Systems Biology for Starters: Reconstruction of Gene Regulatory Networks Using MetaCore

Alexey Eroshkin
Bioinformatics Core

TODAY

- Introduction
- Data input, options
- Enrichment analysis
- Building networks from scratch, results output
- "Combing the hairball“ (pruning the network)
- Network validation
- Overlaying genomics, drug assay, and other data
- Tips and tricks
  - Your participation is welcomed

Bioinformatics Core supports all Sanford|Burnham scientists

Andrew Hodges, PhD,
Bioinformatics Scientist

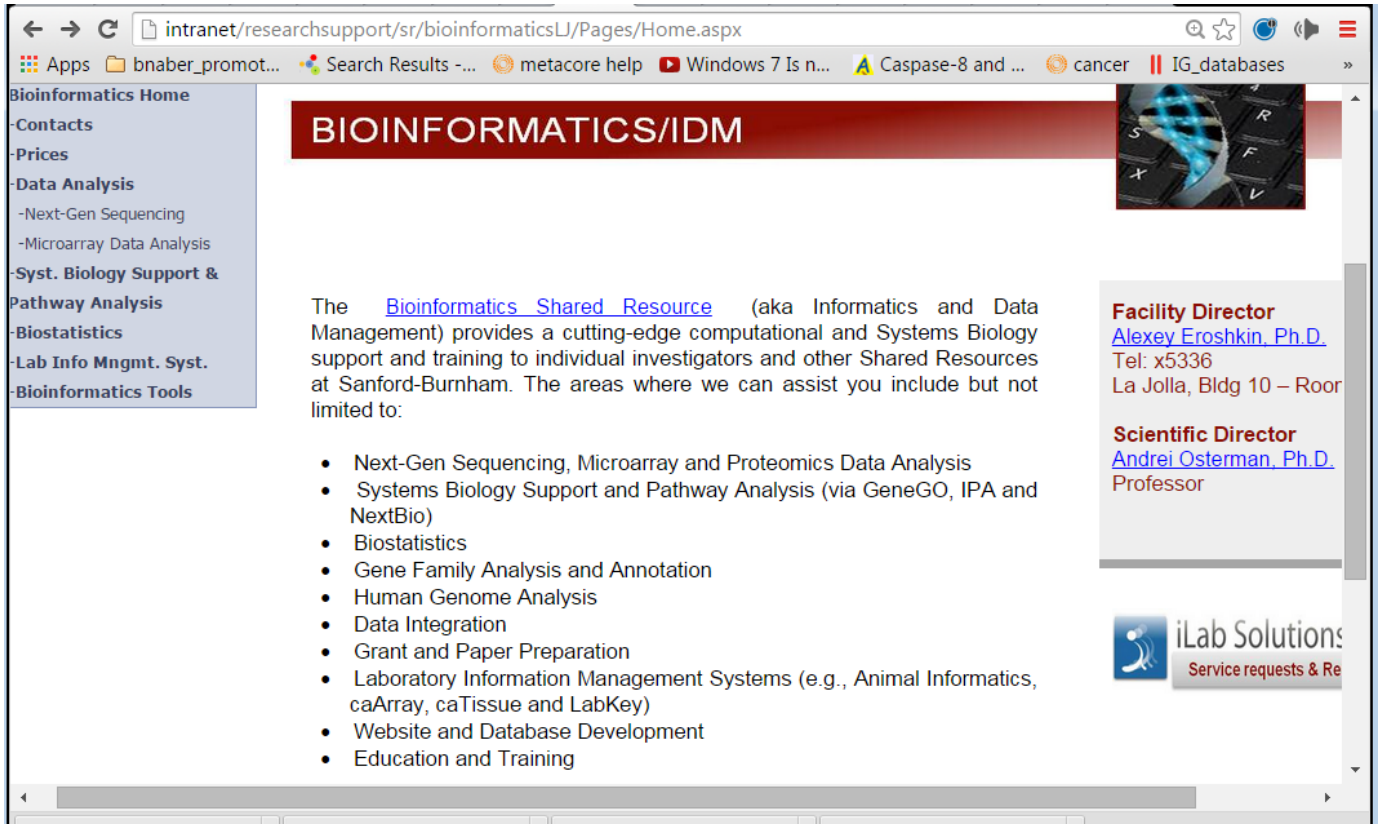
Alexey Eroshkin, PhD,
Bioinformatics Scientist
and Core Director

Xiayu (Stacy) Huang, PhD,
Biostatistician (on contract)

Bioinformatics Core

- http://intranet/researchsupport/sr/bioinformaticsLJ/Pages/Home.aspx
- http://bsrweb.burnham.org

Request our service using iLAB, phone, email
-- Walk-ins are welcome! - Bldg. 10 Rm. 2405/6



Data flood

- What can save you?
- Systems biology approaches i.e., network analysis

Networks (wiki)

A **network** is any system with sub-units that are linked into a whole

Complex biological systems may be represented and analyzed as computable networks

Nodes and edges are the basic components of a network. Nodes represent units in the network, while edges represent the interactions between the units.

Networks in biology

- **Protein-protein interaction networks**
- **Gene regulatory networks (DNA-protein interaction networks)**
- **Gene co-expression networks (transcript-transcript association networks)**
- **Metabolic networks**
- **Signaling networks**

Pathway Databases

- MetaCore (Thomson Reuters)
- Ingenuity Pathway Analysis (www.ingenuity.com)
- NetworkAnalyst (http://www.networkanalyst.ca/)
- Pathway Studio (www. ariadnegenomics.com)
- GenMAPP (www. genmapp.com)
- WikiPathways (www. wikipathways.org)
- cPath (cbio.mskcc.org/cpath)
- BioCyc (www.biocyc.org)
- Pubgene (www.pubgene.org)
- PANTHER (www. pantherdb.org)
- WebGestalt (bioinfo.vanderbilt.edu/webgestalt/)
- ToppGene Suite(/toppgene.cchmc.org/)
- DAVID (david.abcc.ncifcrf.gov/)
- Pathway Painter (pathway.painter.gsa-online.de/

Why use MetaCore? Comparison between different pathway databases and experimentally derived gold-standards for several transcription factors

MetaCore performs much better that other databases

Underlined values in red represent statistically significant intersections

Shmelkov E, et al., Biol Direct. 2011 Feb 28;6:15.

Number of overlapping genes between a gold-standard and a pathway-derived gene set.

Transcription factor	Gold standard ID#	Number of genes (in gold stand.	Ingenuity (Transcription)	Ingenuity (All)	TransPath	TransFac	Biocarta	KEGG	Wikipathways	Cell Signaling Technology	GeneSpring (Expression or Bind.	GeneSpring (Expression and Bin	Pathway Studio	MetaCore
AR	I	513	1	5	5	1	0	1	0	0	1	0	4	39
	II	712	2	6	6	1	0	1	0	0	3	0	5	55
	III	526	1	4	4	1	0	1	0	0	1	0	4	36
	Number of genes (in pathway)		17	83	31	8	0	1	1	0	11	1	51	441
BCL6	I	369	2	5	0	0	0	0	0	0	0	0	3	14
	II	98	1	1	0	0	0	0	0	0	0	0	1	5
	III	271	0	4	0	0	0	0	0	0	0	0	2	14
	IV	76	0	0	0	0	0	0	0	0	0	0	0	3
	Number of genes (in pathway)		8	53	5	2	0	0	0	0	17	4	39	132
MYC	I	1887	15	85	5	3	2	0	1	0	4	1	28	333
	II	1708	19	80	5	2	1	1	0	1	2	1	26	278
	III	3039	30	151	13	6	0	3	4	4	7	4	65	624
	IV	2224	20	112	9	5	0	2	3	1	8	5	39	476
	Number of genes (in pathway)		89	493	29	13	4	14	7	6	26	12	197	1649
NOTCH1	I	414	2	4	1	0	1	1	1	2	1	1	8	16
	II	302	2	4	1	0	1	1	1	2	1	1	7	17
	III	637	2	4	1	1	1	1	1	2	2	1	11	21
	IV	471	2	2	1	0	1	1	1	2	1	1	5	19
	V	167	1	1	1	0	0	0	0	1	0	0	1	4
	VI	166	1	1	1	0	0	0	0	1	0	0	1	4
	Number of genes (in pathway)		8	84	5	1	4	3	3	9	15	4	210	133
RELA	I	1864	2	31	6	5	9	4	1	3	5	0	21	53
	II	1420	0	16	4	4	1	1	0	1	2	0	10	28
	III	188	1	11	6	3	7	4	1	1	3	0	6	15
	IV	136	1	13	7	3	4	3	1	1	3	0	8	12
	Number of genes (in pathway)		18	196	69	42	141	63	13	7	18	0	134	389
STAT1	I	2128	1	5	5	0	1	2	0	1	2	1	8	15
	II	2967	1	6	6	2	0	3	2	0	5	1	15	33
	Number of genes (in pathway)		5	95	67	35	3	33	22	2	19	4	152	285

MetaCore vs IPA: each has own strong points

MetaCore:

- Rich and detailed database content
- 10 algorithms for network reconstruction
- User has complete control of network building
- Easy output of all network-related data
- Metabolomics data analysis
- Detailed interaction annotation
- Multiple interaction filters!
  - Mechanisms (20 total)
  - Effects:
    - Activation
    - Inactivation
    - Unspecified

IPA

- More intuitive interface
- Easy to learn
- Gene Isomer view
- Additional tools
  - Upstream Regulator and Downstream Effects analysis

Access to MetaCore

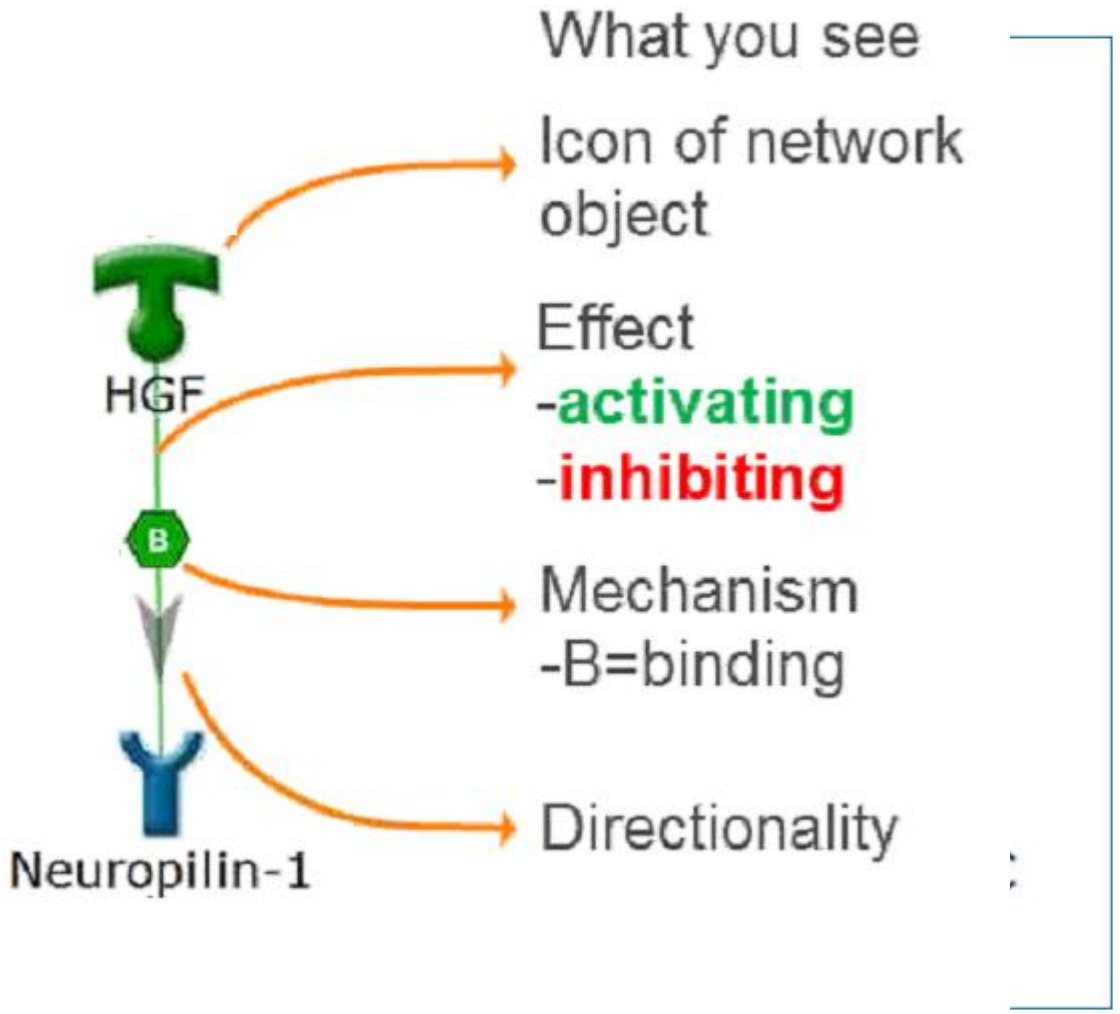
1. Email your request for an account to: genegosupport@thomsonreuters.com
2. You will receive your access keys
3. Go to MetaCore portal (http://portal.genego.com) and type your User ID & Password
4. You are in!

Do not try to remember – I will send you this info

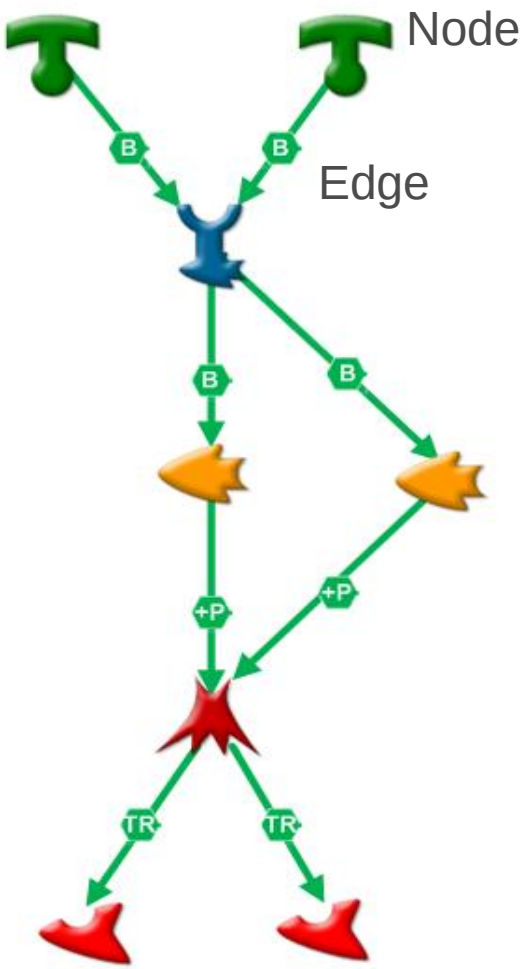
WHAT TYPES OF QUESTIONS CAN BE ANSWERED WITH METACORE?

- What are the ***most relevant biological pathways*** for my data?
- What is known about any ***particular gene/ protein/ compound***? Which ***canonical pathways*** is it involved in? Which ***diseases*** are associated with it? Where is it expressed?
- What are the ***known interactions*** downstream of my favorite gene/protein/microRNA? How does my data reflect this?
- What are the ***differences or similarities*** between multiple experimental conditions or species/cell lines? Or between different data types?
- What are the most ***important genes*** in my gene list?
  - What are they ***interacting*** with?
  - Which genes are involved in my ***disease*** of interest?
  - What are some ***important hubs*** responsible for signal regulation in my data?
  - Are there known ***therapeutic targets*** in my list? How are they connected to each other and to my significant genes?

What is interaction?



SIGNALING PATHWAYS IN NETWORK

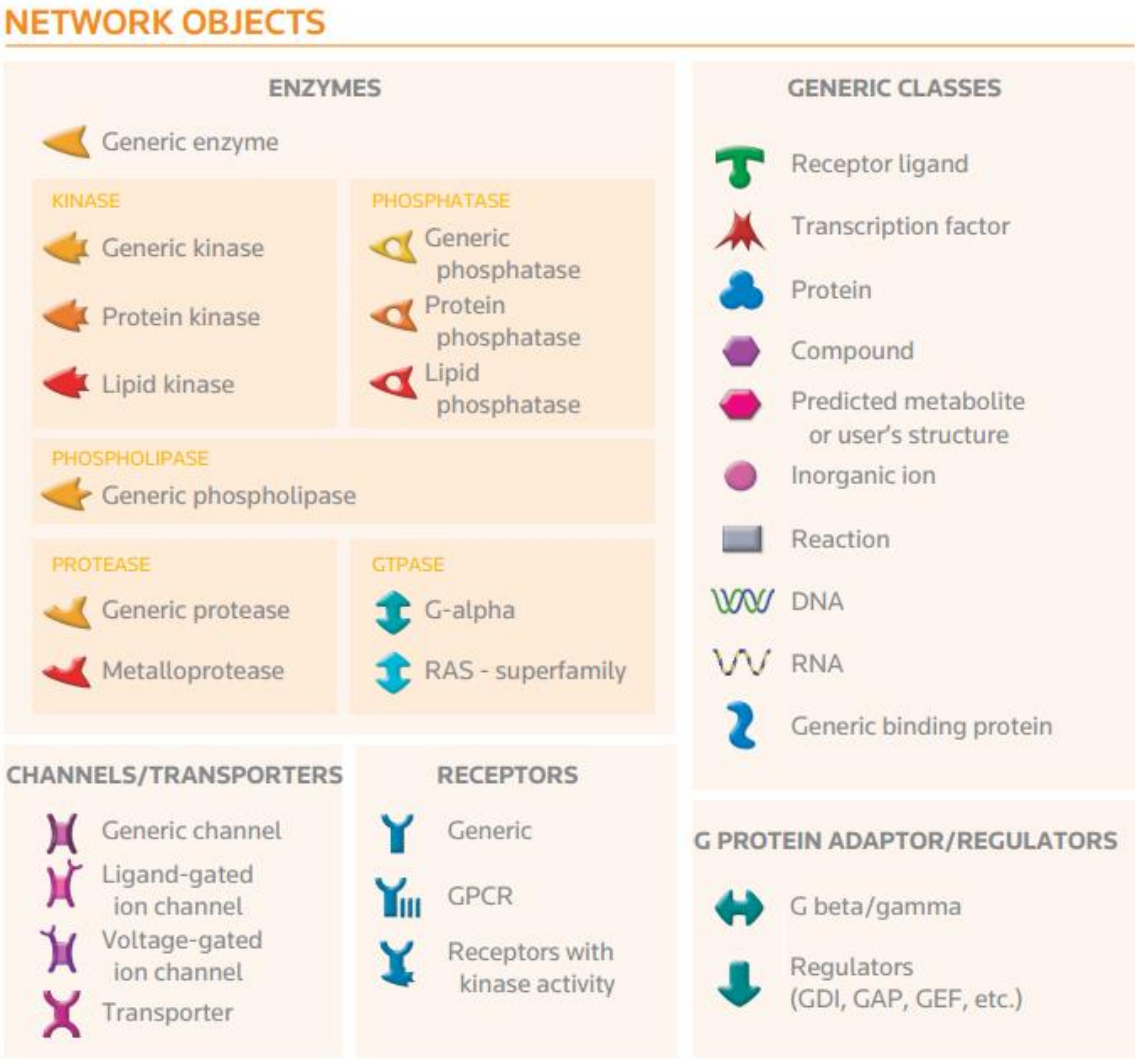


- Edges reflect interaction mechanisms
  - 20 such mechanisms can be defined
  - Consequences of interaction
    - activation
    - inhibition
    - unspecified (you normally should remove this type)
- Nodes reflect the molecular function of corresponding biomolecule (receptor, phosphatase etc.)

.

Network objects in MetaCore

do not try to memorize – help file is ready available :=)



Why do we need network analysis/database?

- Learning tool
  - Just click on any network element
- Data analysis tool
- Hypothesis generation/validation tool

KNOWLEDGE BASE BEHIND METACORE: HOW IT WAS CREATED

Data level

~3,000 journals

45 PhD's and 5 M.D.'s

Manually curated content

Annotation level

High quality

"small experiment evidence"

3 months training

3 levels of supervision

Data storage level

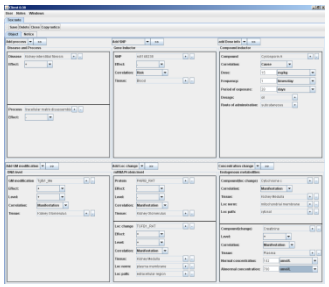
Oracle tables

Ontologies

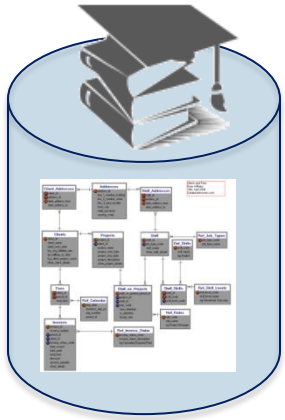
Controlled vocabulary

Knowledge level/VALUE

GUI access to MetaBase content + analysis + reporting solutions

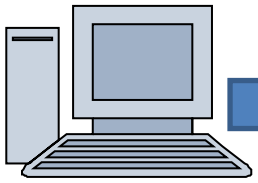


MetaBase



GeneGo Applications

MetaCore™

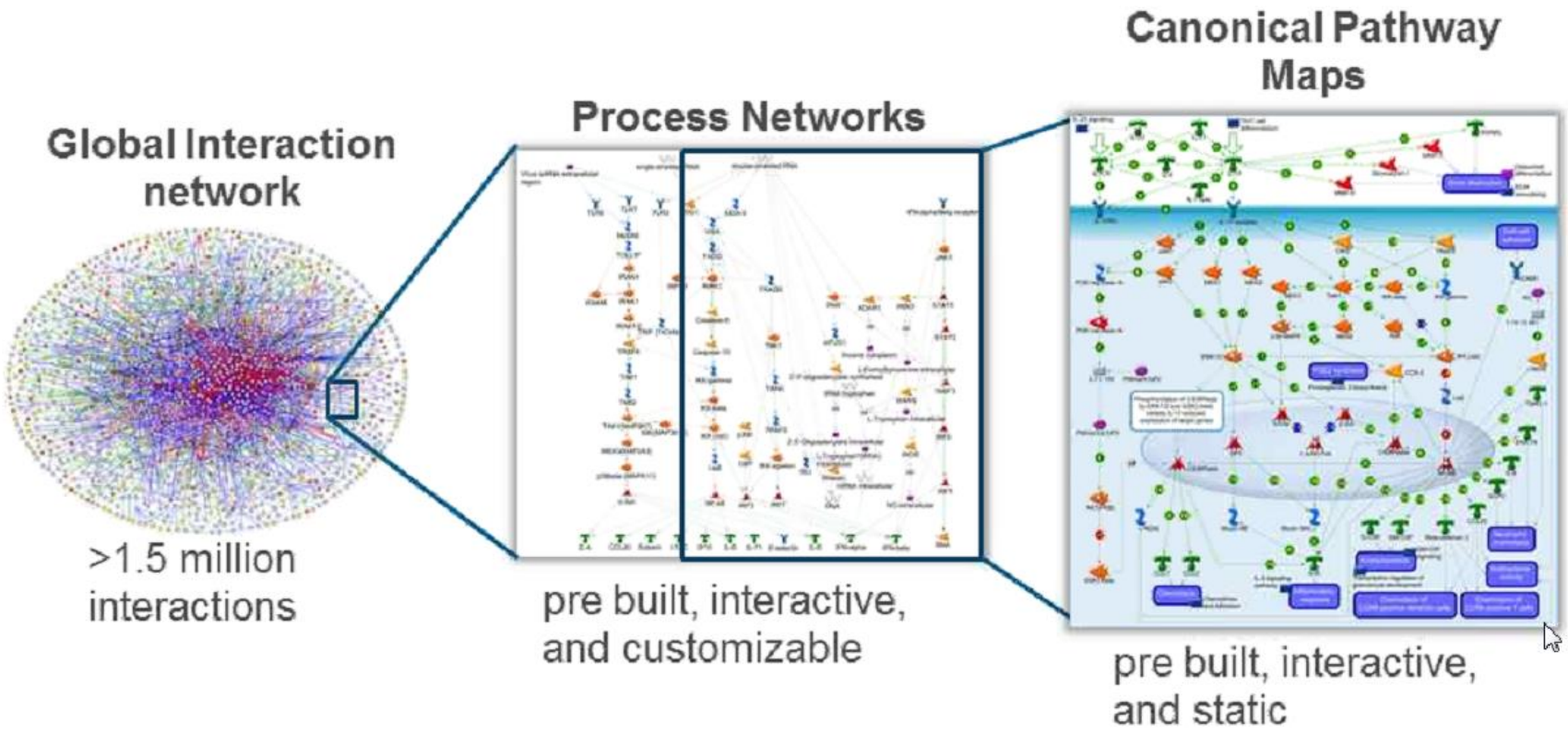


KNOWLEDGE!

and MetaSearch

(Advanced search)

Global interactions, networks and pathway maps

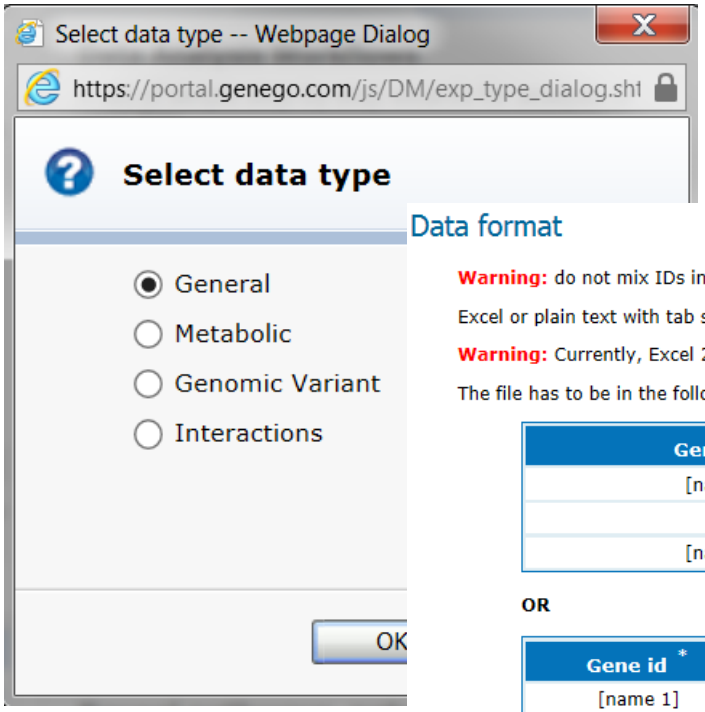


THOMSON REUTERS

General steps in analysis of your data

- Upload your list of genes/proteins/metabolites
- (or) Activate data loaded before
- Run enrichment analysis
- Build network
- Validate network with independent data

Data input is simple



Data format

Warning: do not mix IDs in the same column.

Excel or plain text with tab separated fields formats are supported.

Warning: Currently, Excel 2007 files are not supported. To upload your file, please save it as a text file with tab separated fields or an older Excel version

The file has to be in the following format (Column order is not important):

Gene id *	Exp 1	Exp 2	...
[name 1]	[value 1.1]	[value 2.1]	...
...	...	...	...
[name n]	[value 1.n]	[value 2.n]	...

OR

Gene id *	Exp 1	P-value 1	Exp 2	P-value 2	...
[name 1]	[value 1.1]	[P-value 1.1]	[value 2.1]	[P-value 2.1]	...
...	...	...	...	...	...
[name n]	[value 1.n]	[P-value 1.n]	[value 2.n]	[P-value 2.n]	...

Required fields marked with (*)

Most files are of «General» type. The following identifiers are recognized:

- EntrezGene (LocusLink) IDs — Mouse, Rat, Bovine, Chimpanzee, Dog, Zebra fish, Chicken, Fly, Mosquito, Worm, Arabidopsis, Rice, Blast of rice, Plasmodium, Mold, Bread mold, Candida sphaerica, Fission yeast and Baker's yeast IDs are supported as well (via orthologs)
- Gene symbol (e.g. TP53, etc.)
- Affymetrix tag ID (expression)
- Affymetrix tag ID (exon)
- Affymetrix tag IDs (SNP)
- Illumina tag IDs (expression)
- Agilent tag IDs (expression)
- Codelink tag IDs (expression)
- OMIM IDs
- RefSeq IDs

Data types: general

Gene list

Gene Symbol
HLA-B
HLA-B
SH3BGR
PSMA3
HS3ST1
RIMKLB
EIF4EBP1
NT5E
KIAA0895
FABP4
TNFAIP3
RBFOX1
A2M
AAK1
GOT1
GOT2
ABCA1
ABCA5
ABCB8
ABCB9
ABCC1
ABCC5
ABHD4
ABI2
TNIP2

Gene fold change

Gene Symbol	Fold Change
HLA-B	10
HLA-B	10
SH3BGR	10
PSMA3	10
HS3ST1	10
RIMKLB	10
EIF4EBP1	10
NT5E	10
KIAA0895	10
FABP4	10
TNFAIP3	10
RBFOX1	10
A2M	10
AAK1	10
GOT1	10
GOT2	10
ABCA1	10
ABCA5	10
ABCB8	10
ABCB9	10
ABCC1	10
ABCC5	10
ABHD4	10
ABI2	10
TNIP2	10
ABR	10
ABTB1	10

Differentially expressed genes in two conditions Fold change and p-value

Gene Symbol	Astrocytic vs. Normal Brain (Control)	
YWHAB	-2.20452	0.025215
YWHAG	-4.57555	0.005521
YWHAH	-5.7121	0.007812
YWHAZ	-2.72895	0.015473
YWHAB	-2.20452	0.025215
YWHAH	-5.7121	0.007812
YWHAG	-4.57555	0.005521
YWHAZ	-2.72895	0.015473
HLA-B	2.402827	0.02571
HLA-B	2.402827	0.02571
HLA-B	2.402827	0.02571
SH3BGR	2.714856	0.049934
PSMA3	-2.47434	0.03154
RIMKLB	3.883591	0.02671
EIF4EBP1	3.132838	0.041458
NT5E	3.263703	0.026319
KIAA0895	-2.46505	0.019195
FABP4	6.905603	0.007252
TNFAIP3	2.808823	0.028993
RBFOX1	-7.61465	0.032001

Differentially expressed genes between multiple conditions (Fold change and p-value)

Gene Sym	Anaplastic Oligoastrocytoma vs. Normal Brain (Control)		Astrocytic vs. Normal Brain (Control)		Secondary Glioblastoma vs. Normal Brain (Control)	
HLA-B	-2.26337	0.036326	-2.20452	0.025215	3.132838	0.041458
HLA-B	-4.23796	0.027396	-4.57555	0.005521	3.263703	0.026319
SH3BGR	-3.86437	0.027972	-5.7121	0.007812	-2.46505	0.019195
PSMA3	-2.19096	0.047266	-2.72895	0.015473	6.905603	0.007252
HS3ST1	-2.26337	0.036326	-2.20452	0.025215	3.132838	0.041458
RIMKLB	-3.86437	0.027972	-5.7121	0.007812	3.263703	0.026319
EIF4EBP1	-4.23796	0.027396	-4.57555	0.005521	-2.46505	0.019195
NT5E	-2.19096	0.047266	-2.72895	0.015473	6.905603	0.007252
KIAA0895	3.132838	0.041458	2.402827	0.02571	7.477156	0.030005
FABP4	3.263703	0.026319	2.402827	0.02571	7.477156	0.030005
TNFAIP3	-2.46505	0.019195	2.402827	0.02571	7.477156	0.030005
RBFOX1	6.905603	0.007252	2.714856	0.049934	3.132838	0.041458
A2M	-2.53441	0.03091	-2.47434	0.03154	3.263703	0.026319
AAK1	2.874272	0.047935	3.883591	0.02671	-2.46505	0.019195
RIMKLB	-2.39276	0.027396	3.883591	0.02671	6.905603	0.007252
EIF4EBP1	4.600407	0.049804	3.132838	0.041458	3.597915	0.02473
NT5E	3.485927	0.027396	3.263703	0.026319	5.1936	0.045312
KIAA0895	-6.94798	0.026495	-2.46505	0.019195	9.767623	0.013309
FABP4	4.600407	0.049804	6.905603	0.007252	5.1936	0.045312
TNFAIP3	3.485927	0.027396	2.808823	0.028993	9.767623	0.013309

Metabolic data

- Chemical Name
- Formula
- Molecular Weight
- SMILES
- InChI
- CAS Number
- KeGG ID
- PubChem Compound ID
- Compound ID

An example of a valid file:

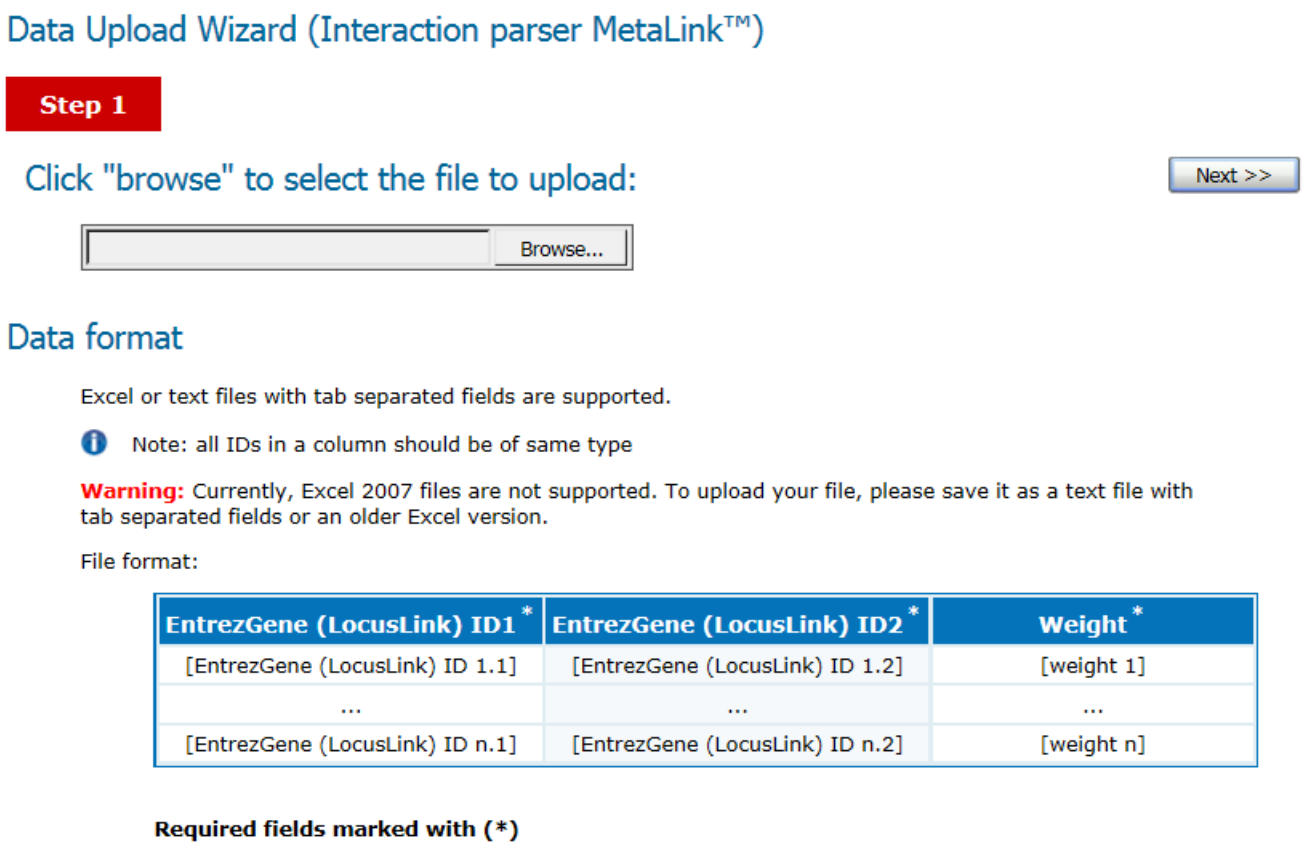
Chemical Name	Formula	Molecular Weight	SMILES	InChI	CAS Number	KeGG ID
Ethanol	C2H6O	46.0	CCO	InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3	64-17-5	C00469
2-hydrazinoethylbenzene	C8H12N2	136.1	NNCCc1ccccc1	InChI=1/C8H12N2/c9-10-7-6-8-4-2-1-3-5-8/h1-5,10H,6-7,9H2	51-71-8	C07430
4-Methyl-1H-pyrazole	C4H6N2	82.1	Cc1c[nH]nc1	InChI=1/C4H6N2/c1-4-2-5-6-3-4/h2-3H,1H3,(H,5,6)	7554-65-6	C07837

PubChem Compound ID	Compound ID	Intensity
702	411	3.857
3675	1077872840	2.275
3406	2052370230	2.451

Genomic data in Variant Call Format (VCF file):

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Interaction data



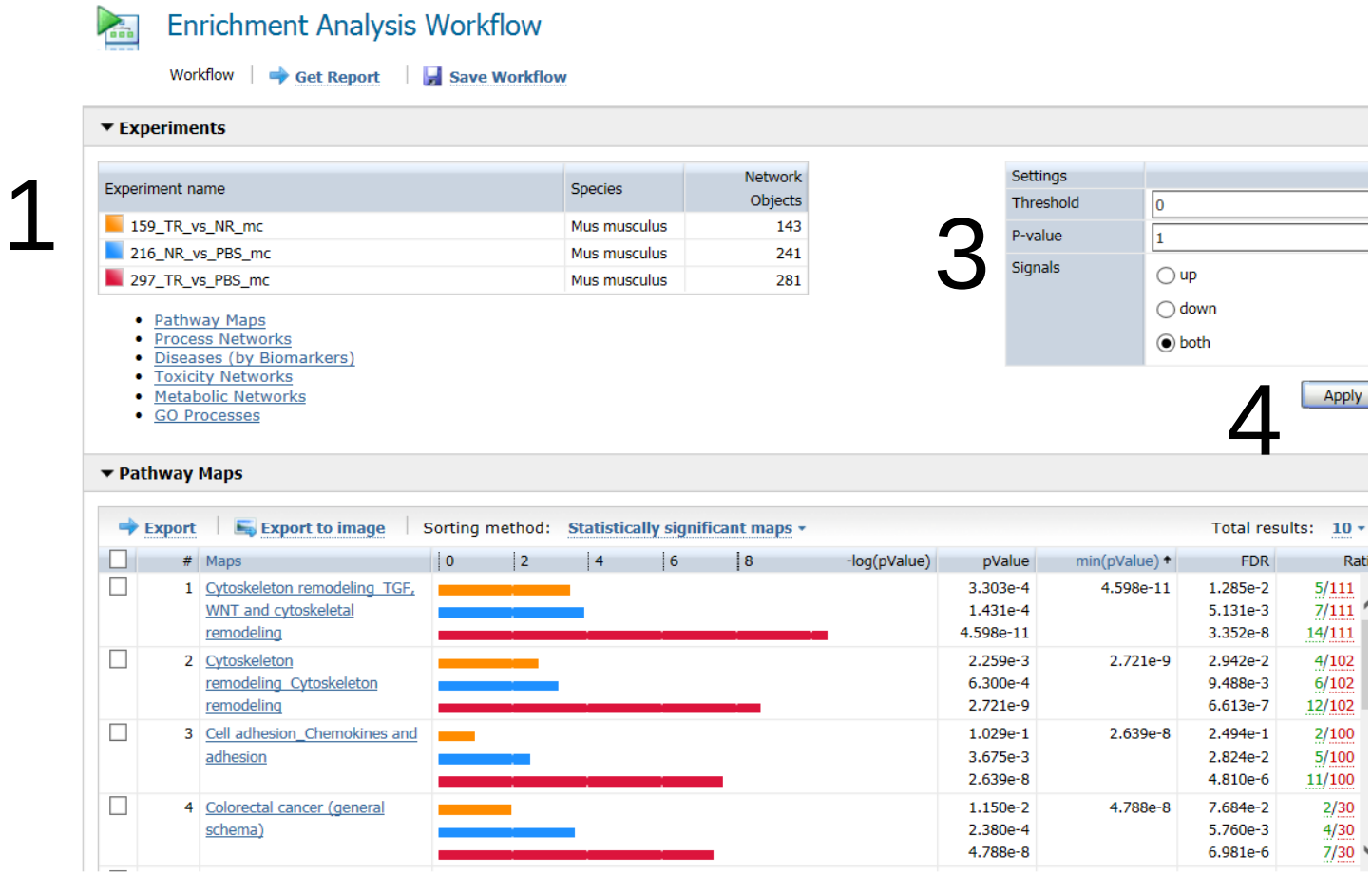
E.g. Co-expression table (weight - correlation)

Enrichment analysis workflow

1. Select Data
2. Select Background (go to Tools)
3. Select Thresholds
4. Run

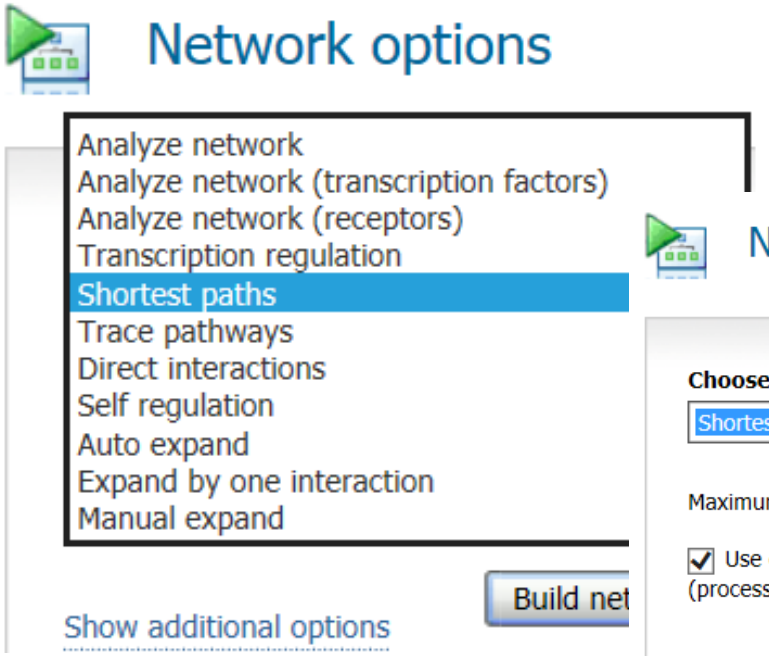
Enrichment categories:

- Pathway Maps
- Process Networks
- Diseases (by Biomarkers)
- Toxicity Networks
- Metabolic Networks
- GO Processes

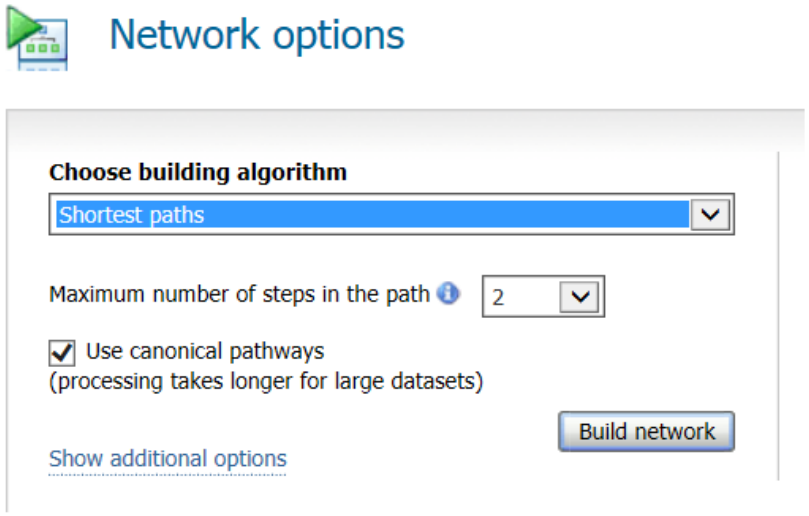


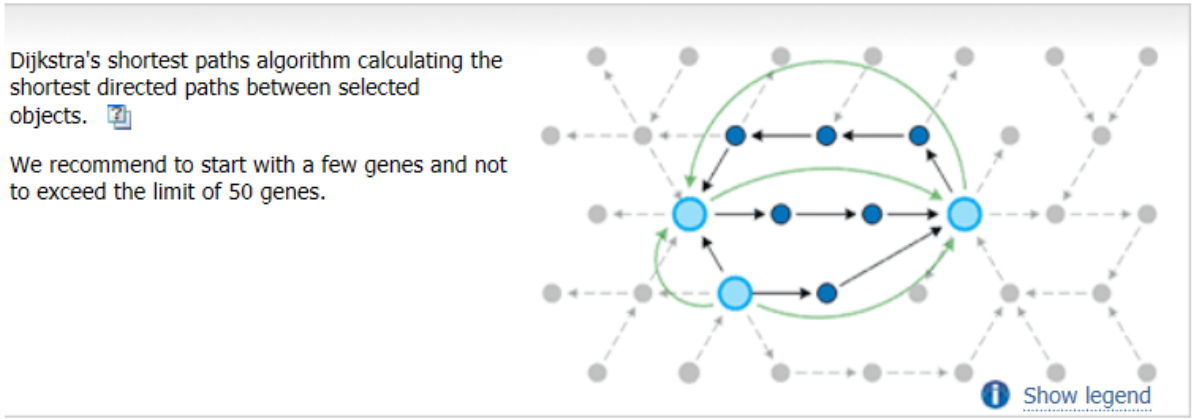
Result
- can be saved and shared as a Word document

Eleven algorithms to build networks

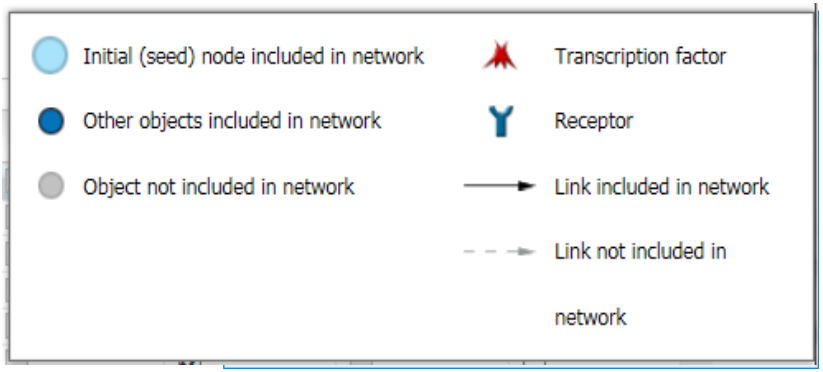


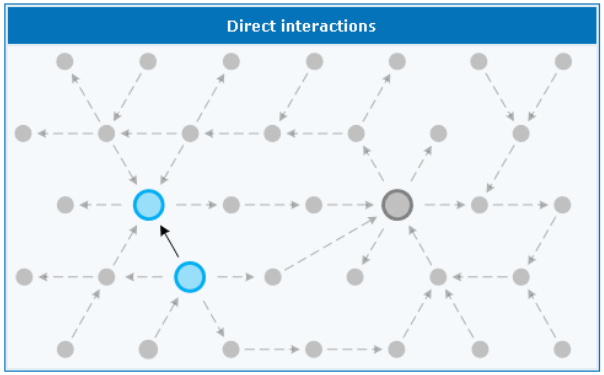
Start with shortest paths





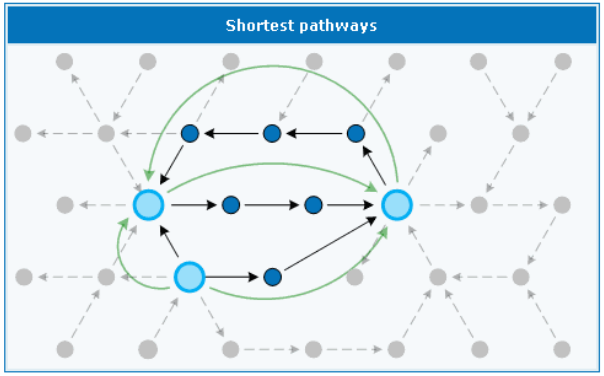
Schematics for the algorithms in MetaCore

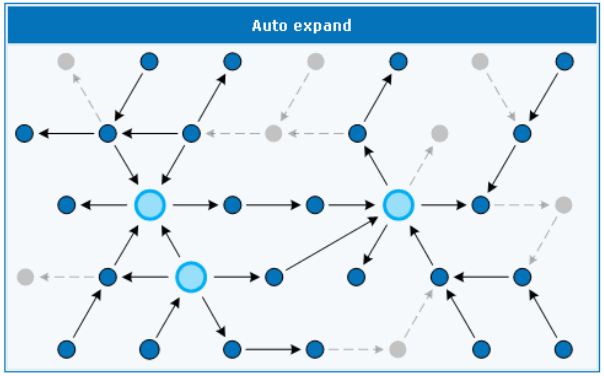




Direct Interactions

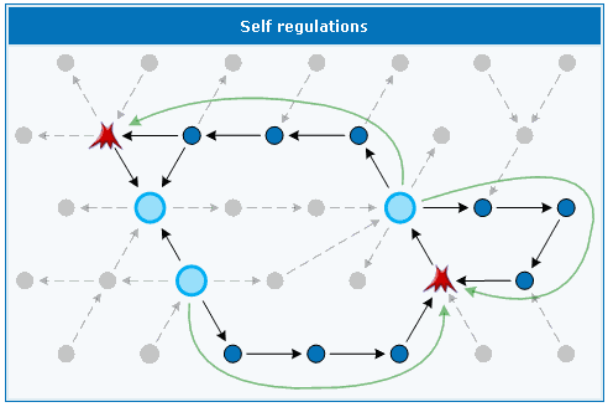
Shortest Pathway

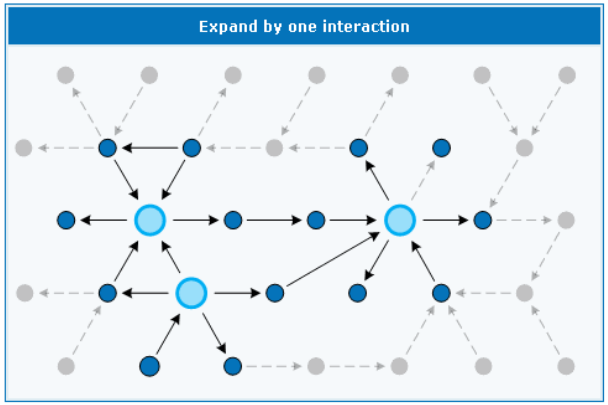




Auto Expand

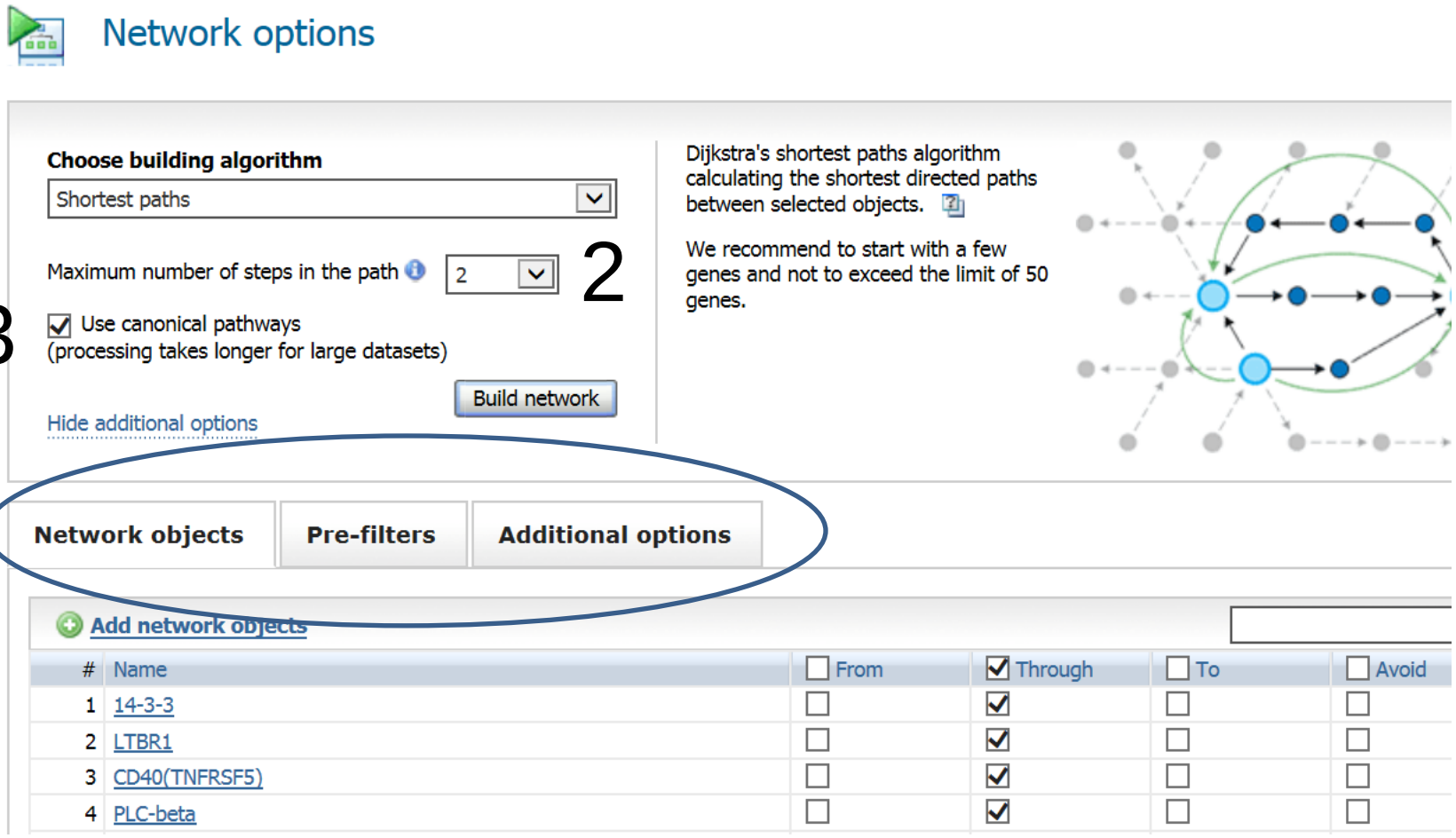
Self Regulation



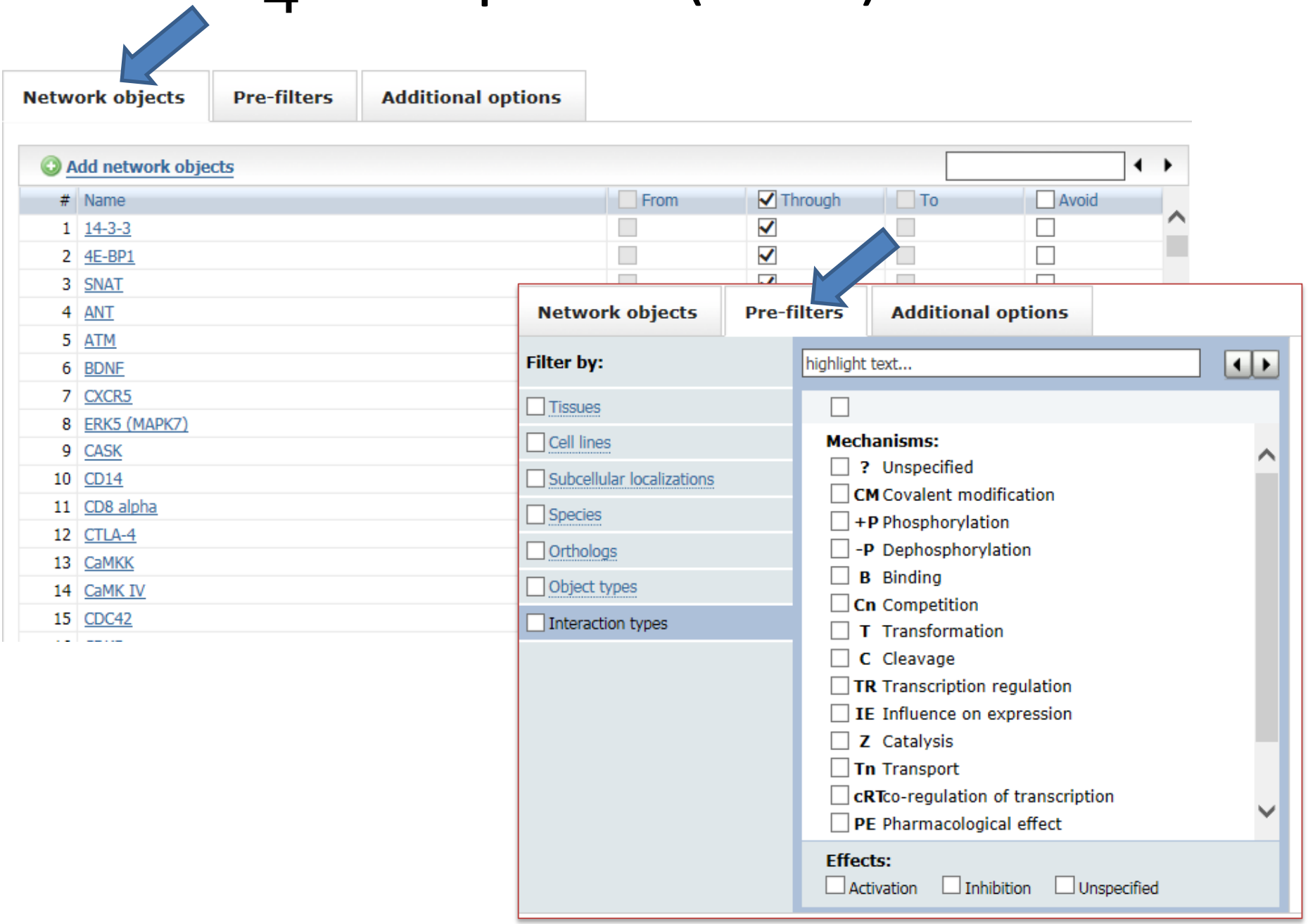


Expand by One Interaction

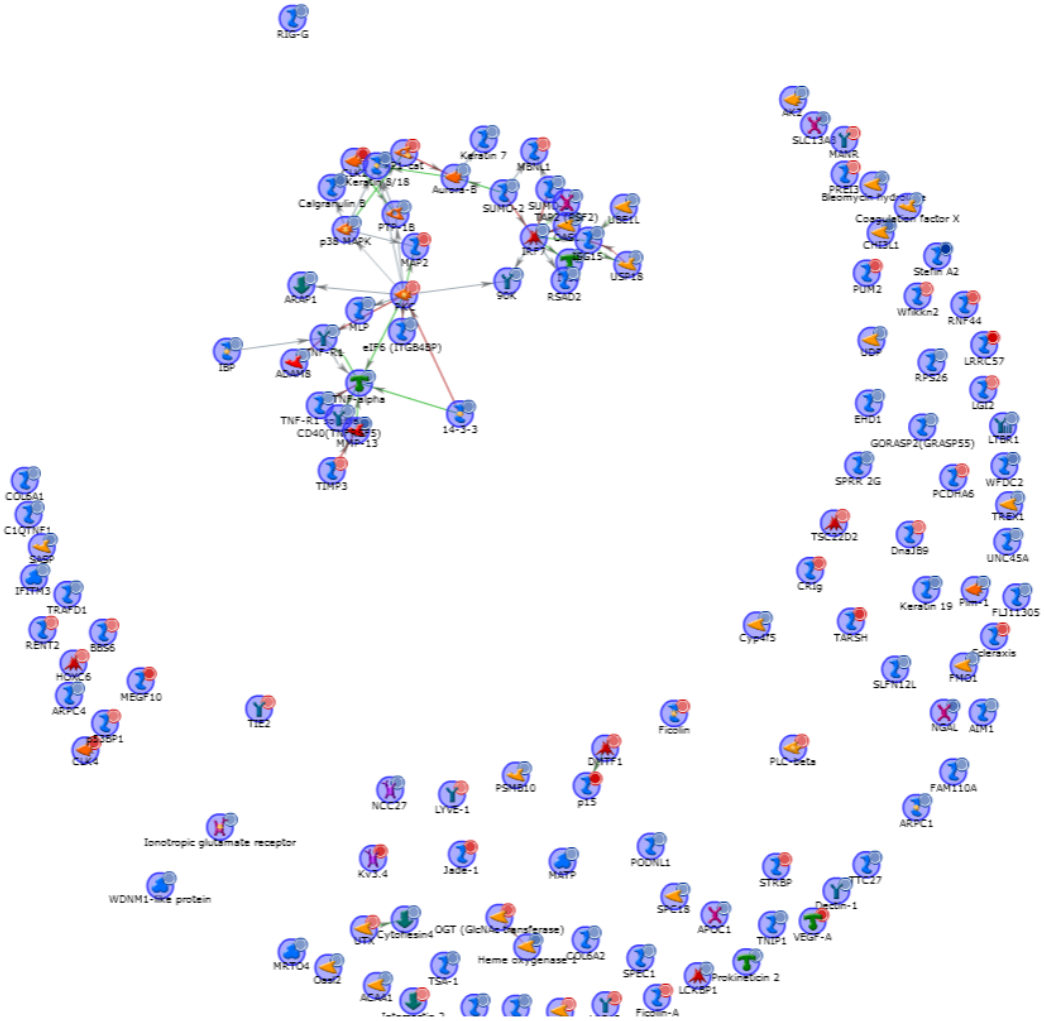
Options in network building



4 Options (cont.)

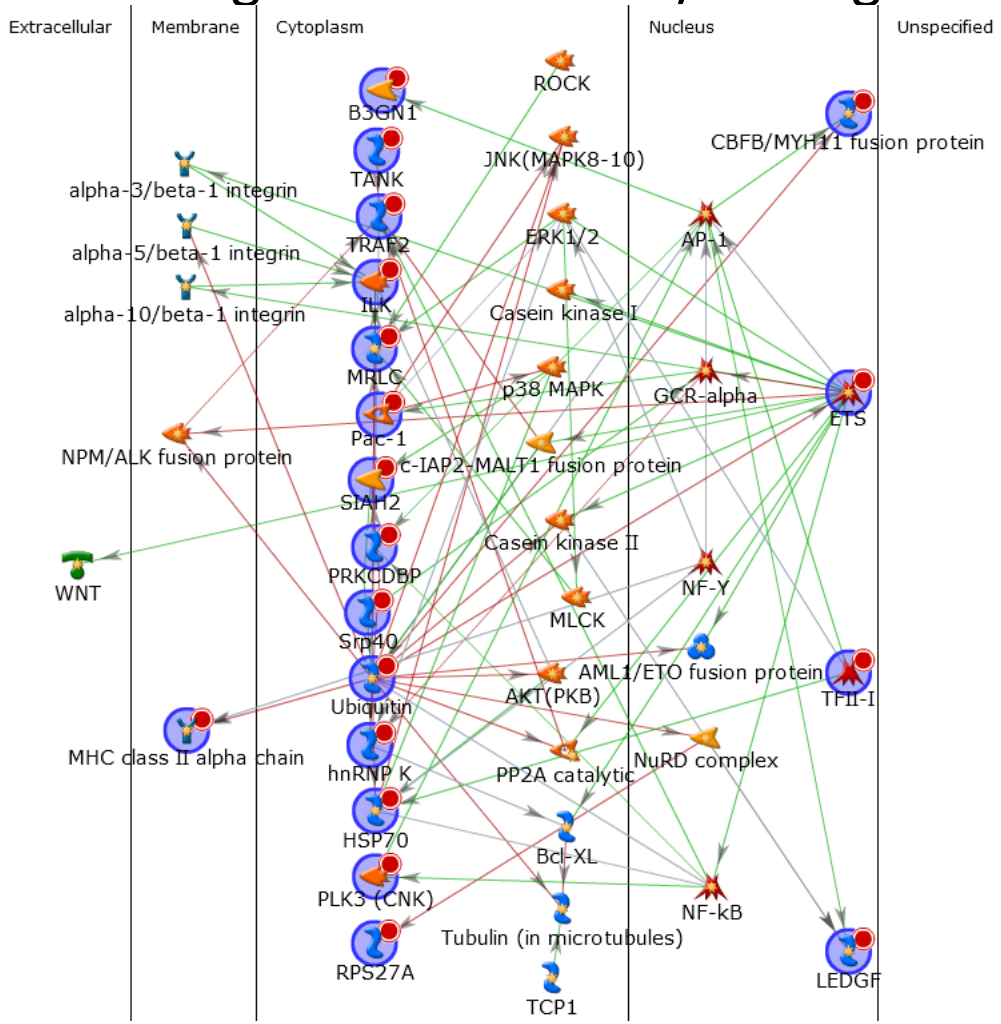


Output: network



Output: network (after some work is done)

- Save picture for your paper
- Save network for sharing and further use/editing



Output - Network statistics (lots of data)

Network statistics:

MetaCore™ version	6.21 build 66768
Name	Tariq_HIV-realted_sp1_Ubiquitin, TRAF2, hnRNP K, ETS2, HSP70
Description	
Total Nodes	
Total Edges	1004
Low Trust Edges	0

- Options
  - Additional options
- Experiments
  - General parser
    - (1) HIV_specific_genes_TR
- Interactions
- Hubs
- Divergence hubs
- Convergence hubs
- Nodes
- Transcription factors
- Membrane receptors
- Secreted proteins & peptides
- GO processes
- Diseases
- Tissues

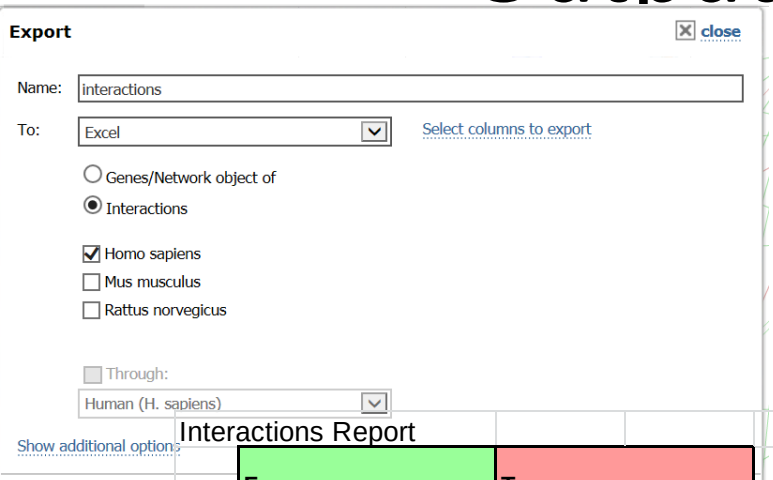
Hubs

#		Visible	Name	Edges
1		☑	Ubiquitin	165/0
2		☑	ETS1	60/0
3		☑	TRAF2	41/0
4		☑	hnRNP K	38/0
5		☑	ETS2	34/0
6		☑	c-Myc	32/0
7		☑	HSP70	28/0

Transcription factors

#		Visible	Name	Edges	Edges In	Edges Out
1		☑	AML1 (RUNX1)	4/0	3/0	1/0
2		☑	Androgen receptor	9/0	5/0	4/0
3		☑	AP-1	1/0	0/0	1/0
4		☑	AP-2A	4/0	2/0	2/0
5		☑	ATF-2	4/0	2/0	2/0
6		☑	ATF-2/c-Jun	0/0	0/0	0/0
7		☑	ATF-6 alpha (50kDa)	3/0	1/0	2/0
8		☑	ATF-6 alpha (90kDa)	2/0	1/0	1/0
9		☑	Bcl-3	2/0	1/0	1/0
10		☑	BOB1	2/0	1/0	1/0
11		☑	c-Fos	6/0	3/0	3/0
12		☑	c-Jun	5/0	3/0	2/0
13		☑	c-Jun/c-Fos	0/0	0/0	0/0
14		☑	c-Jun/c-Jun	0/0	0/0	0/0

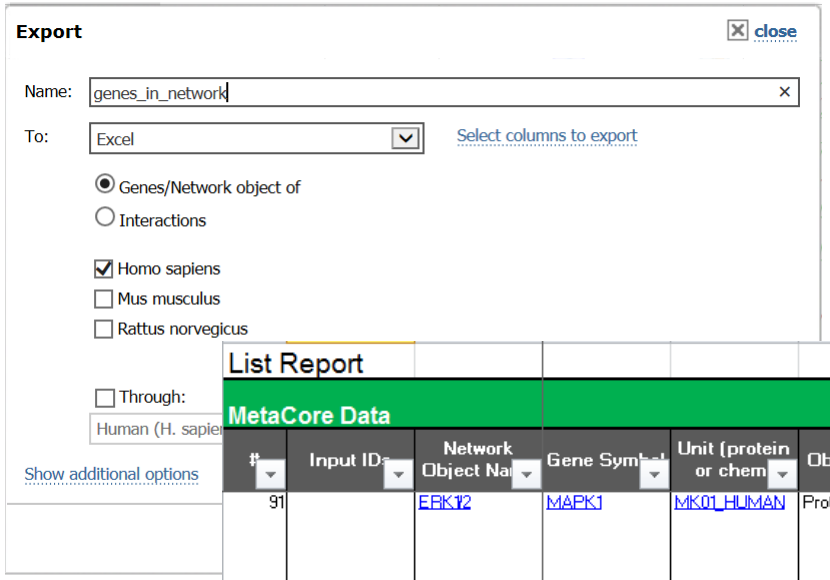
Output: interaction report



Interactions Report

	From		To							From	GSE2223_genes_for_class_Astrocyt		To	GSE2223_genes_for_class_Astrocyt	
#	Network Object "FROM"	Object Type	Network Object "TO"	Object Type	Effect	Mechanism	Homo sapiens	Link Info	References	Input IDs	Signal	p-value	Input IDs	Signal	p-value
1	GRP75	Generic binding protein	CREB1	Transcription factor	Activation	Transport	x	GRP75 transports	16207717						
2	Ubiquitin	Generic binding protein	Caveolin-1	Generic binding protein	Inhibition	Binding	x	Cav-1 S-nitrosylation by	19706615;20923773;21041450;21148404;2182						
3	Ubiquitin	Generic binding protein	ZFP36(Tristetraprolin)	Generic binding protein	Inhibition	Binding	x	MEKK1 mediated	21921033						
4	GPS2	Regulators (GDI, GAP,	TRAF2	Generic binding protein	Inhibition	Binding	x	GPS2 (1-155) interacted	22424771						
5	Ubiquitin	Generic binding protein	Miz-1	Transcription factor	Inhibition	Binding	x	Mule ubiquitinates	20624960;22184250;23699408						
6	TNF-R1	Generic receptor	TRAF2	Generic binding protein	Activation	Binding	x	TNF-R1 binds to and activates	11032840;15047705;15130948;1559091 6;1598						
7	PP1-cat alpha	Protein phosphatase	IP3R1	Ligand-gated ion channel	Inhibition	Dephosphorylation	x	PP1alpha binds to IP3R1.	12533600;16874461				ITPR1	-11.89841	0.0084806
8	TFII-I	Transcription factor	ERK2 (MAPK1)	Protein kinase	Activation	Binding	x	TFII-I physically interacts with	17052463						
9	IRF4	Transcription factor	RAB6IP1	Generic binding protein	Inhibition	Transcription regulation	x	IRF4 probably binds to	21919915						
10	ETS1	Transcription factor	PDGF-A	Receptor ligand	Activation	Transcription regulation	x	ETS1 binds to gene PDGF-A	15297375	ETS1	4.1188454	0.0112976			
11	UHRF1	Generic enzyme	Ubiquitin	Generic binding protein	Activation	Binding	x	UHRF1 is ubiquitinated as	16195352;17370265;17658611;18781797	UHRF1	16.10008	0.007943			
12	Ubiquitin	Generic binding protein	PP2A catalytic	Protein phosphatase	Inhibition	Binding	x	Both the Mid1 binding domain	21454489						
13	Ephrin-A receptor 1	Receptor with enzyme activity	ILK	Protein kinase	Inhibition	Binding	x	EphA1 interacts with integrin-	19118217						
14	PCAF	Generic enzyme	Ubiquitin	Generic binding protein	Activation	Binding	x	PCAF physically	17293853	KAT2B	3.4167676	0.0151011			
15	HIF1A	Transcription factor	HSPA1A	Generic binding protein	Inhibition	Transcription regulation	x	HIF-1 and HIF-2 suppressed	22322648	HIF1A	2.6611633	0.0186205			
16	Bcl-2	Generic binding protein	Tubulin alpha	Generic binding protein	Inhibition	Binding	x	Bcl-2 physically interacts with	16446153						

Output: gene report



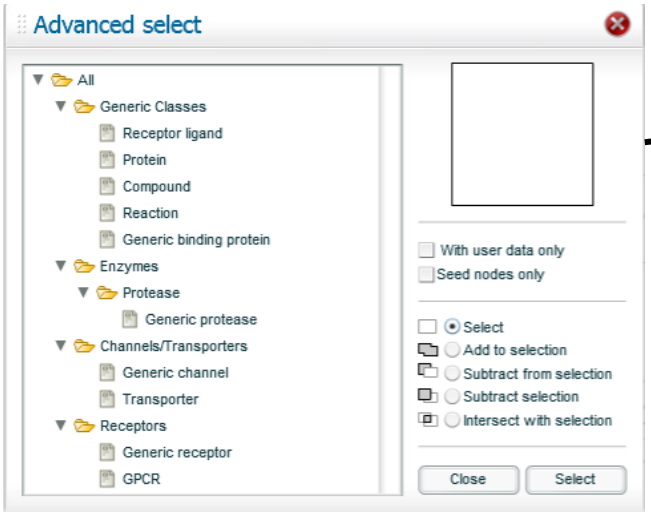
List Report

MetaCore Data								Integrity Biomarkers Data				GSE2223_genes_for_class_Astroc	
#	Input IDs	Network Object Nam	Gene Symbol	Unit (protein or chem	Object Type	Description	Therapeutic Drugs	Integrity Biomarke	Integrity Biomarker Role	Integrity Biomarker Type	Integrity Gene Sym	Signal	p-value
91		ERK1/2	MAPK1	MK01_HUMAN	Protein kinase			Mitogen Activated Protein Kinases	Diagnosis; Differential Diagnosis; Disease Profiling;	Proteomic	MAPK1 variant 1; MAPK10 variant 2; MAPK11; MAPK12;		
92	MAPK3	ERK1/2	MAPK3	MK03_HUMAN	Protein kinase			Mitogen-activated protein kinase 3	Diagnosis; Differential Diagnosis; Disease Profiling;	Proteomic	MAPK3 variant 1	-3.6388392	0.0077539
93	MAPK3	ERK1/2	MAPK3	MK03_HUMAN	Protein kinase			Mitogen Activated Protein Kinases	Diagnosis; Differential Diagnosis; Disease Profiling;	Proteomic	MAPK1 variant 1; MAPK10 variant 2; MAPK11; MAPK12;	-3.6388392	0.0077539
94		ERK2 (MAPK1)	MAPK1	MK01_HUMAN	Protein kinase	Mitogen-activated protein kinase 1		Mitogen-activated protein kinase 1	Diagnosis; Differential Diagnosis; Disease Profiling;	Genomic; Proteomic	MAPK1 variant 1		
95		ERK2 (MAPK1)	MAPK1	MK01_HUMAN	Protein kinase	Mitogen-activated protein kinase 1		Mitogen Activated Protein Kinases	Diagnosis; Differential Diagnosis; Disease Profiling;	Proteomic	MAPK1 variant 1; MAPK10 variant 2; MAPK11; MAPK12;		
96	ETS1	ETS	ETS1	ETS1_HUMAN	Transcription factor			Protein C-ets-1	Diagnosis; Differential Diagnosis; Disease Profiling;	Genomic; Proteomic	ETS1	4.1188454	0.0112976
97	ETS1	ETS	ETS1	ETS1_HUMAN	Transcription factor			45-gene expression kidney transplant rejection panel	Diagnosis	Genomic	ADNP variant 1; ARHGEF7 variant 3; BCL11B variant 1; BRAF; COCH;	4.1188454	0.0112976
98	ETS1	ETS	ETS1	ETS1_HUMAN	Transcription factor			44-gene expression chronic fatigue syndrome panel	Diagnosis	Genomic	AKAP10; ANAPC11 variant 1; APP variant 1; ARL7; ARPC5;	4.1188454	0.0112976

Combing the hairball

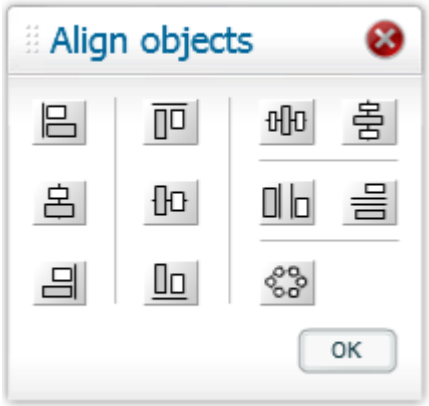
- Remove disconnected nodes
- Remove interaction with irrelevant mechanism and of "unspecified" effect
  - Select nodes with user data and seed nodes
  - Select complement
  - Remove the rest
- Combine related nodes into one new object

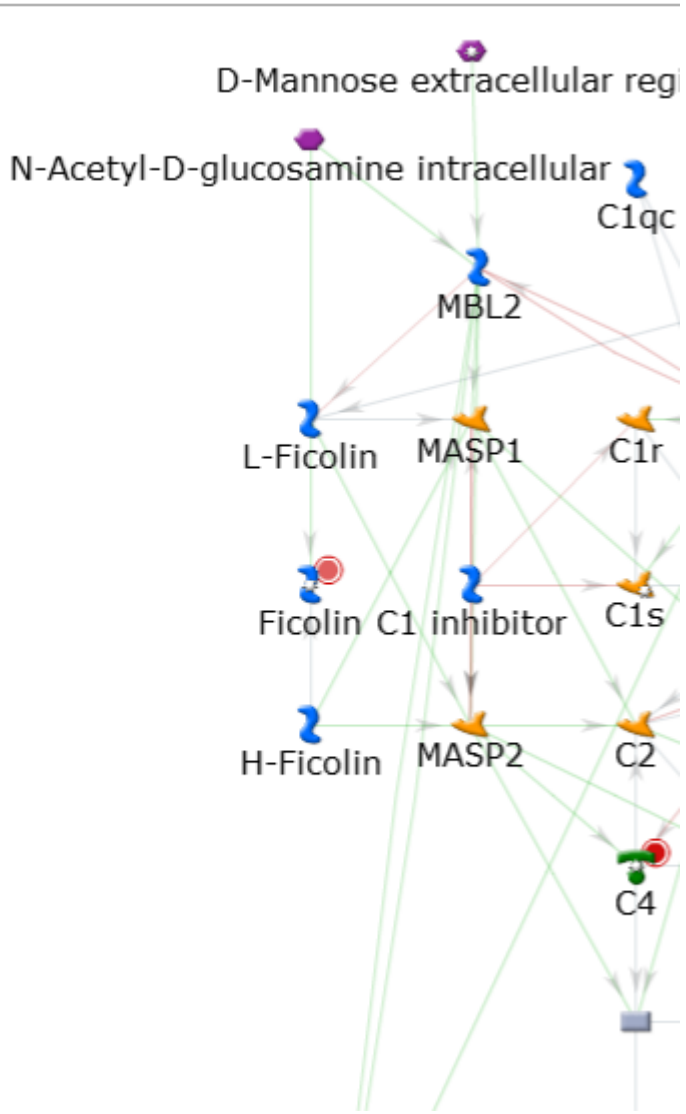
Combing the hairball: multiple tools and filters available



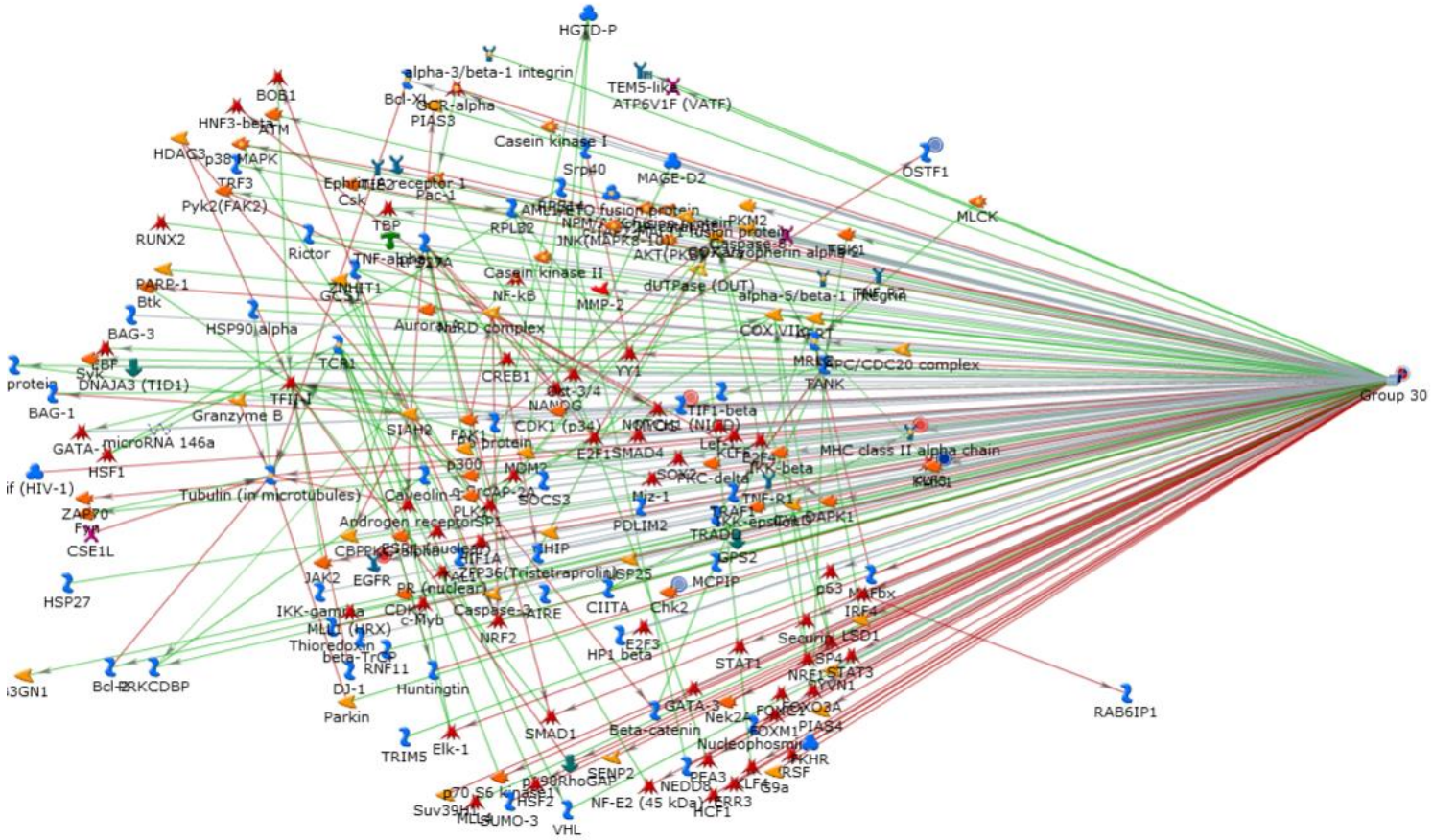
Hide non-connected

Collapse





Any subset of objects can be combined (collapsed) into a new one

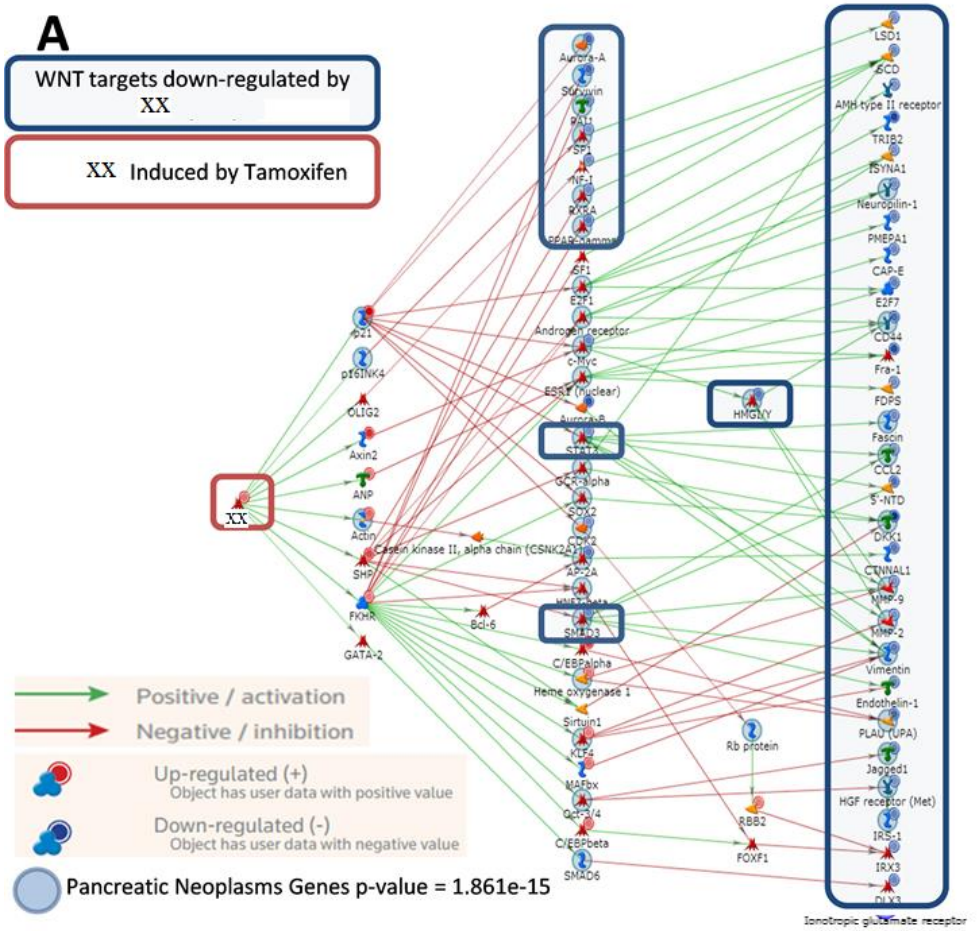


Validation needed:
Developed network is still a hypothesis!

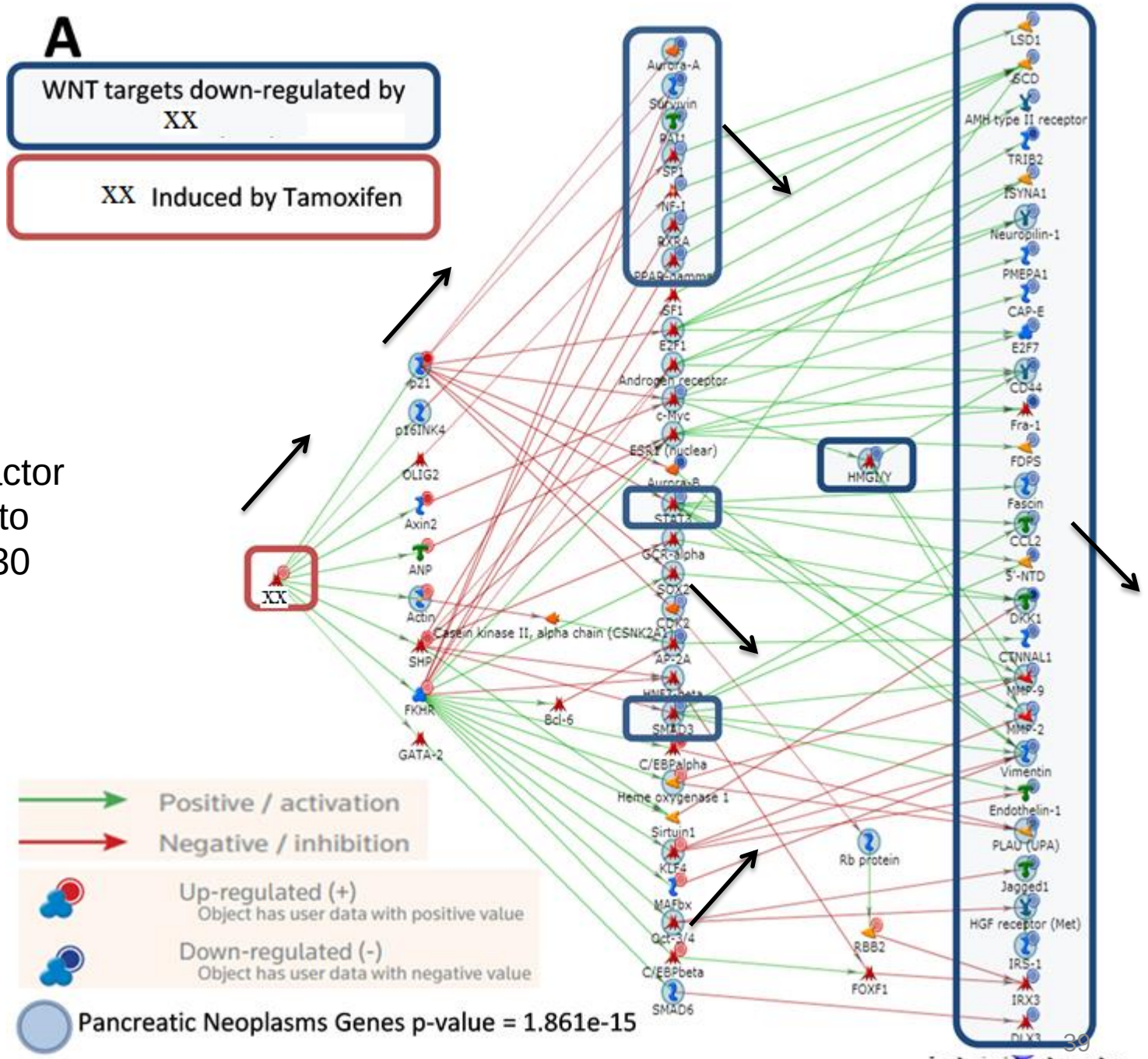
- Use other independent data to validate
- Split your data (samples) in half. Use the first half to develop the network and the other half to validate
- Look for gene overlap with the disease, process or tissue of interest
- Check for concordance with expression data
  - Up or down regulation of genes in agreement with the network signal activation flow
- Experimentally: perturb the network (by drug of gene knockdown) and see the agreement with experiment

Tumor-suppressive mechanisms of transcription factor XX: network concordance with experimental data

By Ally Perlina



Activation of transcription factor (TF) XX leads to inhibition of ~30 other TFs

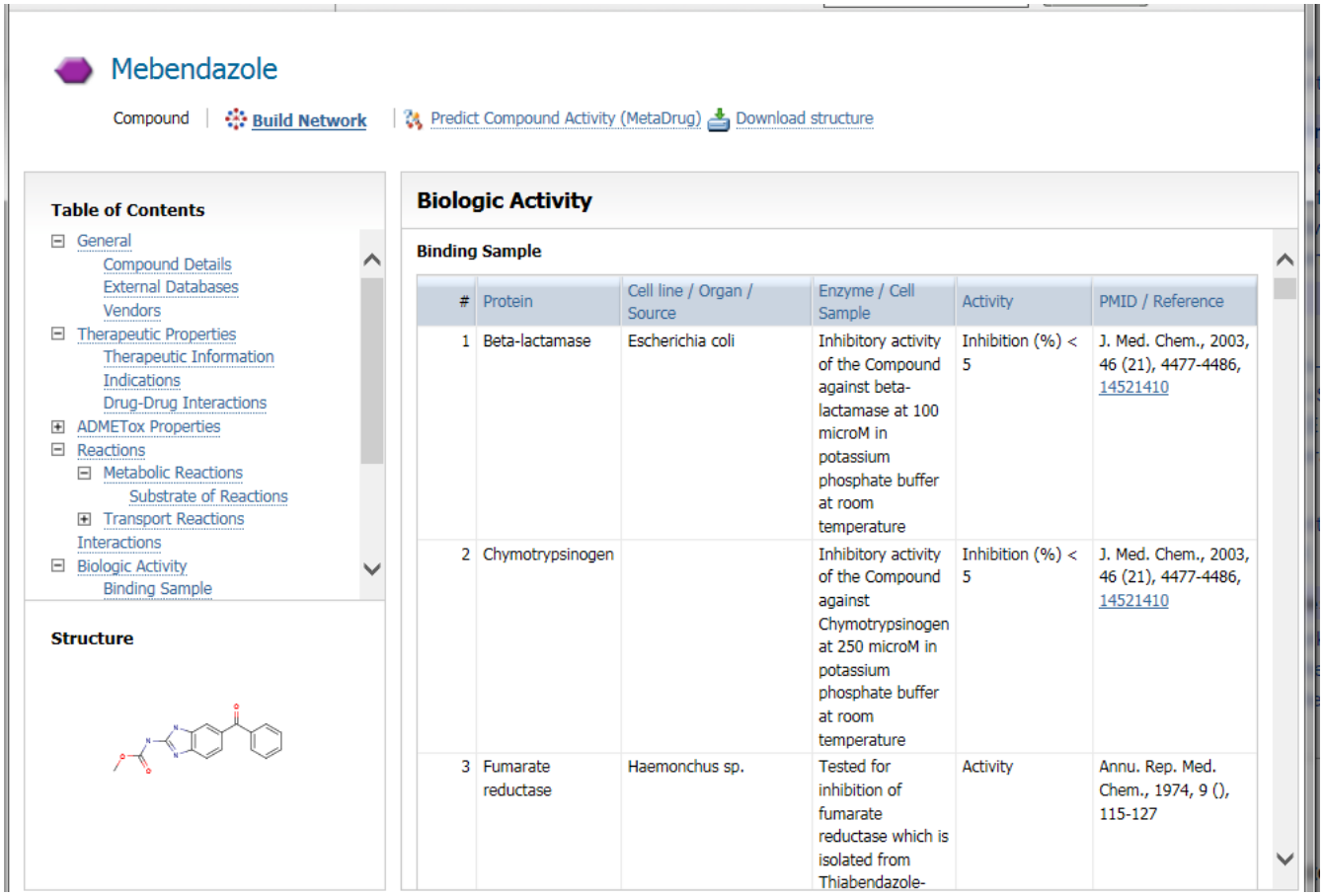


Different types of data can be overlaid and reflected on the network

Genomics, proteomics, metabolomics, drug assay and interaction data

- Microarray – list of differentially expressed genes
- Proteomics – lists of differentially present proteins
- PhosphoProteomics - lists of differentially phosphorylated proteins
- Methylation data - lists of differentially methylated genes
- Metabolomics data – chemical name, SMILES, CAS, MOL file, fold change between conditions
- Any data types that can be presented as gene list

MetaDrug module: new add-on to MetaCore



Systems pharmacology solution:
Extensive manually curated information on biological effects of small molecule compounds.

New: Cancer specialty modules
Add-on to MetaCore (~ 700 prebuilt pathway maps)

- PROSTATE CANCER
- PANCREATIC CANCER
- HEPATOCELLULAR CARCINOMA
- GASTRIC CANCER
- MULTIPLE MYELOMA
- LUNG CANCER
- COLORECTAL CANCER
- BREAST CANCER
- MELANOMA

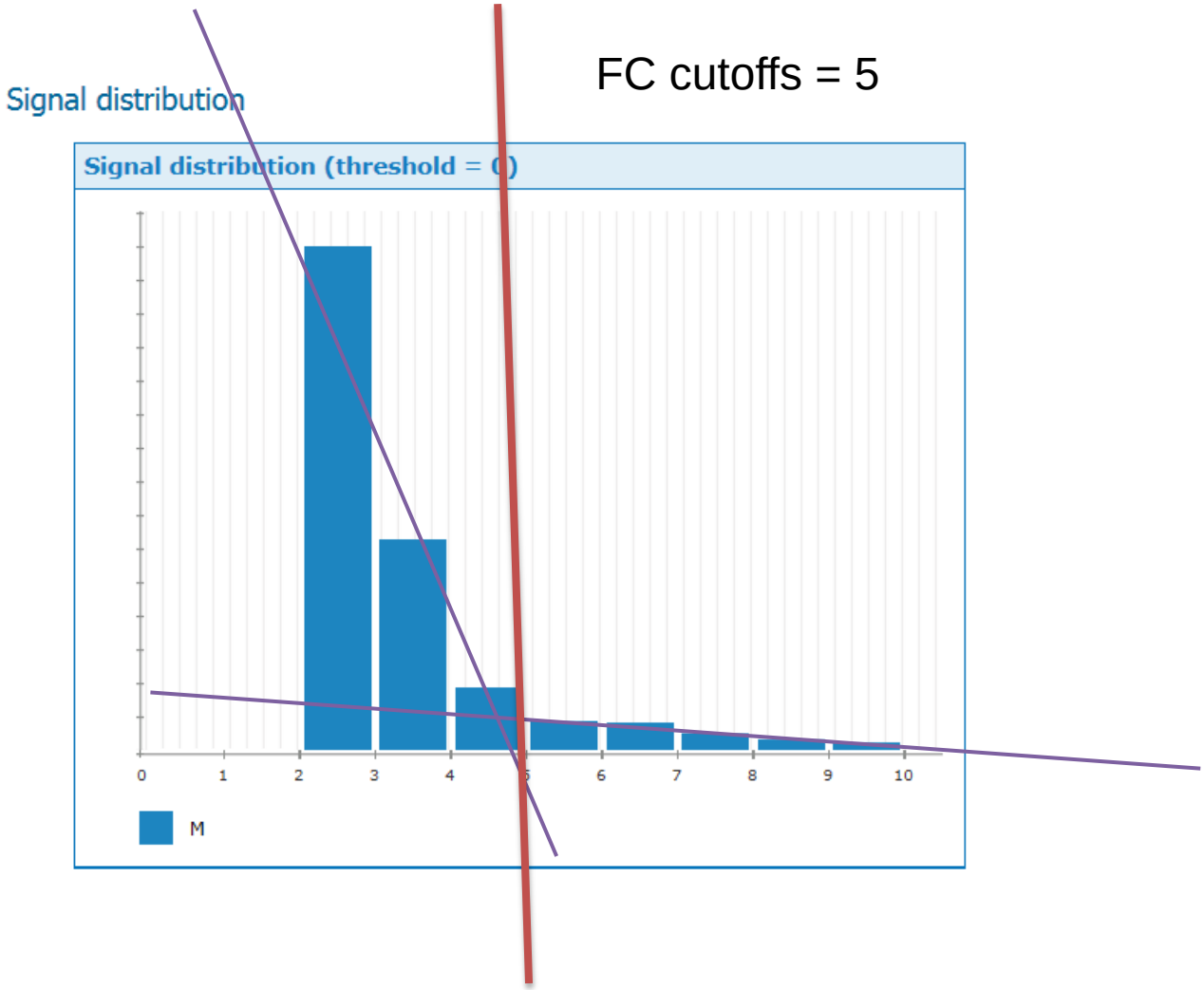
Tips and tricks

- Use IE on PC/Windows, Safari on Mac
- Upload data with relaxed P-value and Fold Change cutoffs; apply more restrictive cutoffs later
- Explore signal distributions to select better FC cutoffs
- Select appropriate background gene set
- Take advantage of multiple tabs (20-30) to accelerate your analysis
- Build multiple networks with increasingly relaxed conditions (more networks objects, more interaction types and less reliable source)
- Comb network down by removing disconnected nodes and interactions with unspecified effect
- Animate your data on the network to easy visualize the effects
- Save network often to protect from computer freeze

Define background gene set for enrichment analysis

- For a microarray experiment:
  - Use appropriate array definition (PROVIDED)
- For proteomics:
  - If your experiment includes only secreted proteins, using the whole genome background will result in a biased enrichment analysis
  - Solution: upload your list of secreted proteins as background

Rapid change of derivative gives a clue to a optimal/better FC cutoffs



- Do you have some useful MetaCore tricks ?
- Questions?
- Suggestions for future classes
- Thanks!