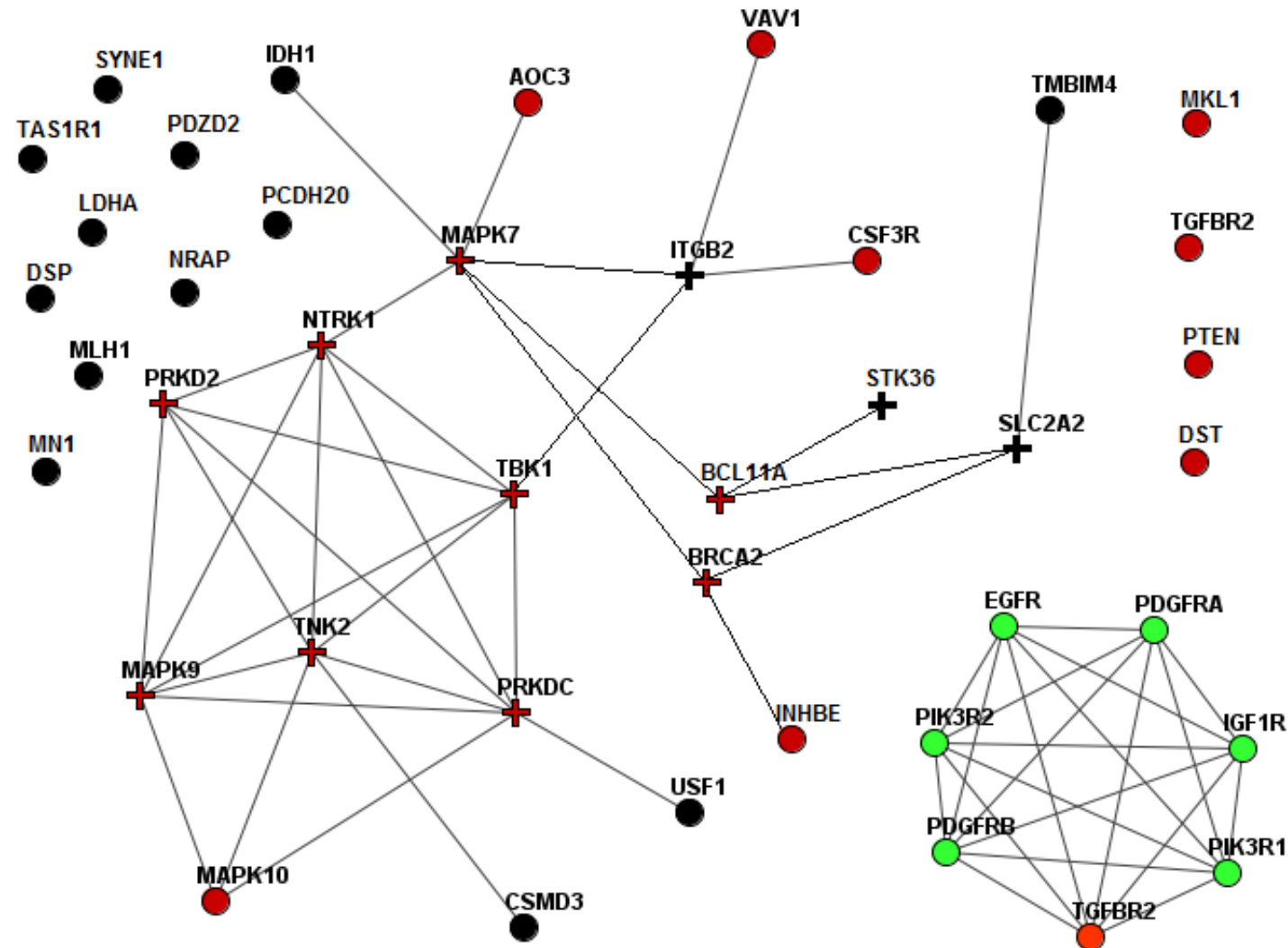
State-of-the-art
method to beat:
Reverse
engineering
from a single
source (usually
transcriptome)

—	Functional coupling
↔	transcription
↔	methylation
↔	methylation
→	methylation
→	transcription
↔	mutation
+	mutated gene

Now:

answer biological questions

Somatic mutations: drivers vs. passengers data from The Cancer Genome Atlas



Validation of candidate disease genes

(work with Jonathan Prince, MEB, KI)

[Genetic association of sequence variants near AGER/NOTCH4 and dementia.](#)

Bennet AM, Reynolds CA, Eriksson UK, Hong MG, Blenn...
Prince JA.

J Alzheimers Dis. 2011;24(3):475-84.

[Genome-wide pathway analysis implicates intracellular transport disease.](#)

Hong MG, Alexeyenko A...
J Hum Genet. 2010 Oct;

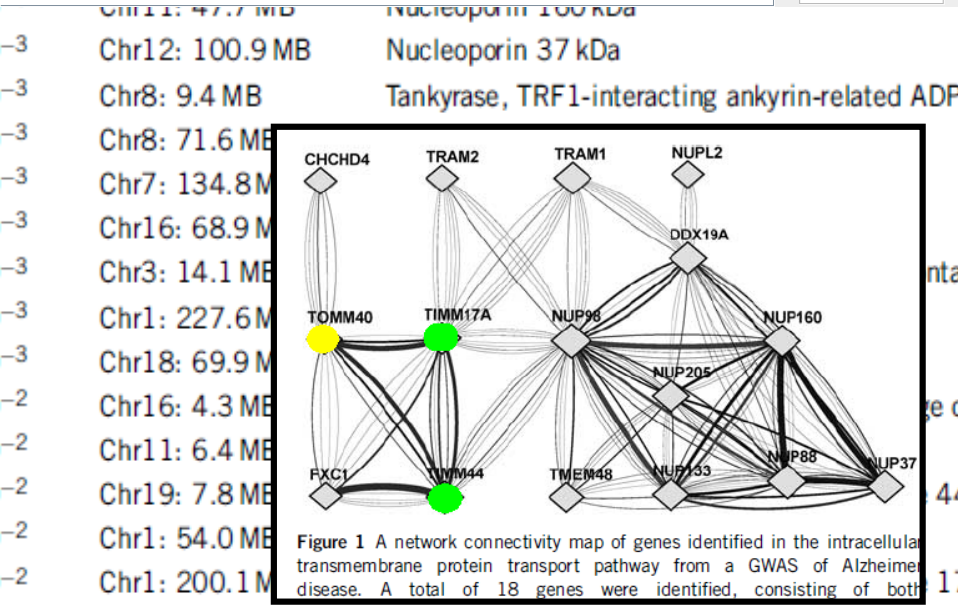
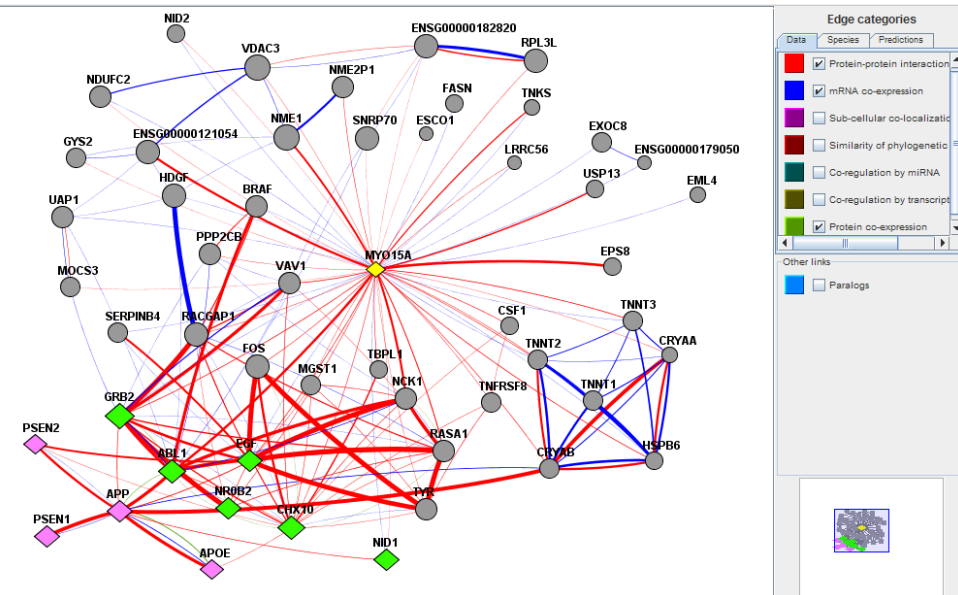
[Analysis of lipid pathway /ATPAF2 with dementia r](#)

Reynolds CA, Hong MG, Alexeyenko A, Grönber...
Hum Mol Genet. 2010 M

Question: Is there extra evidence for GWAS-candidates to be involved?
Answer: Yes, for some...

Table 1 Enriched genes in Alzheimer

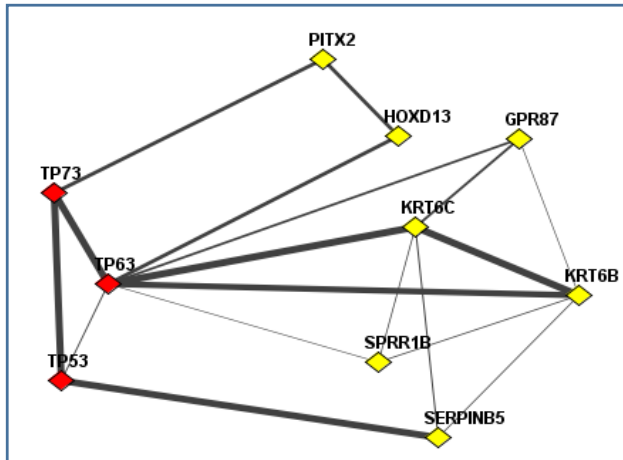
Gene	Best marker	P-value
NUP98	rs276885	6.68×10 ⁻³
NUPL2	rs858238	2.00×10 ⁻³
TRAM2	rs6928665	4.00×10 ⁻³
NUP88	rs6502860	1.00×10 ⁻³
NUP160	rs7951180	1.20×10 ⁻³
NUP37	rs950945	2.70×10 ⁻³
TNKS	rs6601327	2.90×10 ⁻³
TRAM1	rs268652	2.90×10 ⁻³
NUP205	rs11984203	5.00×10 ⁻³
DDX19A	rs8059245	5.40×10 ⁻³
CHCHD4	rs4685078	5.50×10 ⁻³
NUP133	rs927204	5.70×10 ⁻³
C18orf55	rs17062282	9.00×10 ⁻³
Magmas	rs611704	1.10×10 ⁻²
FXC1	rs4758423	1.10×10 ⁻²
TIMM44	rs12983784	1.20×10 ⁻²
TMEM48	rs1181145	1.40×10 ⁻²
TIMM17A	rs2820306	1.60×10 ⁻²



Network enrichment analysis: *applications*

State-of-the-art method to beat: Observational science

Pathway characterization



Question:

Does gene expression in *MAT230414* relate to “response to tumor cell”?

$N \text{ links}_{real} = 6$

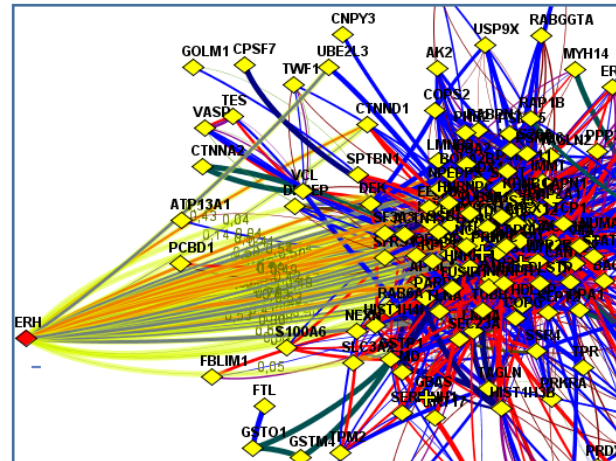
$N \text{ links}_{expected} = 1.00$

Standard deviation = 1.25

$Z = 3.97$

$P\text{-value} = 0.0000356$

Detection of driver mutations



Question:

Could copy number alteration in *EHR* in *HOU501106* lead to changes of its transcriptome and proteome?

$N \text{ links}_{real} = 55$

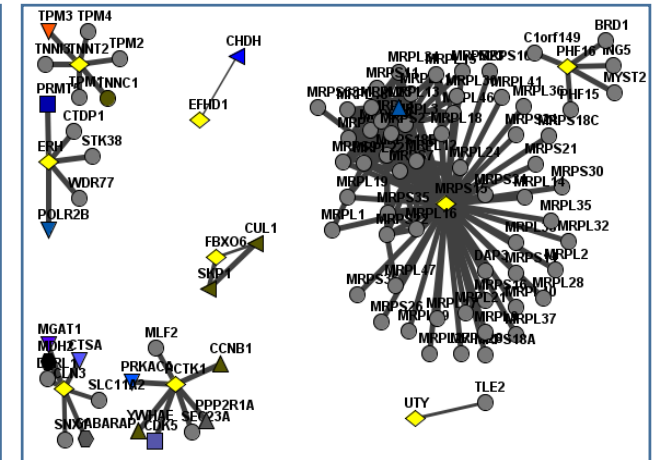
$N \text{ links}_{expected} = 37.05$

Standard deviation = 3.59

$Z = 3.59$

$P\text{-value} = 0.00016$

Coherence of genome alterations



Question:

Are CNA in *HOU501106* coherent?

$N \text{ links}_{real} = 0$

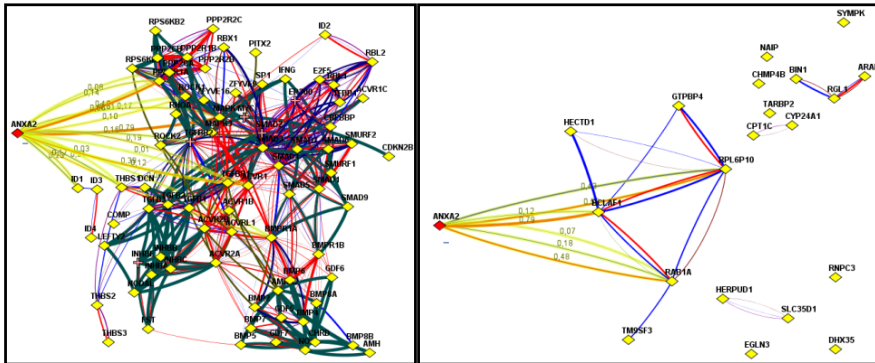
$N \text{ links}_{expected} = 1.05$

Standard deviation = 0.80

$Z = -1.31$

$P\text{-value} = 0.905$

Network enrichment analysis: what to use?



- **R package NEA:** Alexeyenko A, Lee W, ... Pawitan P (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*
- **Perl software :** Simon Merid et al., to be published.
- **C++, the “crosstalk” tool :** Ted McCormack et al., to be published
- Last but not least: <http://FunCoup.sbc.su.se>

