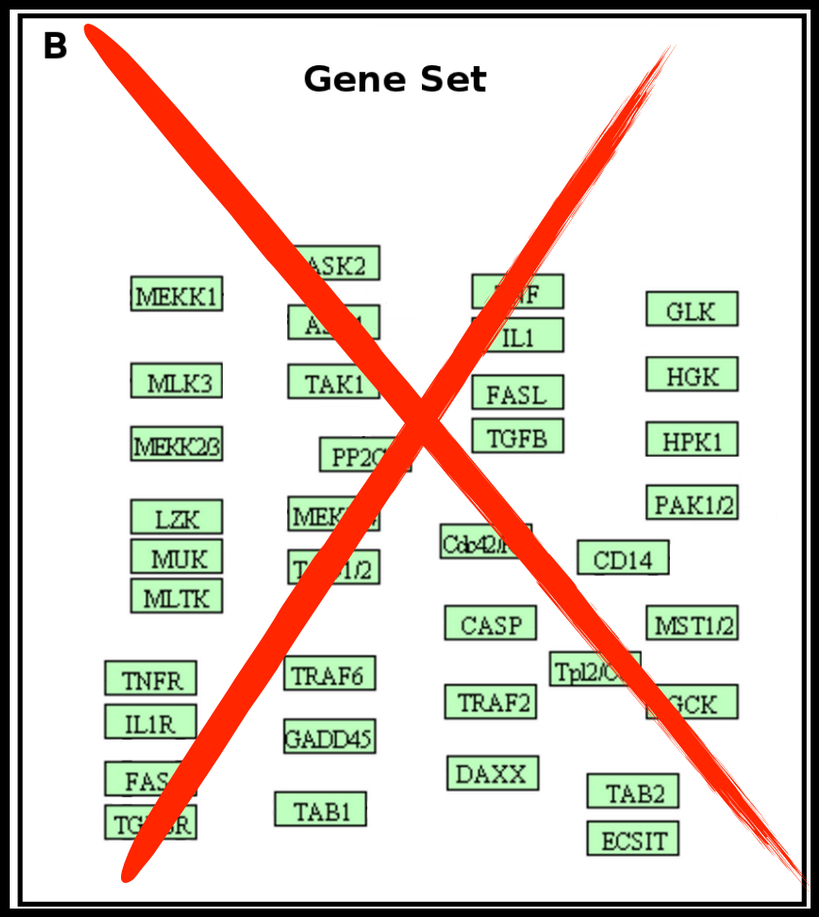
When is a pathway changed?

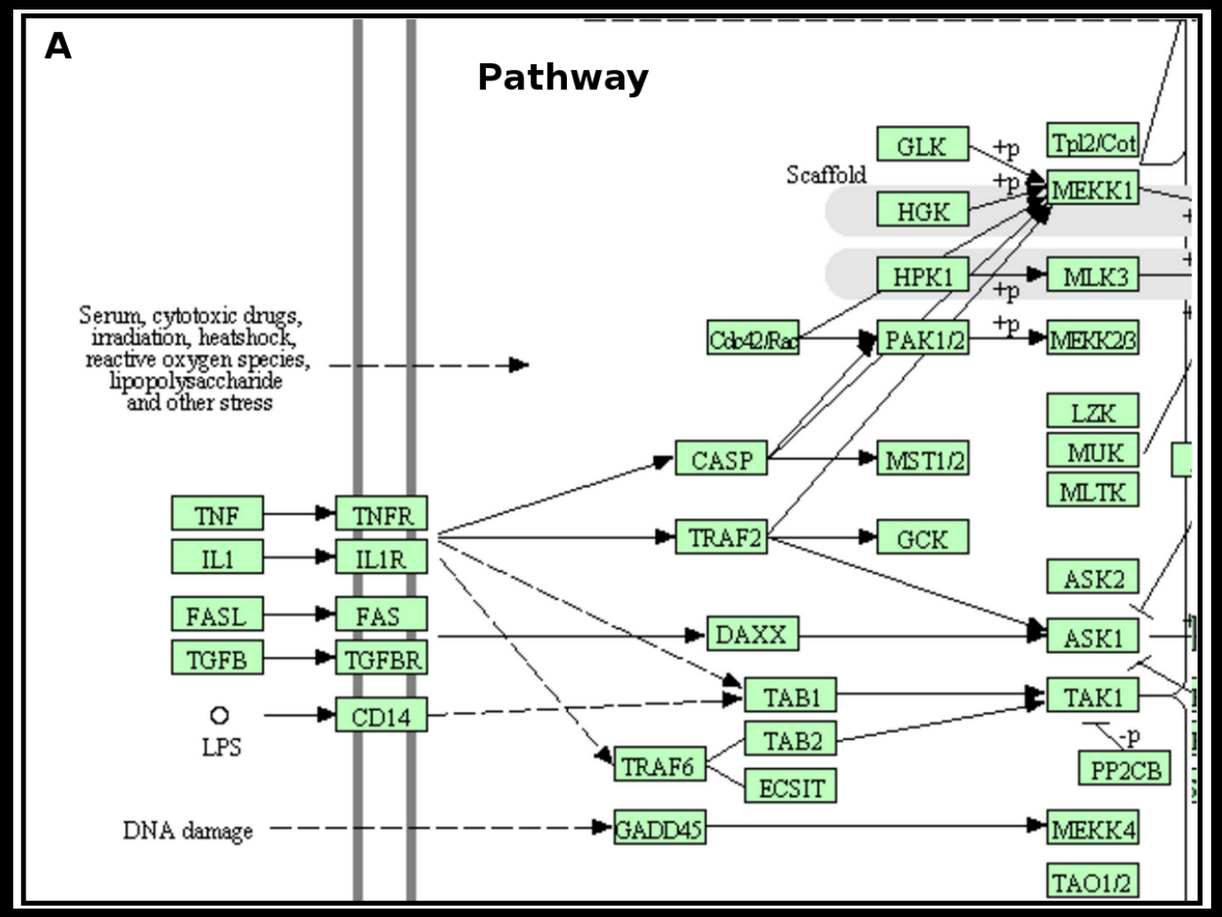
Sorin Draghici

Robert J. Sokol, MD Endowed Chair

Department of Computer Science
Department of Obstetrics and Gynecology
Wayne State University

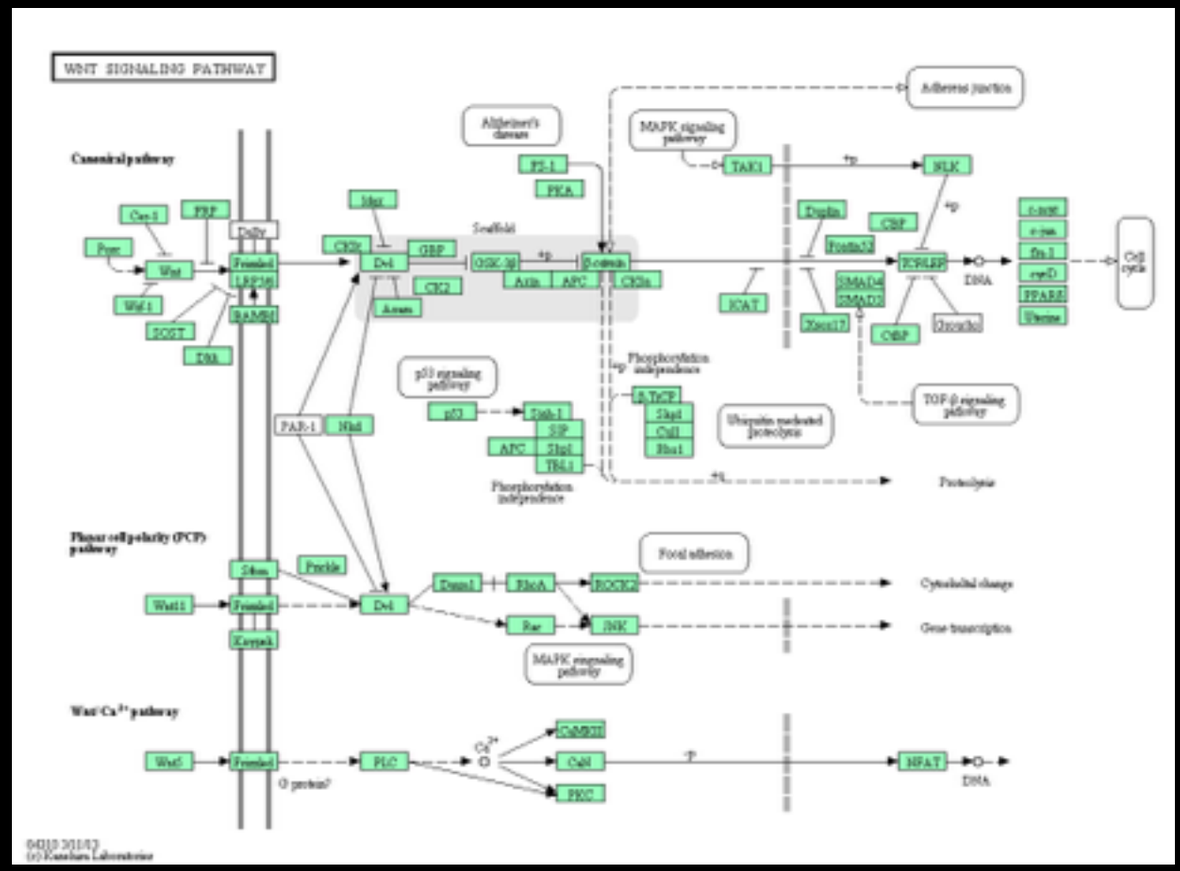
A gene set is not a





What is a pathway?

- **A sub-system with some properties**
  - Has some components (e.g. genes, biochemical compounds, etc)
  - Which are linked by some interactions (eg. activation/repression, reactions, etc.)
  - Interacts with the rest of the system through well defined inputs and outputs
  - There are more shared properties between the components of the sub-system than between these and the rest of the system (all genes related to insulin signaling, all reactions involved in the Krebs cycle, etc.)
- **Can be modeled well by a graph**
  - Components are represented by nodes
  - Interactions are represented by directed edges
  - The edge direction represents signals, reactions, etc.



Problem definition

- Input:
  - A set of variables (e.g., genes, proteins, etc.) for which there are significant measured differences between the given phenotype and control - **measured gene expression changes.**
  - A set of pathways describing sub-systems involving the given variables (e.g., signaling pathways, metabolic pathways) - **pathway database (e.g., KEGG, Reactome).**
- Output:
  - **Rank the sub-systems** in the decreasing order of the amount of disruption suffered
  - If possible, identify those sub-systems for which the **disruption is significant** (i.e. unlikely to be due to chance or noise alone)
  - **Identify potential mechanisms** that led to the measured expression changes.

Overview of existing

- Over-representation analysis (ORA)
  - Relative enrichment (GenMapp, Dahlquist, 2002)
  - Model based statistical approach (Onto-Express, Draghici, 2003)
  - Resampling based approaches (SAFE, Barry, 2005)
- Functional class-scoring (FCS)

Classical approaches

- **Relative enrichment**

  - N – genes on the array used
  - K – differentially regulated genes
  - NP– genes on the pathway
  - NRP– differentially regulated genes on the pathway

$$RE = \frac{\frac{N_{RP}}{N_P}}{\frac{K}{N}}$$

Classical approaches

Model based statistical approach (hypergeometric)

N – genes on the array used

K – differentially regulated genes

N_P– genes on the pathway

N_{RP}– differentially regulated genes on the pathway

The probability of having exactly N_{RP} genes on the given pathway:

$$P(X = N_{RP}|N, M, K) = \frac{\binom{N_P}{N_{RP}}\binom{N-N_P}{K-N_{RP}}}{\binom{N}{K}}$$

The probability of having more than or equal to N_{RP} genes on the given pathway just by chance:

$$p = 1 - \sum_{i=0}^{N_{RP}} \frac{\binom{N_P}{i}\binom{N-N_P}{K-i}}{\binom{N}{K}}$$

Classical approaches

- Resampling based statistical approach
  - SAFE - Significance Analysis of Function and Expression
  - t-test at gene level
  - Wilcoxon rank sum as default global statistic
  - Permutation based method
  - Uses FDR for multiple testing correction

GSEA

Gene Set Enrichment Analysis (GSEA)

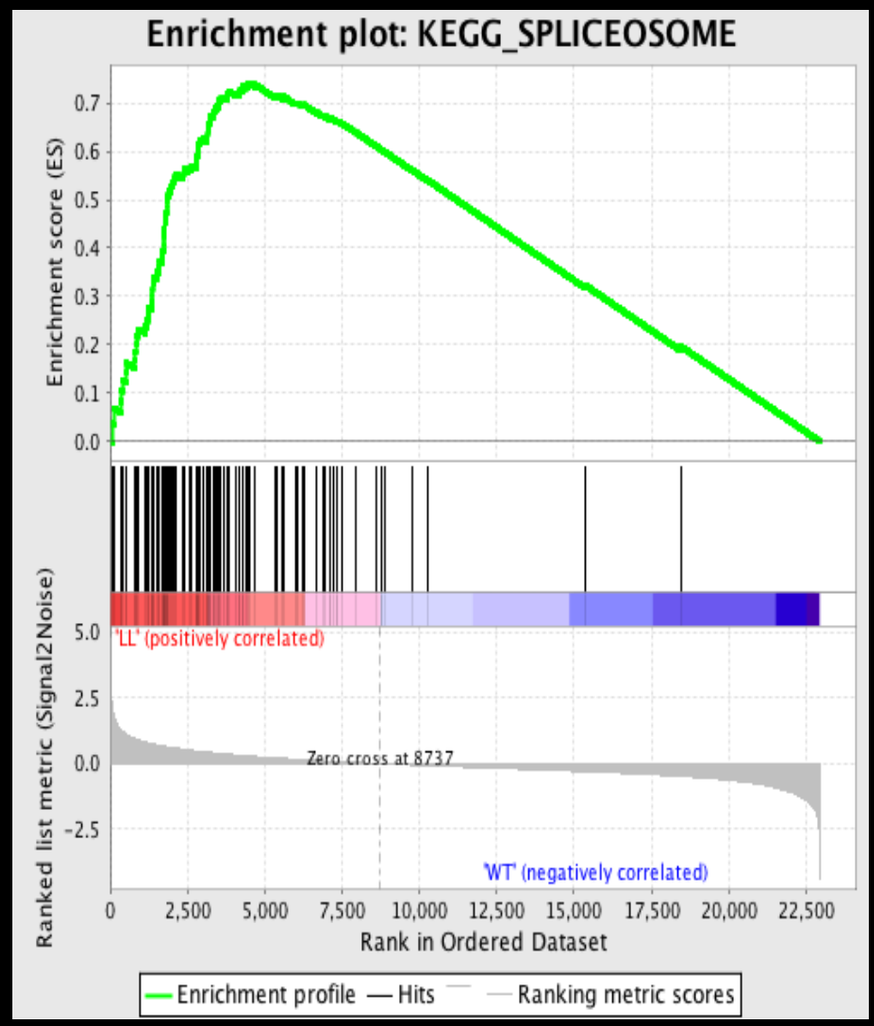
All the genes are ranked based on the correlation to the phenotype

Going down in the ranked list, the Enrichment Score (ES) will be increased if a gene exists in the gene set(P_{hit}), and decreased if not(P_{miss}).

The final Gene Set Enrichment Score is the maximum deviation from zero.

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

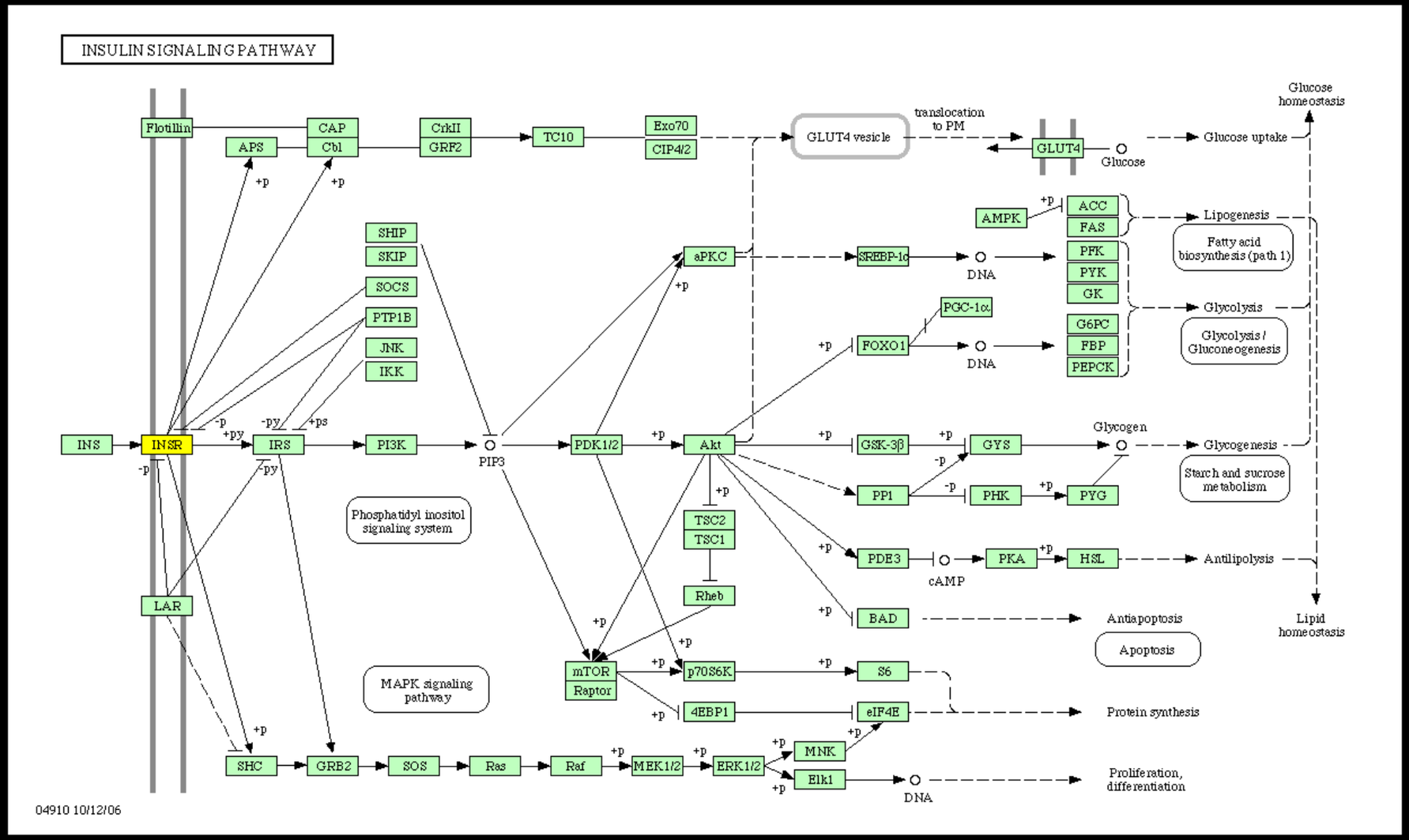
$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$



Limitations of gene set

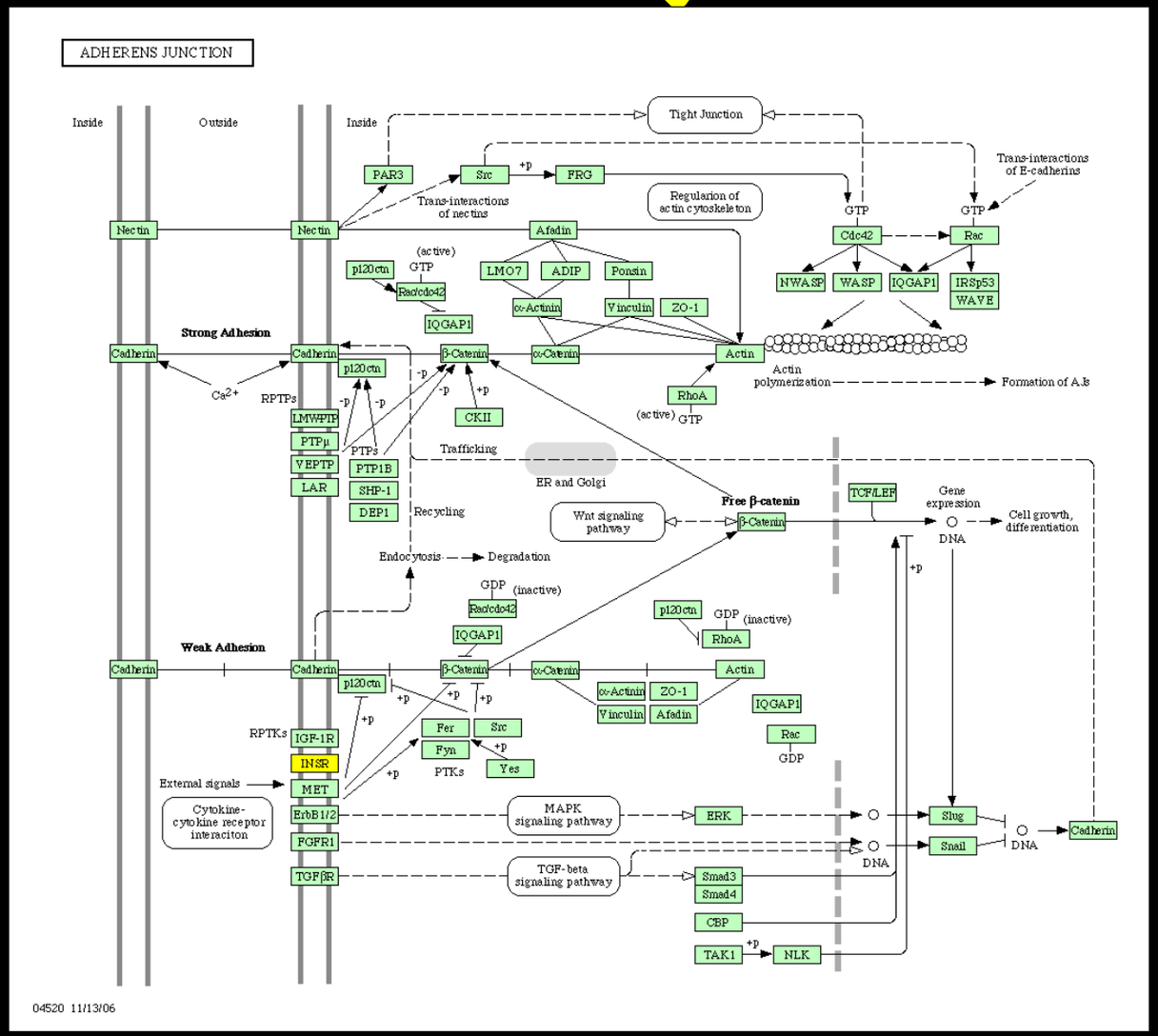
- **Classical techniques only consider the number of genes on a given pathway ignoring other crucial aspects**
  - in the classical approach all genes are the same
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their identity (not good because a small change in a transcription factors could trigger large downstream effects, for instance)
- **The position and role of the DE genes on the pathways**
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their position and role on the pathway (eg INSR on insulin and adherens junction pathways)

Insulin signaling pathway



Insulin receptor (INSR) is the crucial gene on this pathway

Adherens junction



INSR is one of many tyrosine kinase receptors

Limitations of gene set

- **Classical techniques only consider the number of genes on a given pathway ignoring other crucial aspects**
  - in the classical approach all genes are the same
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their identity (not good because a small change in a transcription factors could trigger large downstream effects, for instance)

- **The position and role of the DE genes on the pathways**
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their position and role on the pathway (eg INSR on insulin and adherens junction pathways)
  - genes upstream on the pathway could have a much greater impact than genes downstream

- **The topology of the pathway**
  - all current methods yield the same results even if the topology of the pathway is drastically changed
  - genes with many outgoing regulatory connections could perturb the pathway more than genes with few such connections
  - are the DE genes scattered randomly across the pathway or grouped together in a clique that might suggest a coherent perturbation propagating in the pathway?

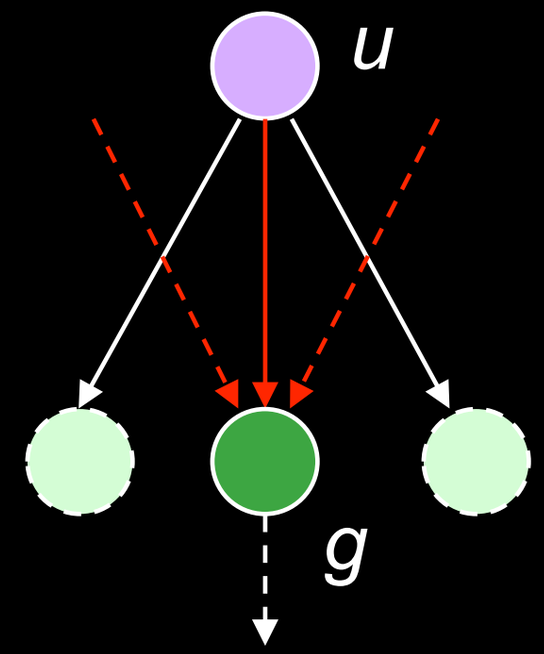
- **The specific expression values measured are completely ignored by all existing techniques**

Impact analysis

For each input gene *g*, a gene perturbation factor PF(g) depends on:

- User-provided normalized fold change of the gene
- Number and amount of perturbation of genes upstream
- The interactions on the pathway (its topology)

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \frac{PF(u)}{N_{ds}(u)}$$



PF(g) – perturbation factor of gene *g*
ΔE(g) – change in expression level of gene *g*
US_g – set of genes directly upstream of *g*
$N_{ds}(u)$ - number of genes directly downstream of u
β_{ug} - efficiency of the connection between u and *g*

Pathway perturbation

For each pathway, the impact of topology, gene interactions, and gene fold changes come into play and are captured through a pathway perturbation factor:

$$PF(P_i) = \frac{\sum_{g \in P_i} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)}$$

$$\overline{|\Delta E|} = \frac{\sum_{k=1}^{N} |\Delta E|}{N}$$

mean fold change over all DE genes

N_{de} (Pi) is the number of DE genes on the given pathway Pi
PF(g) is the perturbation of the gene g

Pathway impact factors

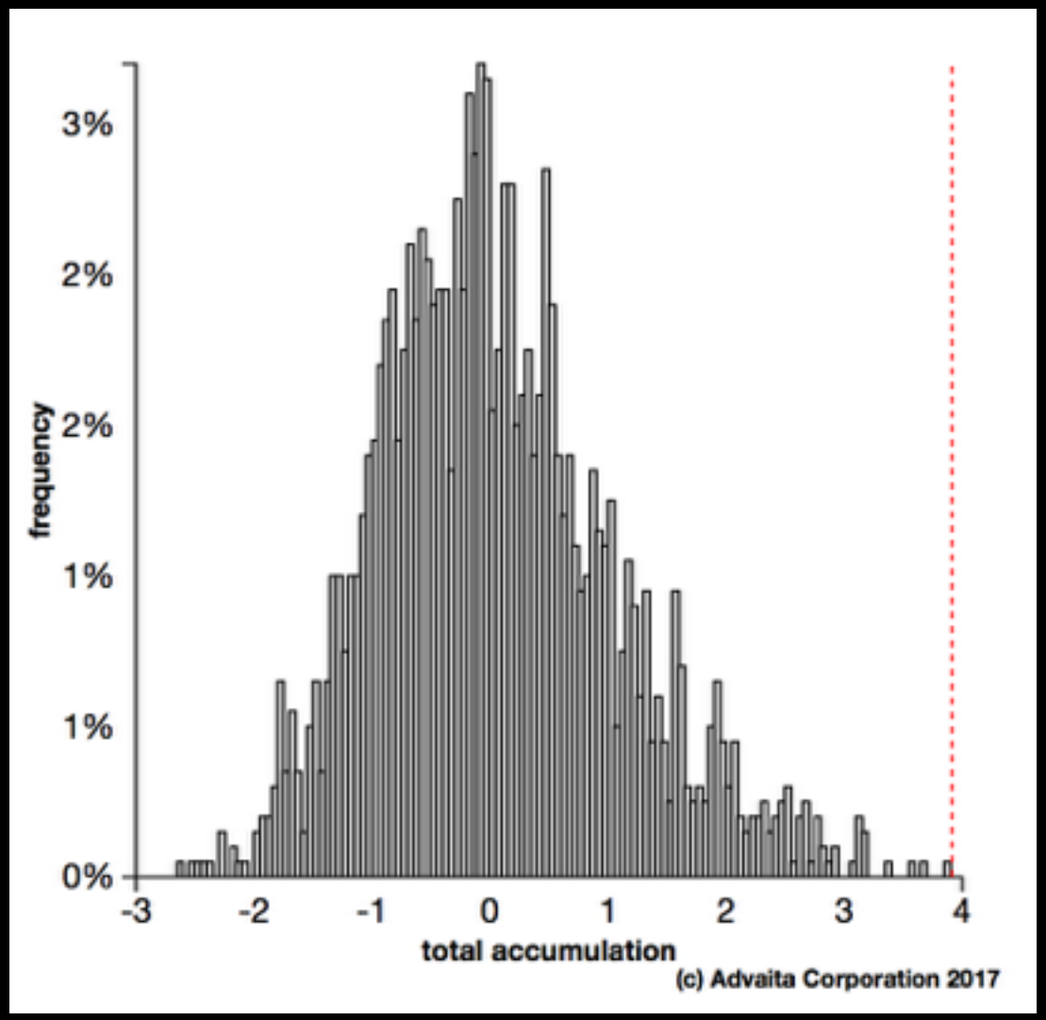
The impact factor of a pathway depends on:

1. A probabilistic term - depends on the proportion of differentially regulated genes on the given pathway
2. A perturbation term - depends on the specific genes that are differentially regulated, the measured amount of expression change and the interactions on the pathway

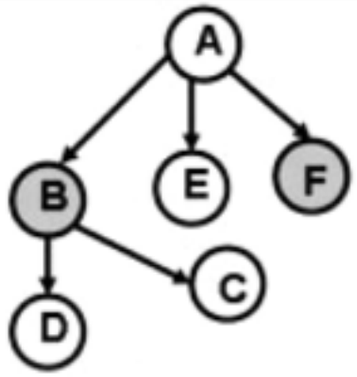
$$IF(P_i) = log\left(\frac{1}{p_i}\right) + PF(P_i)$$

P_i - pathway i

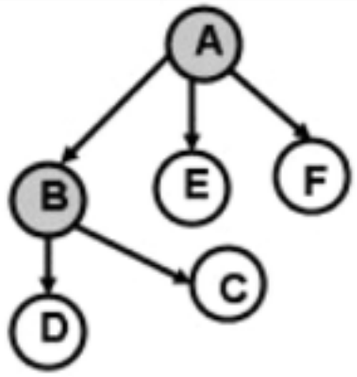
p-value calculated based



What the impact



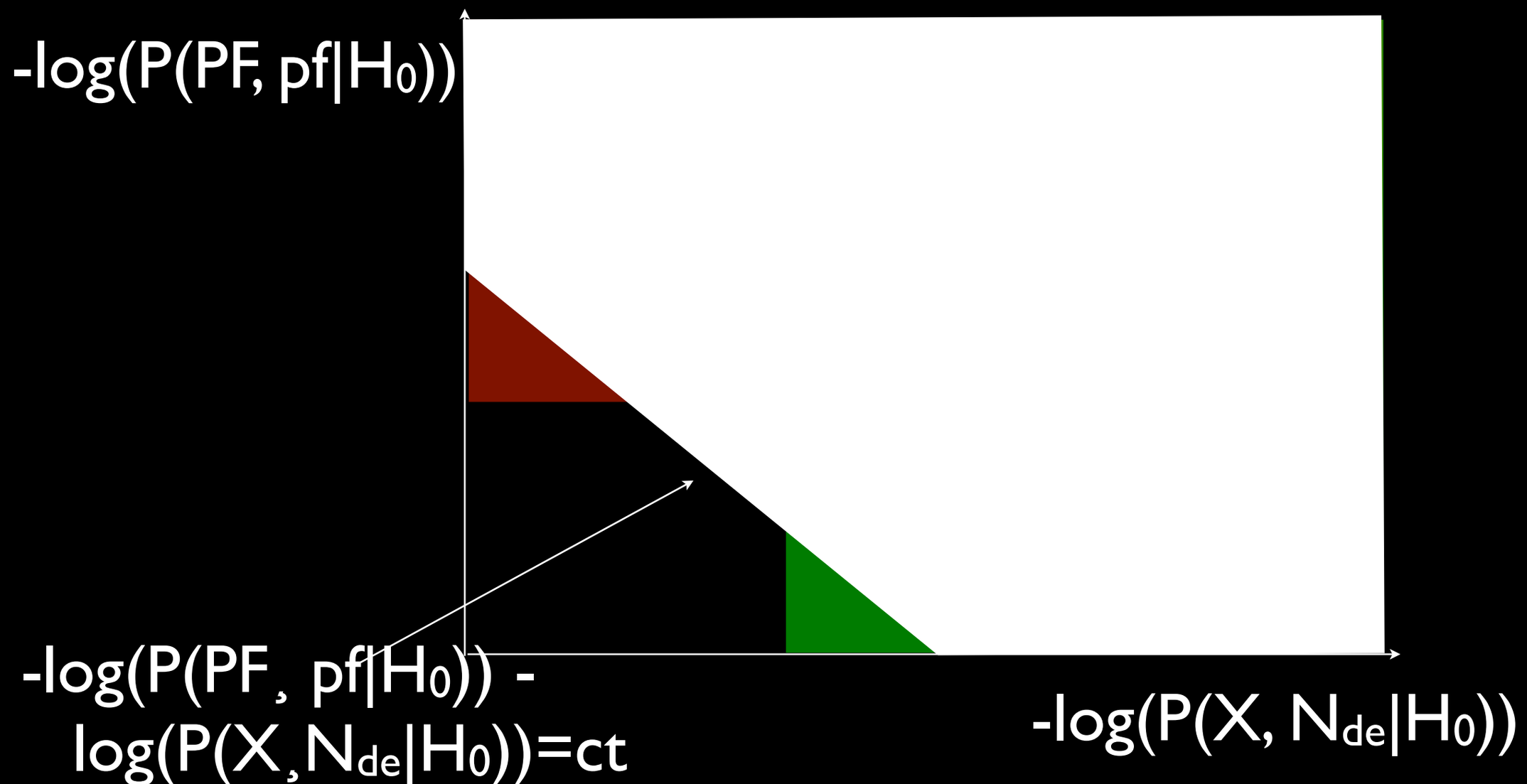
Gene	$\Delta E(g)$	$PF(g)$
A	0	0
B	2	2
C	0	1
D	0	1
E	0	0
F	3	3
$PF(P)$		7.0



Gene	$\Delta E(g)$	$PF(g)$
A	1.5	1.5
B	2	2.5
C	0	1.25
D	0	1.25
E	0	0.5
F	0	0.5
$PF(P)$		7.5

Adapted from: Tarca et al. "A novel signaling pathway impact analysis." *Bioinformatics* 25, no. 1 (2008): 75-82.

Hypergeometric and perturbation analyses



How does one validate pathway analysis methods?

How do we asses the quality of the results obtained with each pathway analysis method? How do we compare different methods?

Possible methods:

1. Validation against known pathways associated to a specific disease in “some” datasets

2. Use “target” pathways

Lung adenocarcinoma

- 97 genes identified as good markers of survival

BAG1	-1.19047619	GRB7	1.38	CTSL	-1.111111111	DBP	1.04
CASP4	-1.063829787	INHA	1.23	FUCA1	1.22	GARS	1.38
FADD	1.57	ITK	-3.125	FUT3	2.14	HRB	1.25
P63	1.37	NACA	-1.098901099	GAPD	2.73	HSU53209	1.06
5T4	2.39	STC1	2.5	GCNT1	1.86	PRDM2	-2
ITGA2	2.09	TNFAIP6	-1.041666667	HMBS	1.56	RELA	-1.075268817
KRT18	2.92	VEGF	2.74	KYNU	1.62	RPS26	1.12
KRT19	2.65	VLDLR	-1.694915254	MLN64	1.32	RPS3	1.12
KRT7	2.26	WNT1	3.52	MSH3	1.12	RPS6KB1	1.03
LAMB1	-1.25	WNT10B	1.31	MT2A	-1.818181818	SUI1	1.05
TMSB4X	-1.492537313	HSPA8	1.08	NME2	1.58	TIEG	-1.315789474
TUBA1	1.65	ERBB2	1.92	NP	1.14	TMF1	1.51
BMP2	-1.265822785	FXYD3	2.11	PACE	1.69	B1	-1.020408163
CDC6	11.7	HLA-B	1	PDE7A	1.33	FEZ2	1.09
H2AFZ	1.31	HPCAL1	1.16	PLGL	-3.125	HPIP	1.31
PDAP1	1.37	P2RX5	1.02	PPIF	2.25	KIAA0005	1.4
POLD3	2.27	PEX7	1.49	PTPRCAP	-1.063829787	KIAA0020	1.35
REG1A	2.12	SLC20A1	1.58	RPC	1.33	KIAA0084	-1.351351351
S100P	16.72	SLC2A1	3.46	SC4MOL	-1.612903226	KIAA0153	1.4
SERPINE1	1.72	VDAC2	1.33	SLC1A6	-1.470588235	KIAA0263	-1.333333333
STX1A	1.54	ALDH8	2.5	UBC	-1.204819277	KIAA0317	1.32
ADM	1.39	ALDOA	1.48	UGP2	1.13	MGB1	2.25
AKAP12	-1.886792453	ATP2B1	-1.724137931	UQCRC2	-1.030927835		
ARHE	-1.639344262	CDS1	1.03	COPEB	-1.492537313		
DEFB1	1.41	CSTB	1.5	CRK	1.32		

Lung cancer

- Poor vs. good survival in lung adenocarcinoma (Beer,

ORA

Pathway name	ORA (hypergeometric)		
	p-value	FDR	Bonferroni
Prion disease	0.14964	0.627132	1
Focal adhesion	0.15542	0.627132	1
Parkinson's disease	0.16484	0.627132	1
Dentatorubropallidoluysian atrophy	0.17978	0.627132	1
Calcium signaling pathway	0.26288	0.627132	1
Alzheimer's disease	0.27710	0.627132	1
Apoptosis	0.28374	0.627132	1
TGF-beta signaling pathway	0.30368	0.627132	1
Huntington's disease	0.32749	0.627132	1
Toll-like receptor signaling pathway	0.33008	0.627132	1
Wnt signaling pathway	0.36914	0.637613	1
Regulation of actin cytoskeleton	0.43939	0.695701	1
MAPK signaling pathway	0.56081	0.762988	1
Phosphatidylinositol signaling system	0.57239	0.762988	1
Adherens junction	0.60235	0.762988	1
Complement and coagulation cascades	0.68033	0.766820	1
Cell cycle	0.68610	0.766820	1
Cytokine-cytokine receptor interaction	0.82065	0.866242	1
Neuroactive ligand-receptor interaction	0.97299	0.972996	1

GSEA

Enriched in cancer			
Pathway Name	NOM p-val	FDR q-val	FWER p-val
Cell cycle	0.038	0.118	0.140
Huntington's disease	0.074	0.217	0.546
Dentatorubropallidoluysian atrophy	0.149	0.291	0.751
Alzheimer's disease	0.189	0.344	0.877
Parkinson's disease	0.373	0.485	0.984
Adherens junction	0.583	0.651	0.998
Wnt signaling pathway	0.861	0.785	1

Enriched in normal			
Pathway Name	NOM p-val	FDR q-val	FWER p-val
MAPK signaling pathway	0.007	0.170	0.361
Apoptosis	0.019	0.175	0.304
Complement and coagulation cascades	0.037	0.255	0.298
Phosphatidylinositol signaling system	0.189	0.343	0.823
Regulation of actin cytoskeleton	0.010	0.356	0.223
Focal adhesion	0.160	0.384	0.817
Cytokine-cytokine receptor interaction	0.241	0.420	0.910
Toll-like receptor signaling pathway	0.330	0.451	0.963
Calcium signaling pathway	0.308	0.489	0.960
Prion disease	0.474	0.563	0.986
TGF-beta signaling pathway	0.631	0.699	0.998
Neuroactive ligand-receptor interaction	0.947	0.957	1

Impact Analysis

Pathway name	Impact Factor			
	IF	p-value	FDR	Bonferroni
Cell cycle	19.259	8.76E-08	1.66E-06	1.66E-006
Focal adhesion	7.414	0.005072	0.048180	0.095683
Wnt signaling pathway	6.780	0.008840	0.055988	0.167964
Dentatorubropallidoluysian atrophy	5.535	0.025786	0.122495	0.489981
Huntington's disease	4.543	0.058988	0.203925	1
Apoptosis	4.407	0.065921	0.203925	1
Regulation of actin cytoskeleton	4.246	0.075130	0.203925	1
TGF-beta signaling pathway	3.511	0.134730	0.319984	1
Complement and coagulation cascades	3.161	0.176357	0.354145	1
Adherens junction	2.953	0.206279	0.354145	1
Alzheimer's disease	2.752	0.239376	0.354145	1
Parkinson's disease	2.631	0.261459	0.354145	1
Toll-like receptor signaling pathway	2.576	0.272054	0.354145	1
Prion disease	2.572	0.272839	0.354145	1
Calcium signaling pathway	2.538	0.279588	0.354145	1
Cytokine-cytokine receptor interaction	2.353	0.318815	0.366952	1
Phosphatidylinositol signaling system	2.311	0.328326	0.366952	1
MAPK signaling pathway	2.205	0.353353	0.372984	1
Neuroactive ligand-receptor interaction	0.576	0.885936	0.885936	1

Draghici et al, [A systems biology approach for pathway level analysis](). Genome Research, 17:1537-1545, 2007

Implemented in the R-OntoTools and SPIA Bioconductor packages, iPathway-Guide

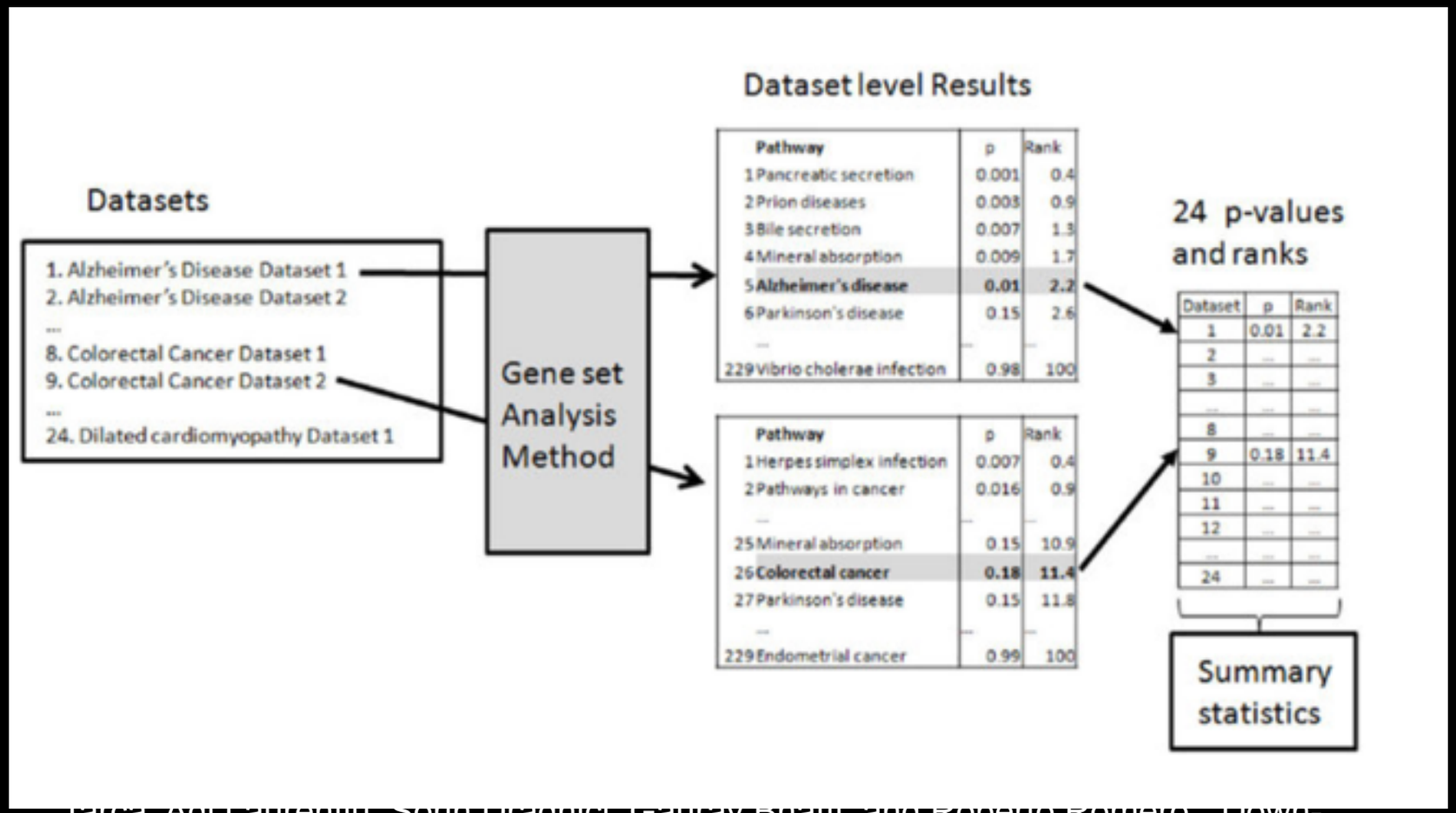
Datasets

	GEOID	Pubmed	Ref.	Disease/Target pathway	KEGGID	Tissue
1	GSE1297	14769913	[20]	Alzheimer's Disease	hsa05010	Hippocampal CA1
2	GSE5281	17077275	[21]	Alzheimer's Disease	hsa05010	Brain, Entorhinal Cortex
3	GSE5281	17077275	[21]	Alzheimer's Disease	hsa05010	Brain, hippocampus
4	GSE5281	17077275	[21]	Alzheimer's Disease	hsa05010	Brain, Primary visual cortex
5	GSE20153	20926834	[22]	Parkinson's disease	hsa05012	Lymphoblasts
6	GSE20291	15965975	[23]	Parkinson's disease	hsa05012	Postmortem brain putamen
7	GSE8762	17724341	[24]	Huntington's disease	hsa05016	Lymphocytes (blood)
8	GSE4107	17317818	[25]	Colorectal Cancer	hsa05210	Mucosa
9	GSE8671	18171984	[26]	Colorectal Cancer	hsa05210	Colon
10	GSE9348	20143136	[27]	Colorectal Cancer	hsa05210	Colon
11	GSE14762	19252501	[28]	Renal Cancer	hsa05211	Kidney
12	GSE781	14641932	[29]	Renal Cancer	hsa05211	Kidney
13	GSE15471	19260470	[30]	Pancreatic Cancer	hsa05212	Pancreas
14	GSE16515	19732725	[31]	Pancreatic Cancer	hsa05212	Pancreas
15	GSE19728		-	Glioma	hsa05214	Brain
16	GSE21354		-	Glioma	hsa05214	Brain, Spine
17	GSE6956	18245496	[32]	Prostate Cancer	hsa05215	Prostate
18	GSE6956	18245496	[32]	Prostate Cancer	hsa05215	Prostate
19	GSE3467	16365291	[33]	Thyroid Cancer	hsa05216	Thyroid
20	GSE3678		-	Thyroid Cancer	hsa05216	Thyroid
21	GSE9476	17910043	[34]	Acute myeloid leukemia	hsa05221	Blood, Bone marrow
22	GSE18842	20878980	[35]	Non-Small Cell Lung Cancer	hsa05223	Lung
23	GSE19188	20421987	[36]	Non-Small Cell Lung Cancer	hsa05223	Lung
24	GSE3585	17045896	[37]	Dilated cardiomyopathy	hsa05414	Heart

- 24 datasets
- 12 different conditions

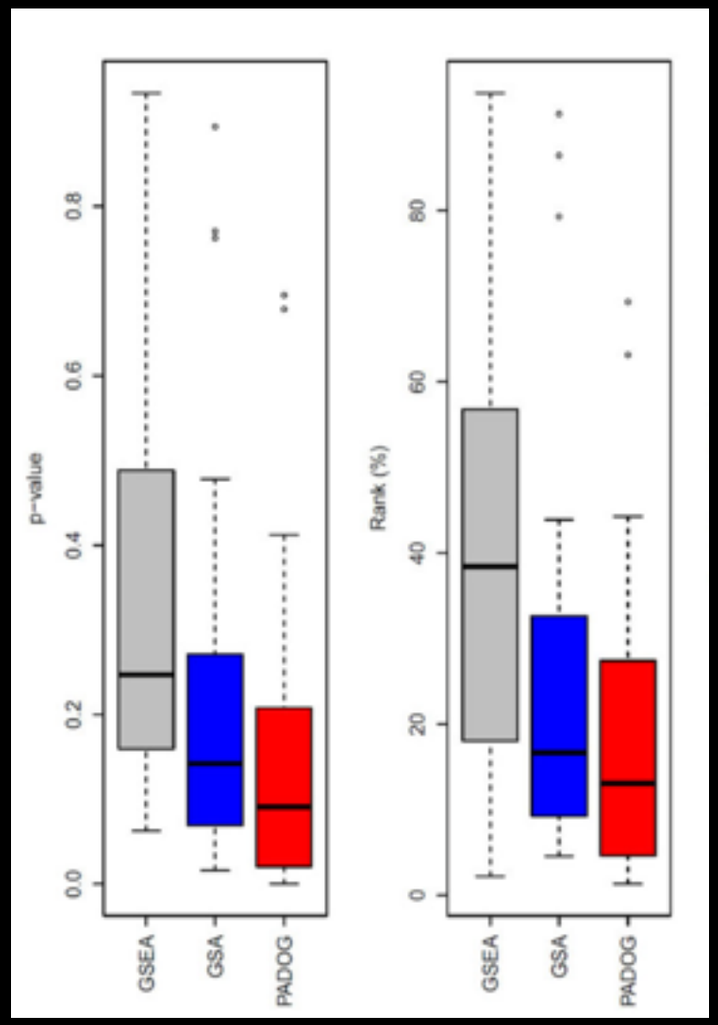
Tarca et al, ”Down-weighting overlapping genes improves gene set analysis." *BMC Bioinformatics* 13, no. 1 (2012): 136.

Target pathway



Tarca, Adi Laurentiu, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. "Down-weighting overlapping genes improves gene set analysis." *BMC bioinformatics* 13, no. 1 (2012): 136.

Comparing analysis



- 24 datasets
- 229 KEGG pathways
- 12 different conditions

Tarca, Adi Laurentiu, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. "Down-weighting overlapping genes improves gene set analysis." *BMC bioinformatics* 13, no. 1 (2012): 136.

Is there maybe a gold standard?

- Knock-out (KO) data sets
  - In a KO experiment, the precise cause of the phenotype changes is known
  - The target pathways (true positives) include all pathways that contain the KO gene
  - A good pathway analysis method would be able to retrieve as many TPs as possible and ideally, pinpoint

Data set: KO of gene Myd88

- The knockout gene of GSE19793 dataset is myd88 in mice
- Wild type mice of identical background were used as control groups.
- 10 biological repeats are performed for the wild type and myd88 samples.
- The platform that is used in this experiment is Affymetrix Mouse Genome 430 2.0 Array.

	Target Pathways	Entry
1	NF-kappa B signaling pathway	4064
2	Apoptosis	4210
3	Toll-like receptor signaling pathway	4620
4	Salmonella infection	5132
5	Pertussis	5133
6	Legionellosis	5134
7	Leishmaniasis	5140
8	Chagas disease (American trypanosomiasis)	5142
9	African trypanosomiasis	5143
10	Malaria	5144
11	Toxoplasmosis	5145
12	Tuberculosis	5152
13	Hepatitis B	5161
14	Measles	5162
15	Influenza A	5164
16	Herpes simplex infection	5168

Enrichment vs. Impact Analysis on the

Results

Rank	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pG FDR
2	Mineral absorption	04978	43	2	0	0.0058...	1	0.0358...	0.5028...
13	Taste transduction	04742	29	1	0	0.0742...	1	0.2672...	0.7048...
12	Calcium signaling pathway	04020	178	2	-8.7729...	0.0811...	0.816	0.2458...	0.7048...
14	African trypanosomiasis	05143	34	1	0	0.0864...	1	0.2981...	0.7048...
15	Malaria	05144	44	1	0	0.1104...	1	0.3538...	0.7048...
16	Type II diabetes mellitus	04930	49	1	7.31029...	0.1222...	0.965	0.3700...	0.7048...
17	Legionellosis	05134	55	1	-6.3680...	0.1361...	0.901	0.3800...	0.7048...
18	Long-term depression	04730	60	1	-7.8598...	0.1475...	0.908	0.4033...	0.7048...
20	Colorectal cancer	05210	64	1	-2.3659...	0.1566...	0.884	0.4122...	0.7048...
3	Leishmaniasis	05140	66	1	2.229	0.1611...	0.04	0.0389...	0.5028...
19	Olfactory transduction	04740	67	1	-6.8572...	0.1633...	0.847	0.4119...	0.7048...
4	Pertussis	05133	70	1	1.486	0.1700...	0.068	0.0631...	0.5028...
22	RNA degradation	03018	70	1	0	0.1700...	1	0.4712...	0.7048...

Hypergeometric p-value (eg. DAVID, Ingenuity Pathway

Results

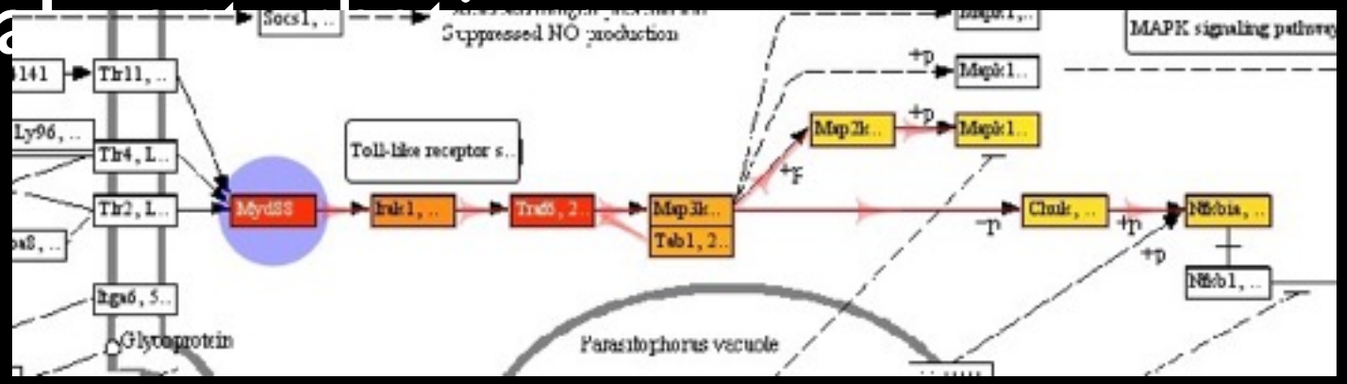
Rank	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pG FDR
1	Toxoplasmosis	05145	111	1	2.72433...	0.2560...	0.02	0.0321...	0.5028...
5	Regulation of actin cytoskelet	04810	210	1	-5.6448	0.4292...	0.027	0.0632...	0.5028...
3	Leishmaniasis	05140	66	1	2.229	0.1611...	0.04	0.0389...	0.5028...
6	Measles	05162	132	1	2.229	0.2966...	0.043	0.0683...	0.5028...
4	Pertussis	05133	70	1	1.486	0.1700...	0.068	0.0631...	0.5028...
9	Influenza A	05164	157	1	1.486	0.3421...	0.071	0.1146...	0.5730...
7	Toll-like receptor signaling p	04620	94	1	4.37839...	0.2214...	0.075	0.0846...	0.5028...
8	Chagas disease (American trypa	05142	100	1	1.486	0.2338...	0.076	0.0893...	0.5028...
10	Tuberculosis	05152	170	1	2.229	0.3646...	0.119	0.1795...	0.7048...
11	NF-kappa B signaling pathway	04064	87	1	1.51935...	0.2068...	0.216	0.1835...	0.7048...
21	ECM-receptor interaction	04512	82	1	-4.8276...	0.1961...	0.766	0.4350...	0.7048...
23	PI3K-Akt signaling pathway	04151	322	2	-0.5306...	0.2104...	0.813	0.4732...	0.7048...
12	Calcium signaling pathway	04020	178	2	-8.7729...	0.0811...	0.816	0.2458...	0.7048...

Perturbation p-value (iPathway-Guide, R-Onto-Tools, SPIA)

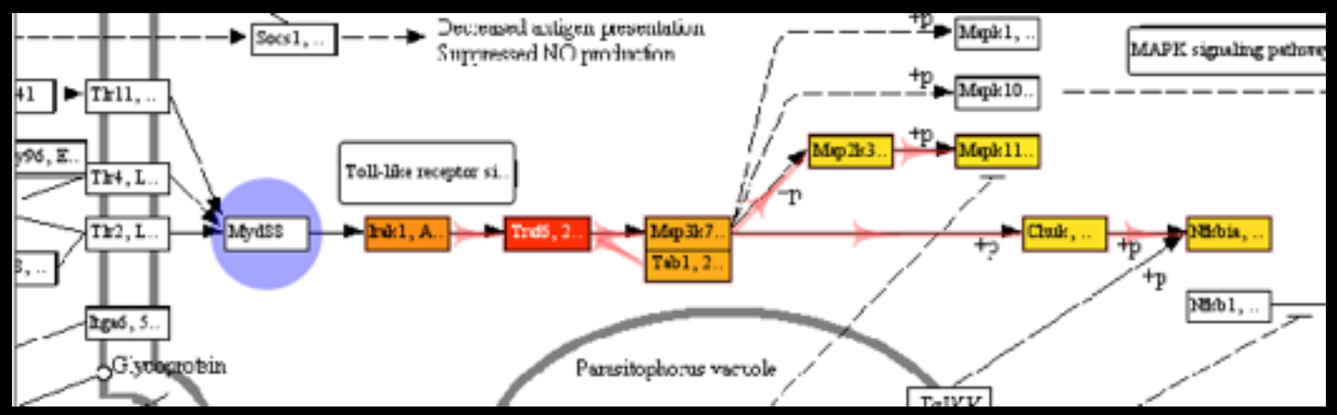
Impact Analysis results on the Myd88 KO

Coherent Change of perturbation propagation in Toxoplasmosis pathway.

Based on tota



Based on total accumulation:



Impact pathway analysis

- Draghici et al, A systems biology approach for pathway level analysis.
Genome Research, 17:1537-1545, 2007
> 700 citations

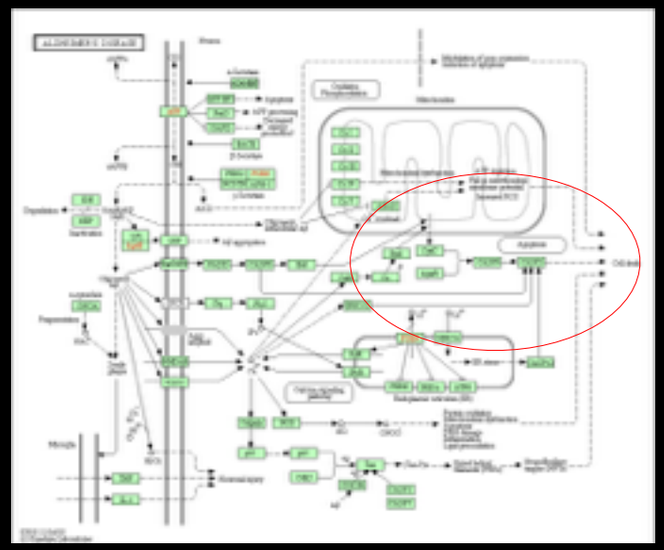
- Tarca et al. A novel signaling pathway impact analysis
Bioinformatics 25, no. 1 (2008): 75-82.
> 500 citations

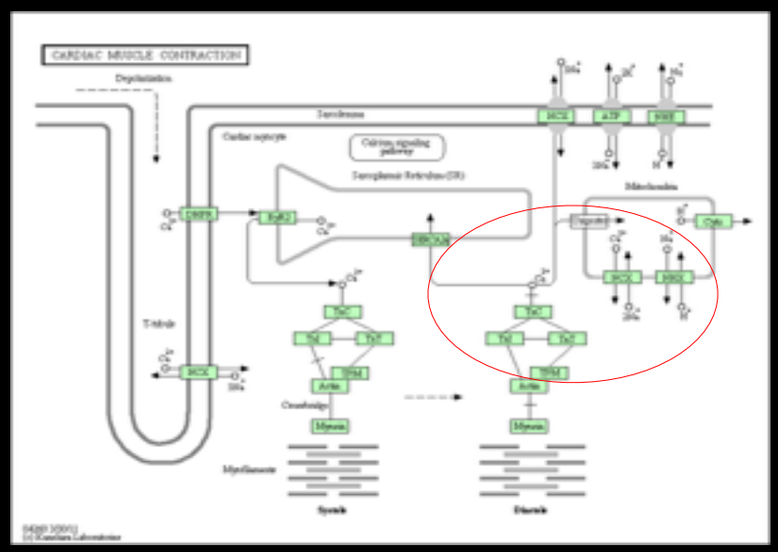
Pathway cross-talk phenomena
Fat remodeling experiment

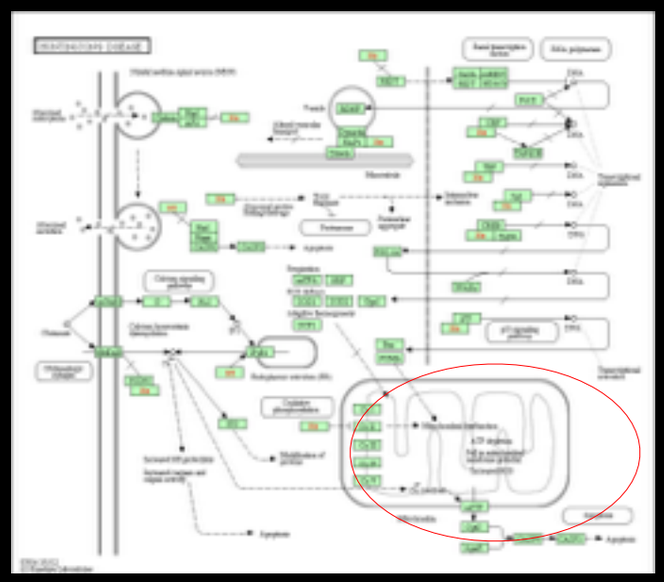
title	p.adj.fdr.
Parkinson's disease	2.05E-06
Alzheimer's disease	3.57E-06
Huntington's disease	3.36E-05
Leishmaniasis	0.000326241
Phagosome	0.000667716
Cell cycle	0.001153498
Oocyte meiosis	0.001680805
Cardiac muscle contraction	0.001680805
Toll-like receptor signaling pathway	0.001875685
PPAR signaling pathway	0.001875685
Chemokine signaling pathway	0.015452048
Lysosome	0.021129303
B cell receptor signaling pathway	0.025295923
Systemic lupus erythematosus	0.029207716
Complement and coagulation cascades	0.0342643
Cytokine-cytokine receptor interaction	0.034618019
Chagas disease	0.046607705
Progesterone-mediated oocyte maturation	0.053020989
Fc epsilon RI signaling pathway	0.05480725
Leukocyte transendothelial migration	0.05480725

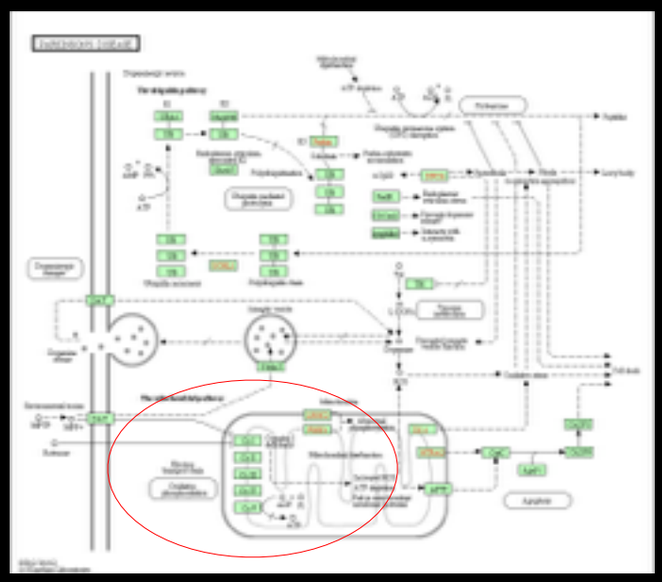
Before eliminating cross-talk

Case Study 1: Fat Remodeling Experiment in Obese Mice (day 3 vs. day 0)

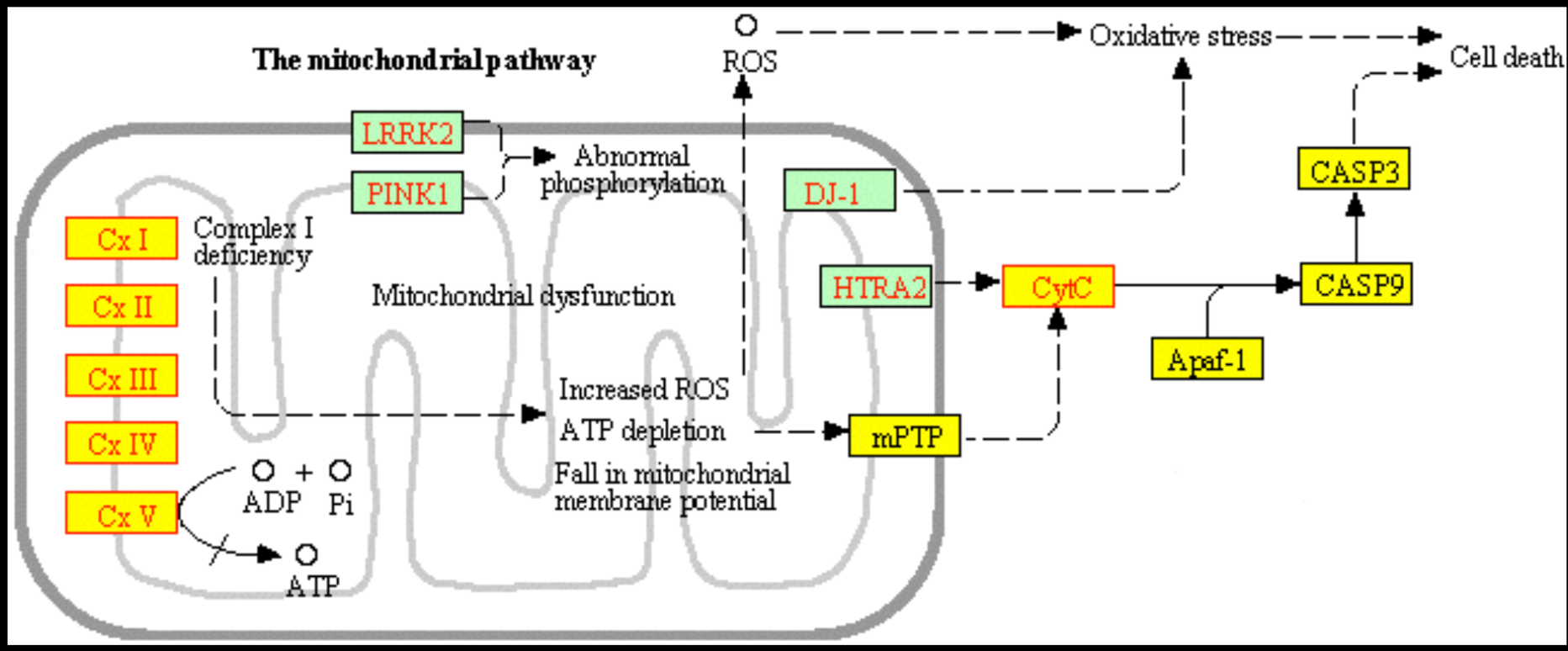








Independent functional module present in the Parkinson's, Huntington's, Alzheimer's, and Cardiac Muscle Contraction pathways



Yellow Background genes belonging to the module

Red Font differentially Expressed genes

Eliminate pathway cross-talk

Parkinson's disease	2.05E-06
Alzheimer's disease	3.57E-06
Huntington's disease	3.36E-05
Leishmaniasis	0.000326241
Phagosome	0.000667716
Cell cycle	0.001153498
Oocyte meiosis	0.001680805
Cardiac muscle contraction	0.001680805
Toll-like receptor signaling pathway	0.001875685
PPAR signaling pathway	0.001875685
Chemokine signaling pathway	0.015452048
Lysosome	0.021129303
B cell receptor signaling pathway	0.025295923
Systemic lupus erythematosus	0.029207716
Complement and coagulation cascades	0.0342643
Cytokine-cytokine receptor interaction	0.034618019
Chagas disease	0.046607705
Progesterone-mediated oocyte maturation	0.053020989
Fc epsilon RI signaling pathway	0.05480725
Leukocyte transendothelial migration	0.05480725

Mitochondria	8.08E-10
Phagosome	9.33E-09
Cell Cycle+Oocyte Meiosis	5.76E-08
PPAR signaling pathway	0.001031617
Compleme+Systemic	0.002154387
* Cytokine-cytokine rec. interaction	0.043021812
Toll-like receptor signaling pathway	0.051196718
MAPK signaling pathway	0.115839005
B cell receptor signaling pathway	0.145781479
Lysosome	0.187446378
Natural killer cell mediated cytotox.	0.187446378
* Cell cycle	0.229087797
Calcium signaling pathway	0.229087797
Cell adhesion molecules (CAMs)	0.2583244
NOD-like receptor signaling pathway	0.2583244
Vascular smooth muscle contraction	0.424980819
Dilated cardiomyopathy	0.424980819
* Oocyte meiosis	0.432407943
Type I diabetes mellitus	0.432407943

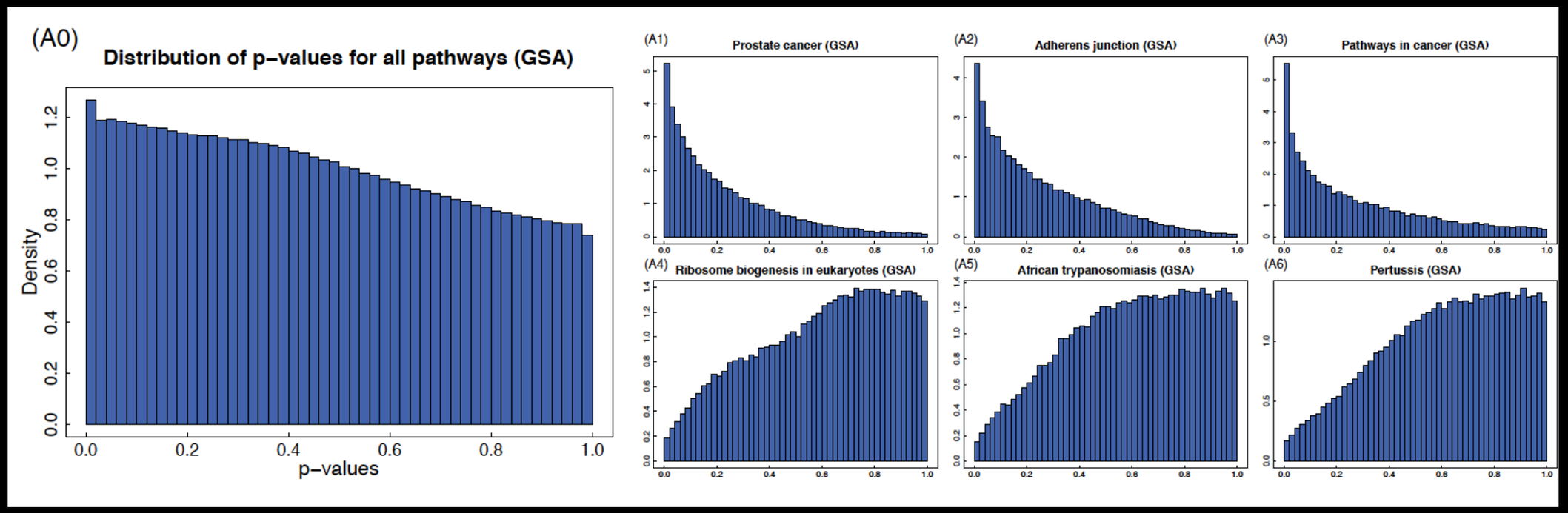
Before eliminating

After eliminating

- Donato et al, Genome Research 23 (11), 1885-1893, 2013

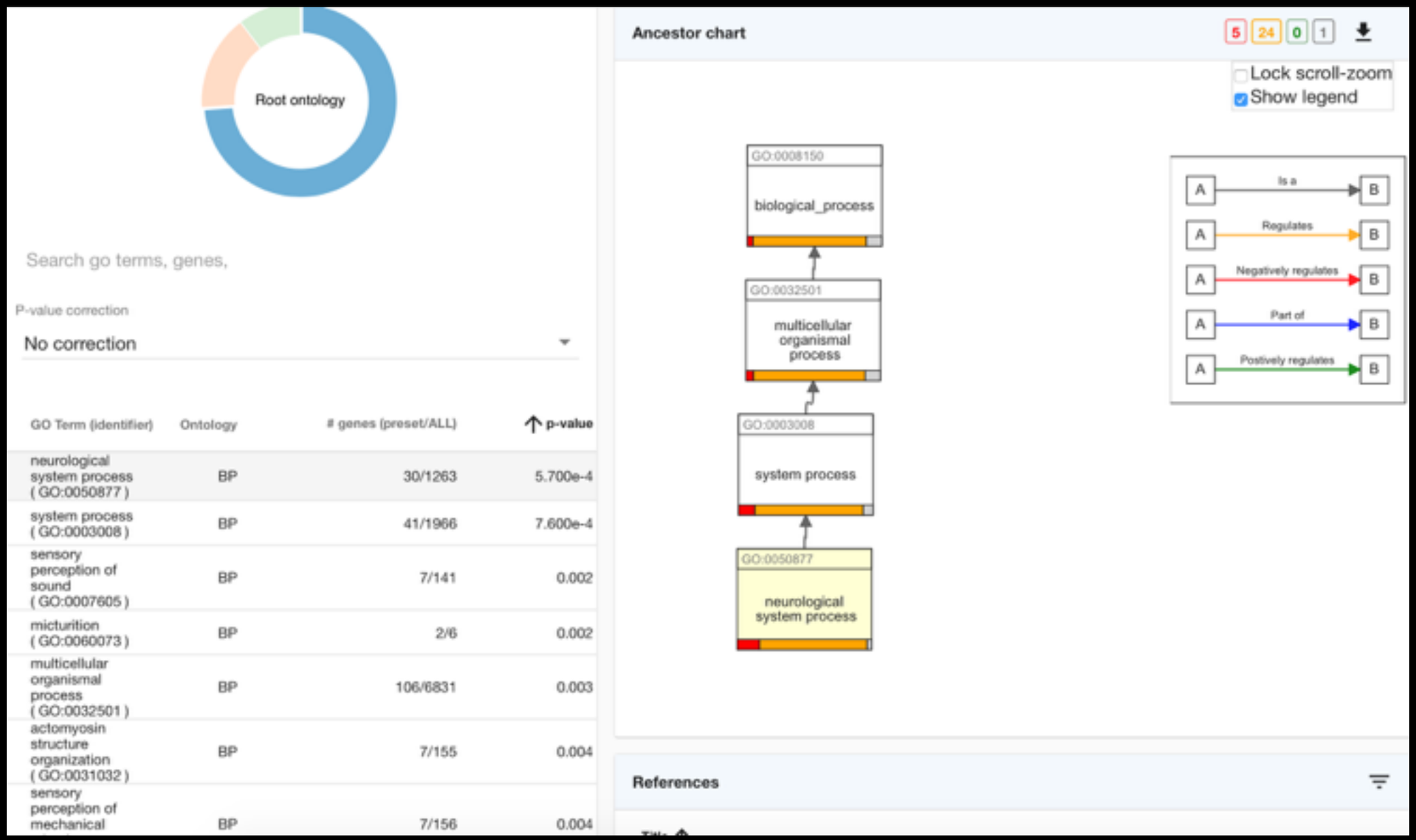
Individual pathway bias

Null distributions (GSA)

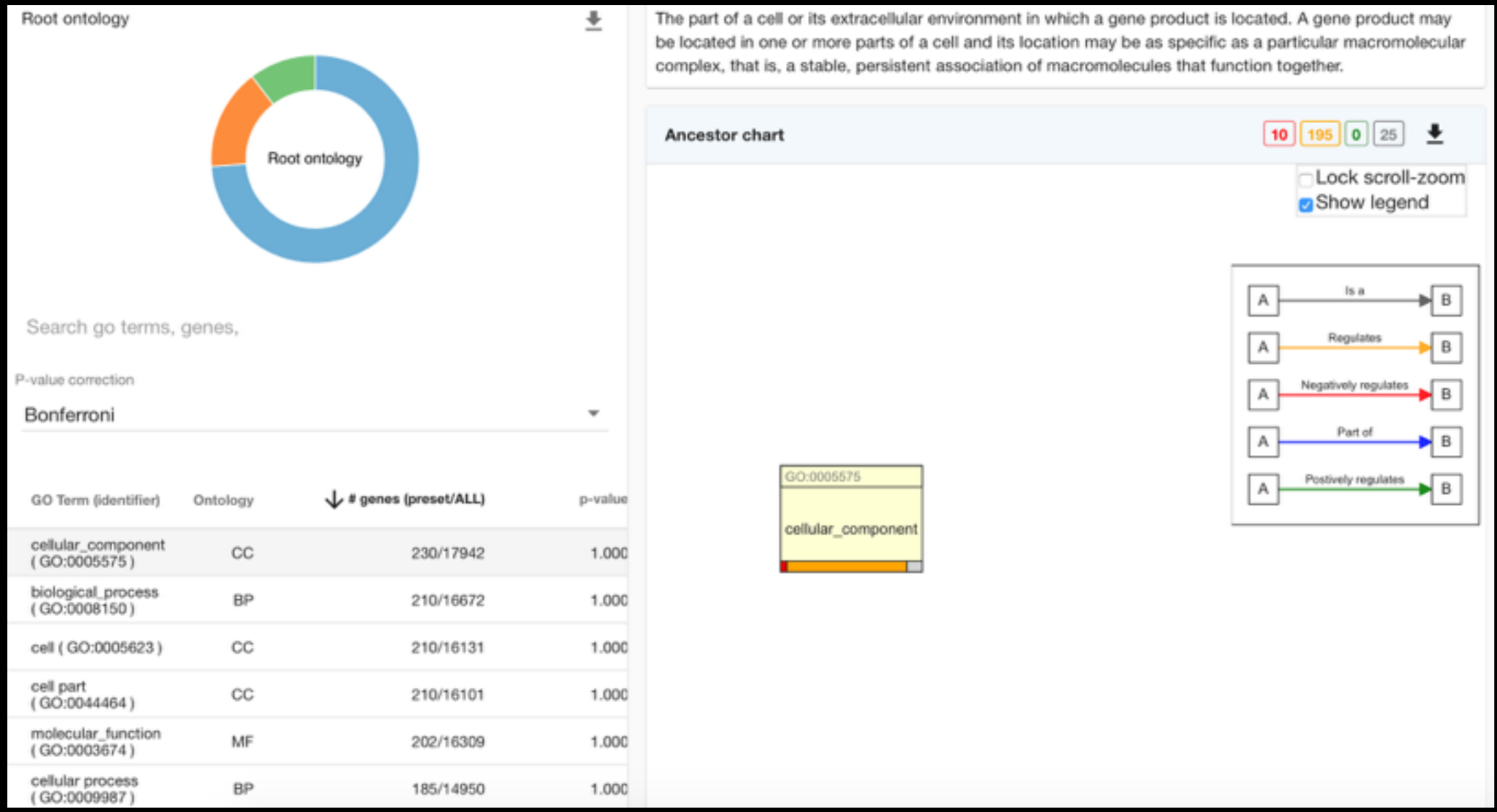


T Nguyen, C Mitrea, R Tagett, S Draghici, Proceedings of the IEEE 105 (3), 496-515

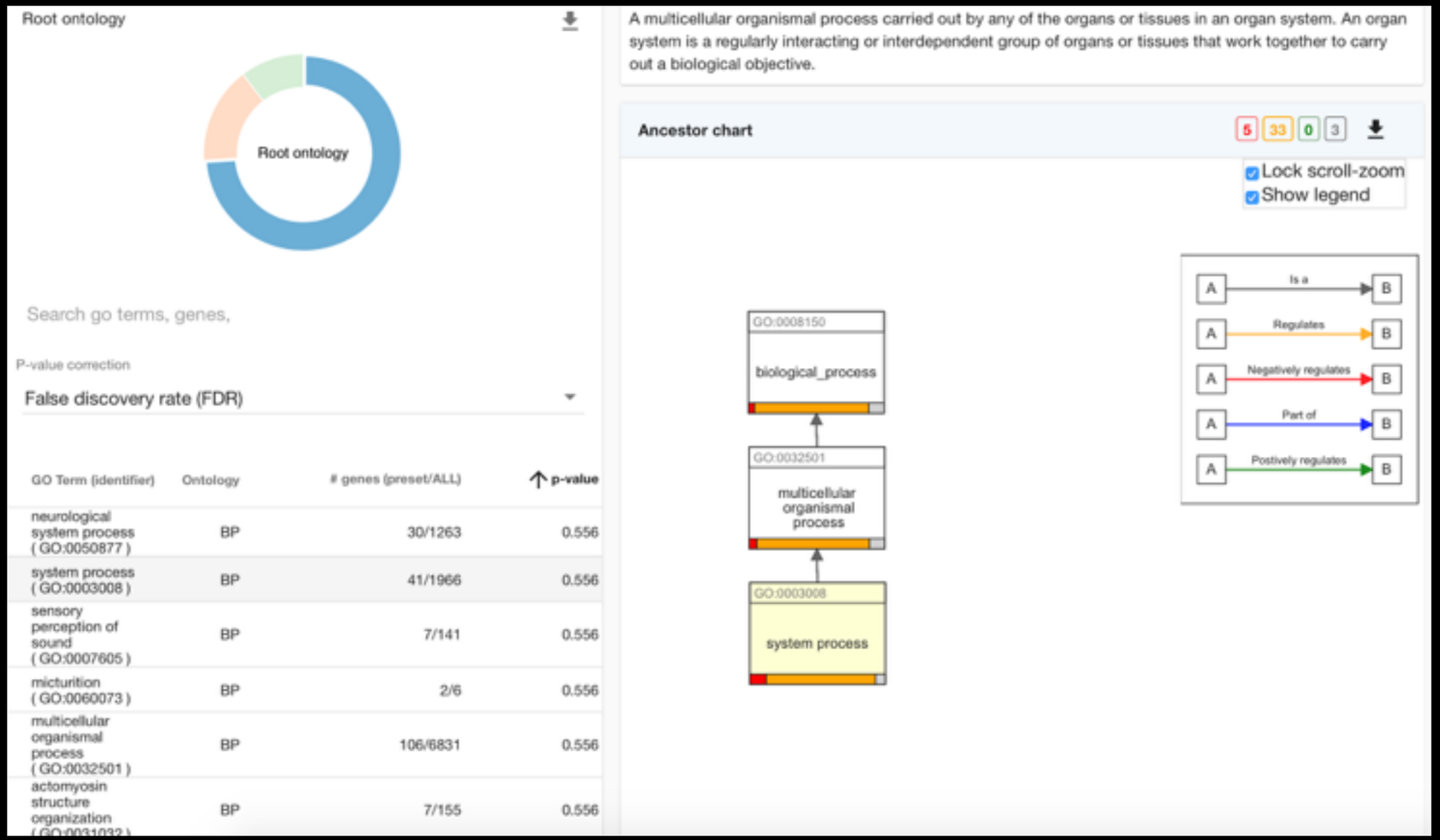
GO term dependencies and overlap



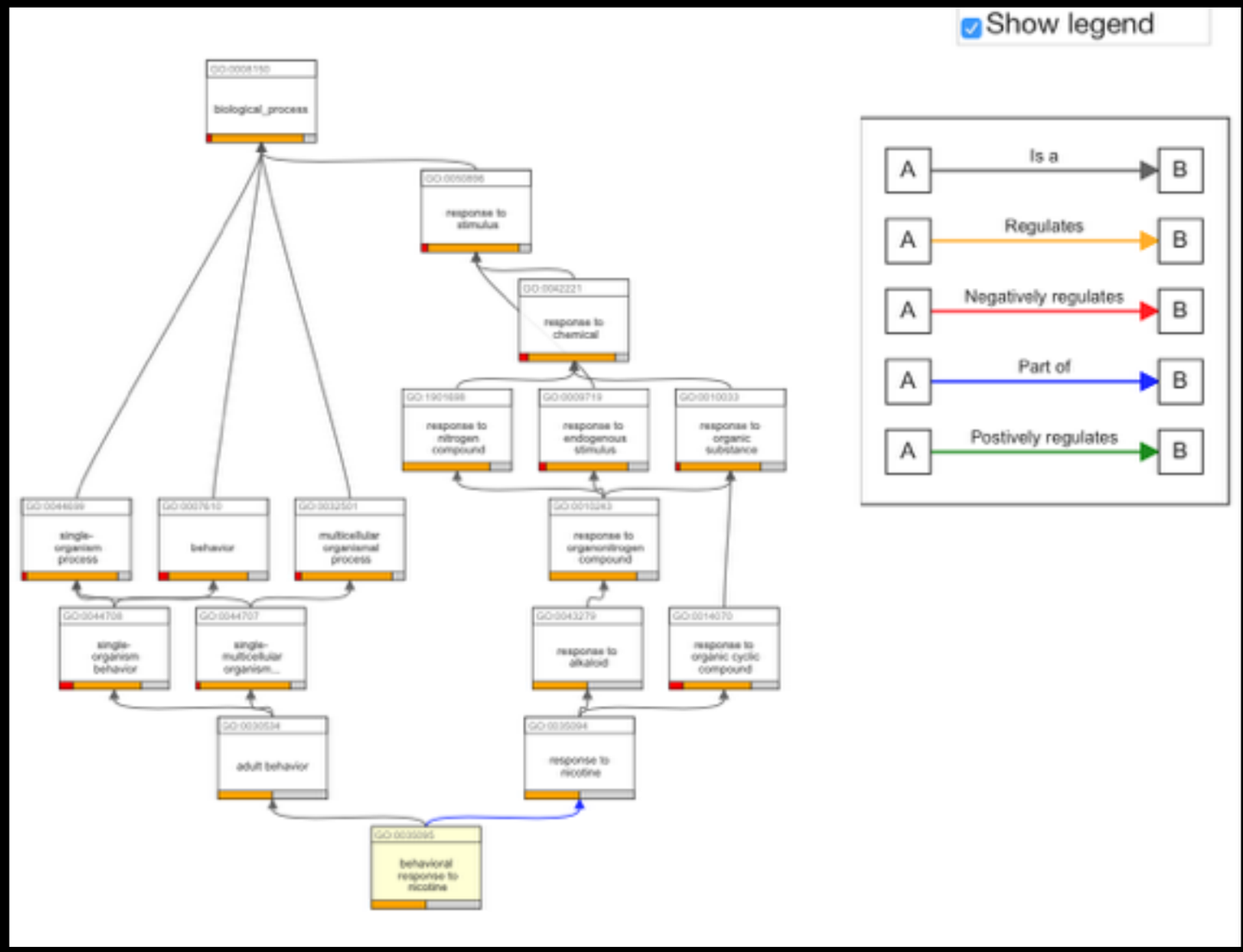
GO term dependencies and overlap



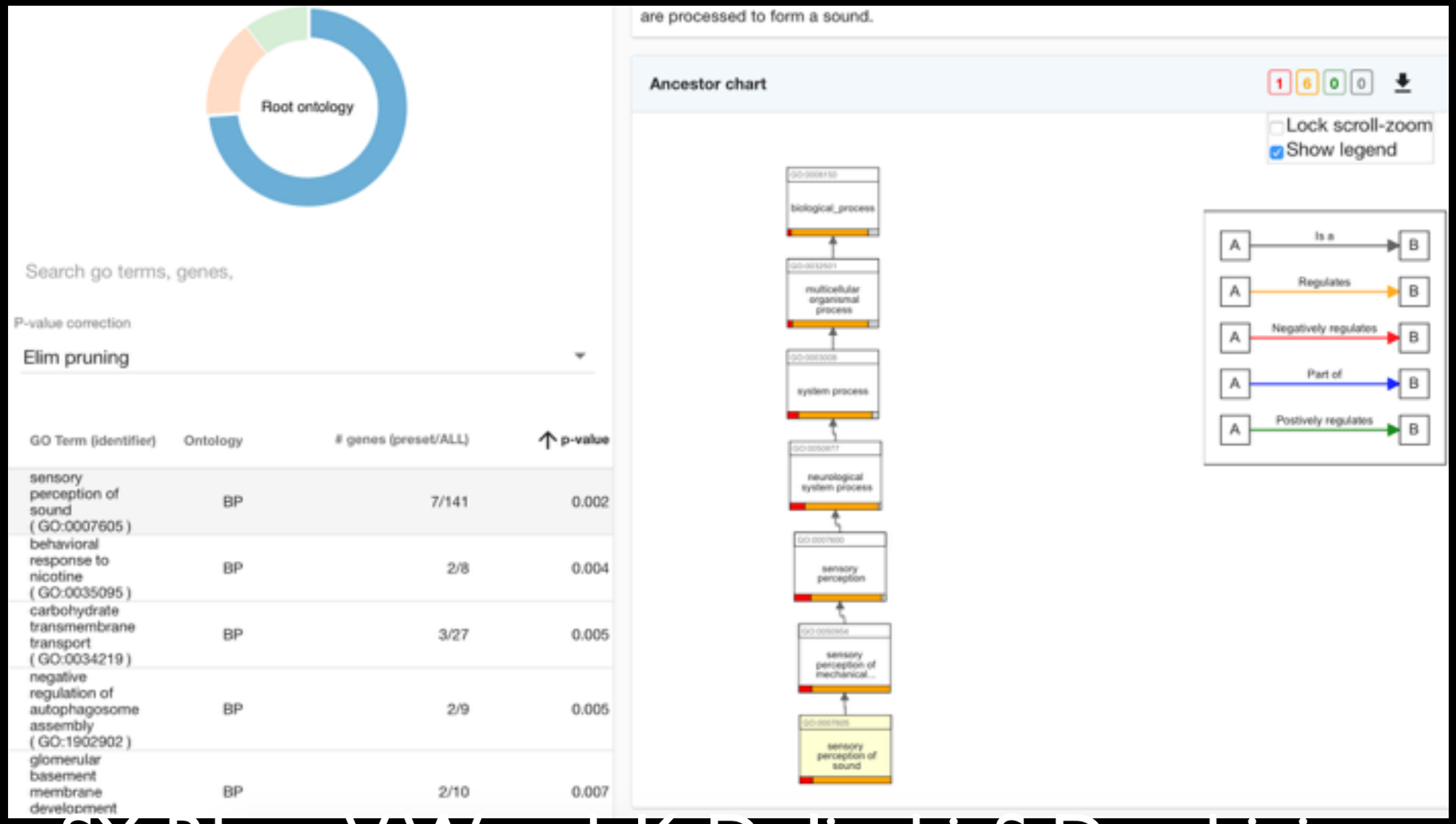
GO term dependencies and overlap



The Gene Ontology true path rule



GO term dependencies and overlap



SY Rhee, V Wood, K Dolinski, S Draghici
Nature Reviews Genetics 9 (7), 509-515, 2008

Recommendations for the

- Use all knowledge available i.e. use pathways, not gene sets if possible (do consider the signals and interactions between genes - they are important!!)

- Use methods that can assess pathway impact based on the topology and calculate significance based on resampling (e.g. impact analysis), not simple enrichment
  - Draghici et al, Genome Research 17 (10), 1537-1545

- Use methods that can identify putative mechanisms based on known pathway topology (you have just seen one)

- Take into consideration and eliminate individual pathway bias
  - Nguyen et al, Proceedings of the IEEE 105 (3), 496-515

Thank you!

Sorin Draghici

sorin@wayne.edu