A dynamic programing approach to integrate gene expression data and network information for pathway model generation

Yuexu Jiang, Yanchun Liang, Duolin Wang, Dong Xu, and Trupti Joshi

Bioinformatics 2020

Students: Mahdi Rostami ,Mahdi Jamshid por

Why active pathways?

- ✓ Disease
- ✓ Drug design
- ✓ Lung cancer
- ✓ ….

Why dynamic programming?

- ✓ Step by step
- ✓ Trace back
- ✓ Economical system

Previous works

Paper	Explanation
Tuncbag et al. (2013)	Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J. Comput. Biol., 20, 124–136.
Ideker,T. et al. (2001)	Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 292, 929–934.
Kim et al., (2011)	Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput. Biol., 7, e1001095.
Lan et al. (2011)	ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res., 39, W424–W429.
Min et al. (2018)	Edge-group sparse PCA for network-guided high dimensional data analysis. Bioinformatics, 34, 3479–3487.

Limitations

Previous works problems

- Unclear pathway
- Hard to trace back
- Time-consuming
- information loss
- Related pathways

Input data

- KEGG

  Kyoto Encyclopedia of Genes and Genomes

- PPI

  protein-protein interaction

- GEO

  Gene Expression Omnibus

- Omics data

MAPK signaling pathway - yeast - Reference pathway

Pathway menu | Pathway entry | Show description | Image (png) file | Help]

Change pathway type

Option

Scale: 100%

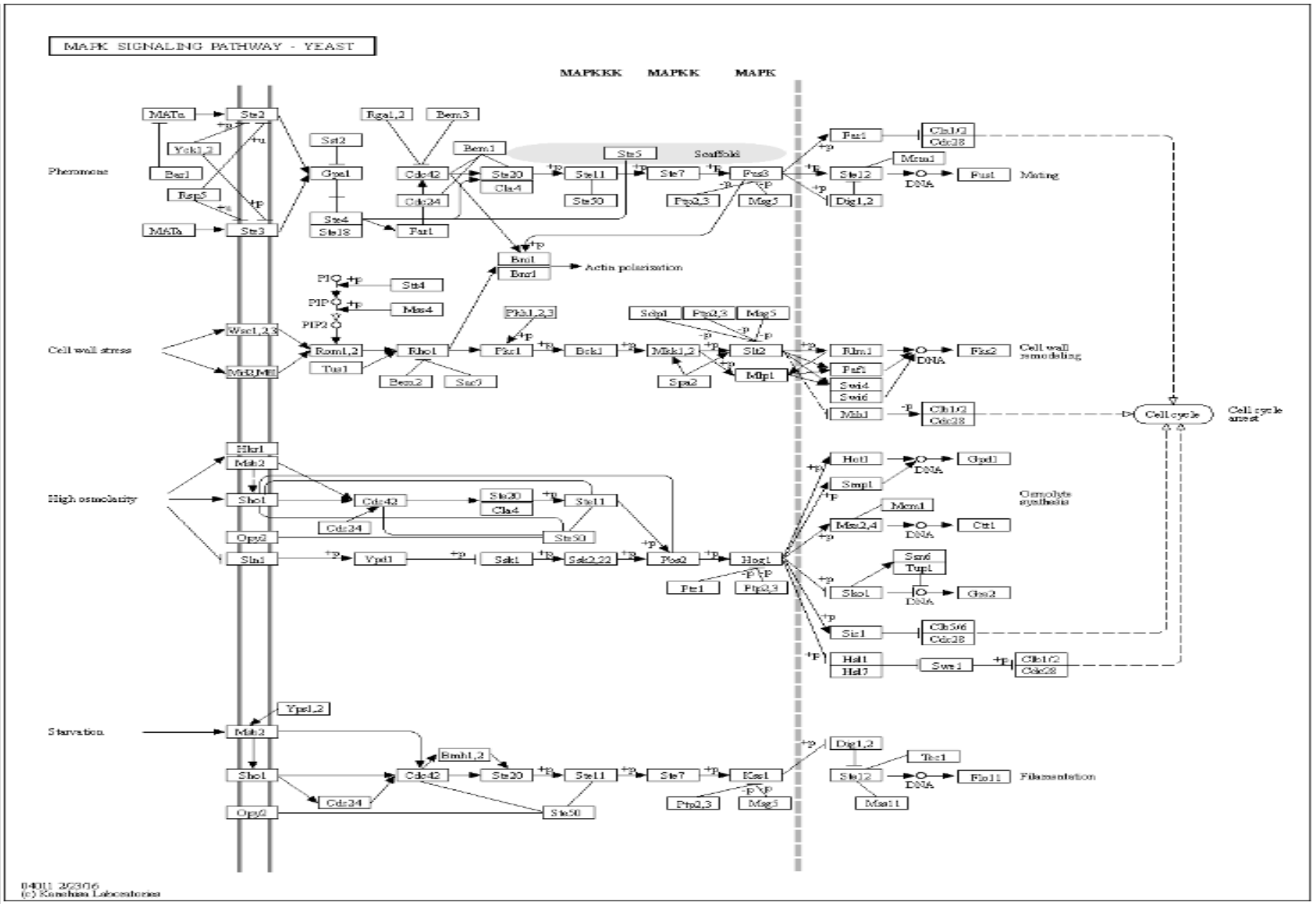
Search

Go

ID search

Go

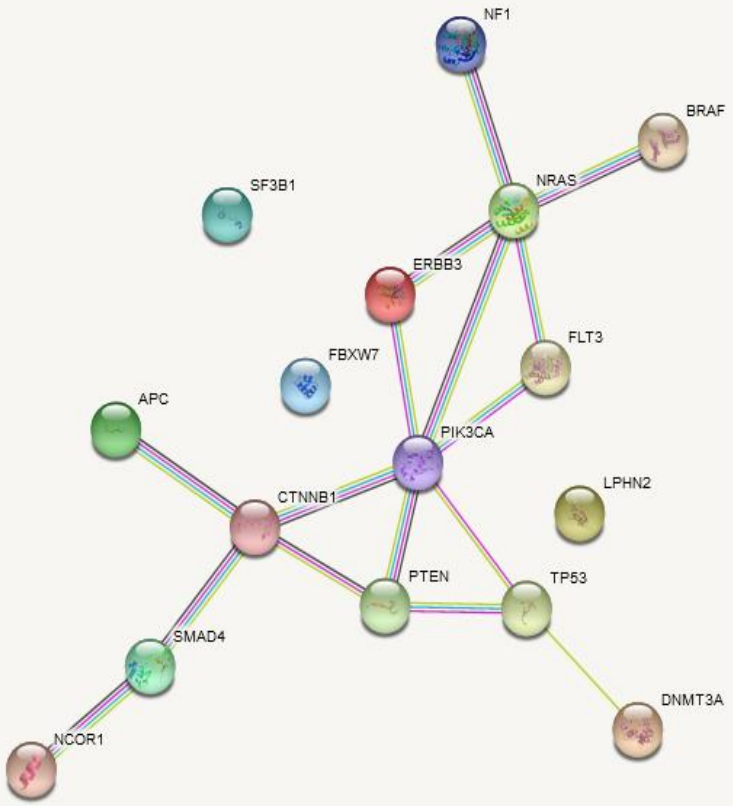
Color



Version: 11.5 LOGIN | REGISTER

STRING

Search Download Help My Data



Viewers | Legend | Settings | Analysis | Exports | Clusters | More | Less

Network currently showing
Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

Cooccurrence
Gene families whose occurrence patterns across genomes show similarities.

Experiments
Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

Coexpression
Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

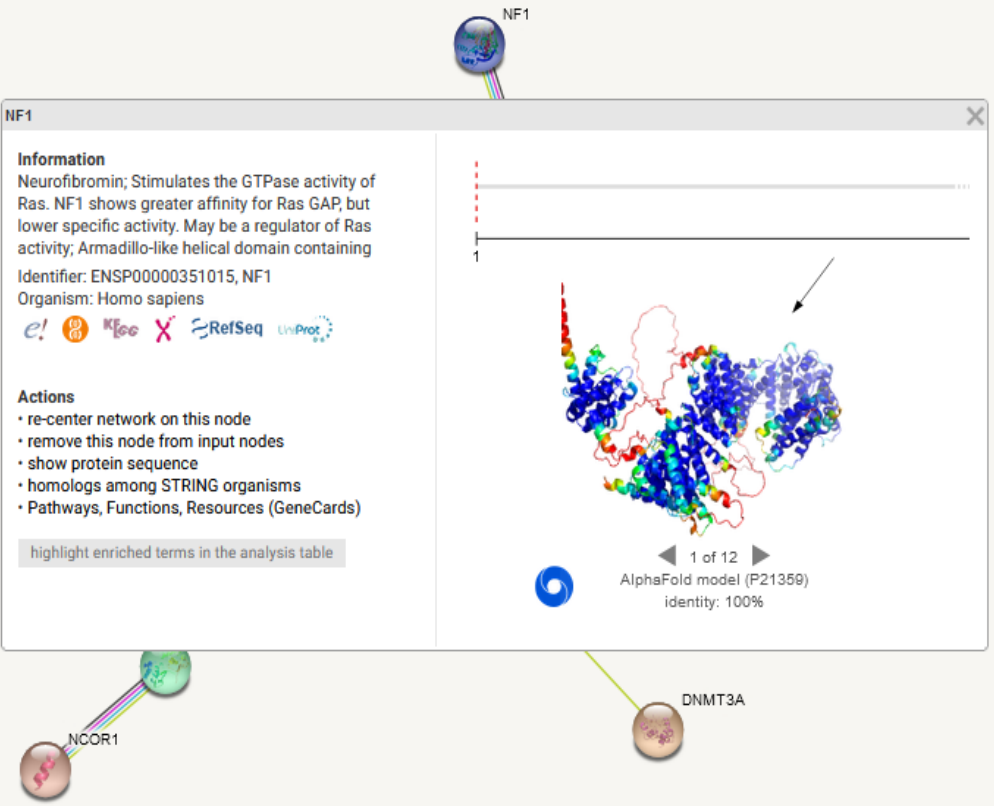
Databases
Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

Neighborhood
Groups of genes that are frequently observed in each other's genomic neighborhood.

Version: 11.5 LOGIN | REGISTER

STRING

Search Download Help My Data



Viewers | Legend | Settings | Analysis | Exports | Clusters | More | Less

Network currently showing
Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

Cooccurrence
Gene families whose occurrence patterns across genomes show similarities.

Experiments
Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

Coexpression
Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

Databases
Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

Neighborhood
Groups of genes that are frequently observed in each other's genomic neighborhood.

GEO

GEO	Title
GSE19804	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan
GSE31176	Expression data from yeast (wild type, rlm1 and swi3 mutants) exposed to Congo Red
GSE71433	Expression data from yeast (wild type and gcn5 mutants) exposed to Congo Red (CR)
GSE13097	mRNA amount analysis of wild type strain subjected to osmotic stress

NCBI > GEO > Accession Display Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE19804 GO

Series GSE19804 Query DataSets for GSE19804

Status	Public on Jan 30, 2011
Title	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Although smoking is the major risk factor for lung cancer, only 7% of female lung cancer patients in Taiwan have a history of cigarette smoking, extremely lower than those in Caucasian females. This report is a comprehensive analysis of the molecular signature of non-smoking female lung cancer in Taiwan.
Overall design	RNA was extracted from paired tumor and normal tissues for gene expression analysis.
Contributor(s)	Lu T, Lai L, Chuang EY
Citation(s)	Lu TP, Tsai MH, Lee JM, Hsu CP et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev* 2010 Oct;19(10):2590-7. PMID: 20802022 Lu TP, Hsiao CK, Lai LC, Tsai MH et al. Identification of regulatory SNPs associated with genetic modifications in lung adenocarcinoma. *BMC Res Notes* 2015 Mar 24;8:92. PMID: 25889623
Submission date	Jan 08, 2010
Last update date	Jan 15, 2020
Contact name	Tzu-Pin Lu
E-mail(s)	tplu@ntu.edu.tw
Organization name	National Taiwan University, Taiwan
Department	Department of Public Health, Institute of Epidemiology and Preventive Medicine
Street address	No. 1, Sec. 4, Roosevelt Road
City	Taipei
ZIP/Postal code	10617
Country	Taiwan
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (120) More...	GSM494556 Lung Cancer 2T GSM494557 Lung Cancer 3T GSM494558 Lung Cancer 6T

This SubSeries is part of SuperSeries:

GSE33356 Genome-wide screening of genomic alterations and transcriptional modulation in non-smoking female lung cancer in Taiwan

Relations

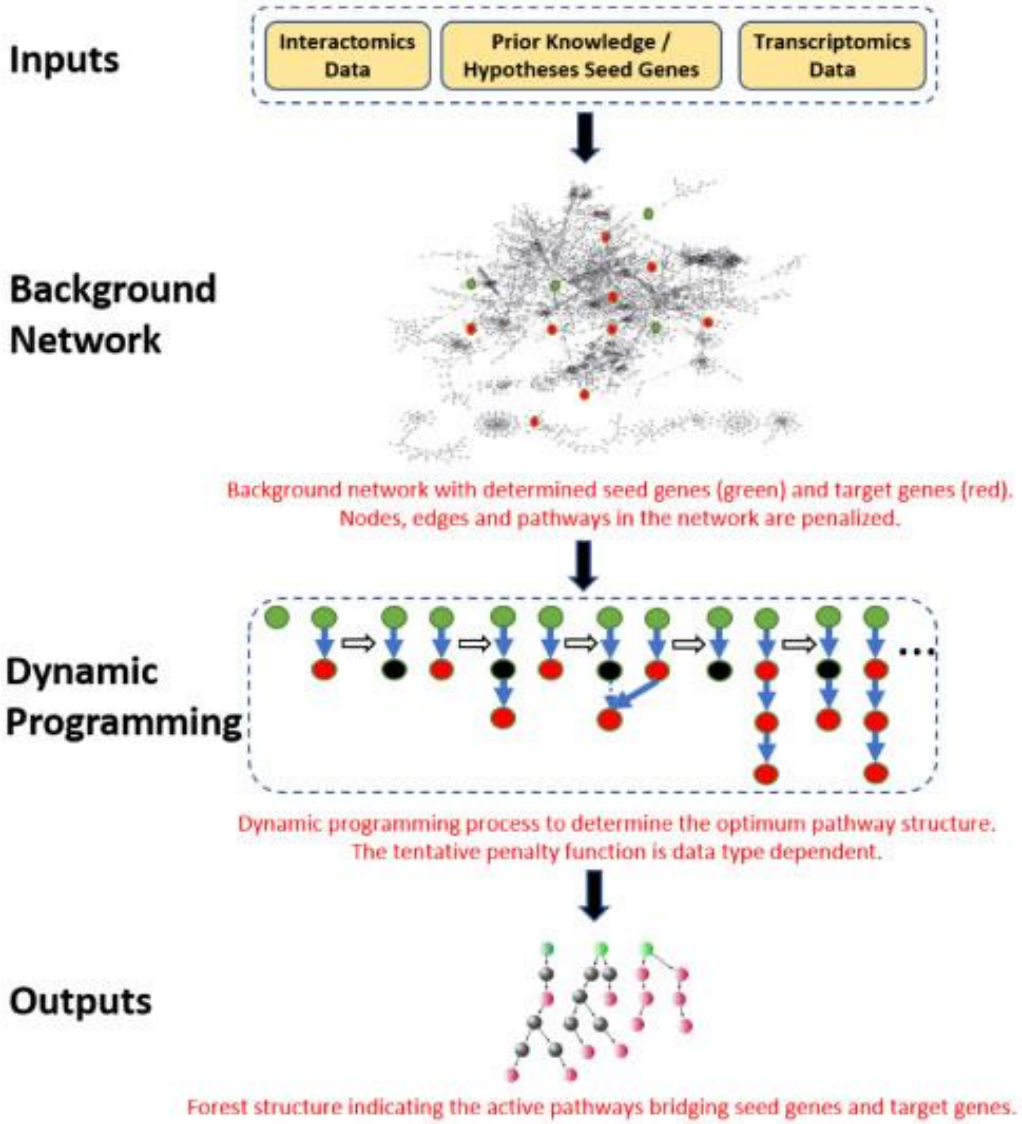
BioProject PRJNA153899

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML

Method

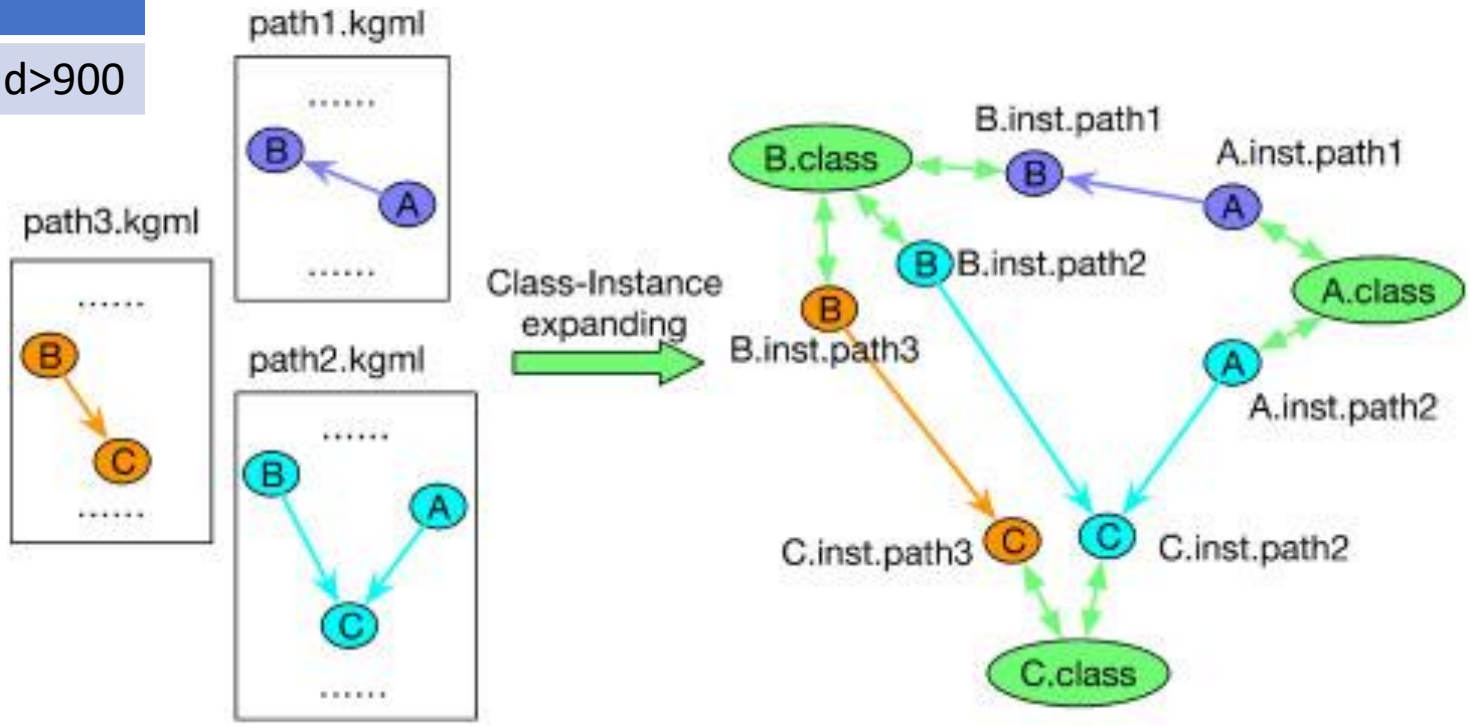
1. Constructing a merged pathway network
2. Define seed genes and target genes
3. Assign penalties based on omics data
4. Calculating minimum involvement score using a shortest path algorithm
5. Detecting final active pathways by truncating and backtracking



Method

Constructing a merged pathway network

1	KEGG		
2	PPI	STRING v10	Threshold>900



Method

Define seed genes and target genes

Seed genes	Perspective
Root nodes	Algorithm
Mutated	Biological
Receptor genes	Biological
Effective	Natural potential

Target genes	
Case-control	q-value threshold of 0.01
time-series	maSigPro

Method

Assign penalties based on omics data

Static penalty:

	Case control	Time series	Equation
Node penalty	+		1
Edge penalty		+	2
Pathway penalty	+	+	3

1. t-test -> P-value

$$\text{penalty}(v) = \frac{1}{-\log2(P_{value}(v))} \quad (1)$$

2. v1, v2 = vectors

$$\text{penalty}(e) = 1 - |\text{correlation}(v1,\ v2)| \quad (2)$$

3. Fisher test -> P-value

$$\text{penalty}(p) = \frac{1}{-\log2(P_{value}(p))} \quad (3)$$

4. Unknown pathway = 1

Method

Calculating minimum involvement score using the shortest path algorithm (Dijkstra)

Dynamic penalty:

$$\text{tentative penalty}(s, v) = \begin{cases} penalty(s) + Sigmoid\Big(fc(s) \cdot fc(v)\Big) \cdot \Big(penalty(v) + penalty(p_v)\Big), & negative\ relation \\ penalty(s) + Sigmoid\Big(-fc(s) \cdot fc(v)\Big) \cdot \Big(penalty(v) + penalty(p_v)\Big), & positive\ relation \end{cases}$$

Case control	
fc	fold
Pv	pathway
relation	KEGG

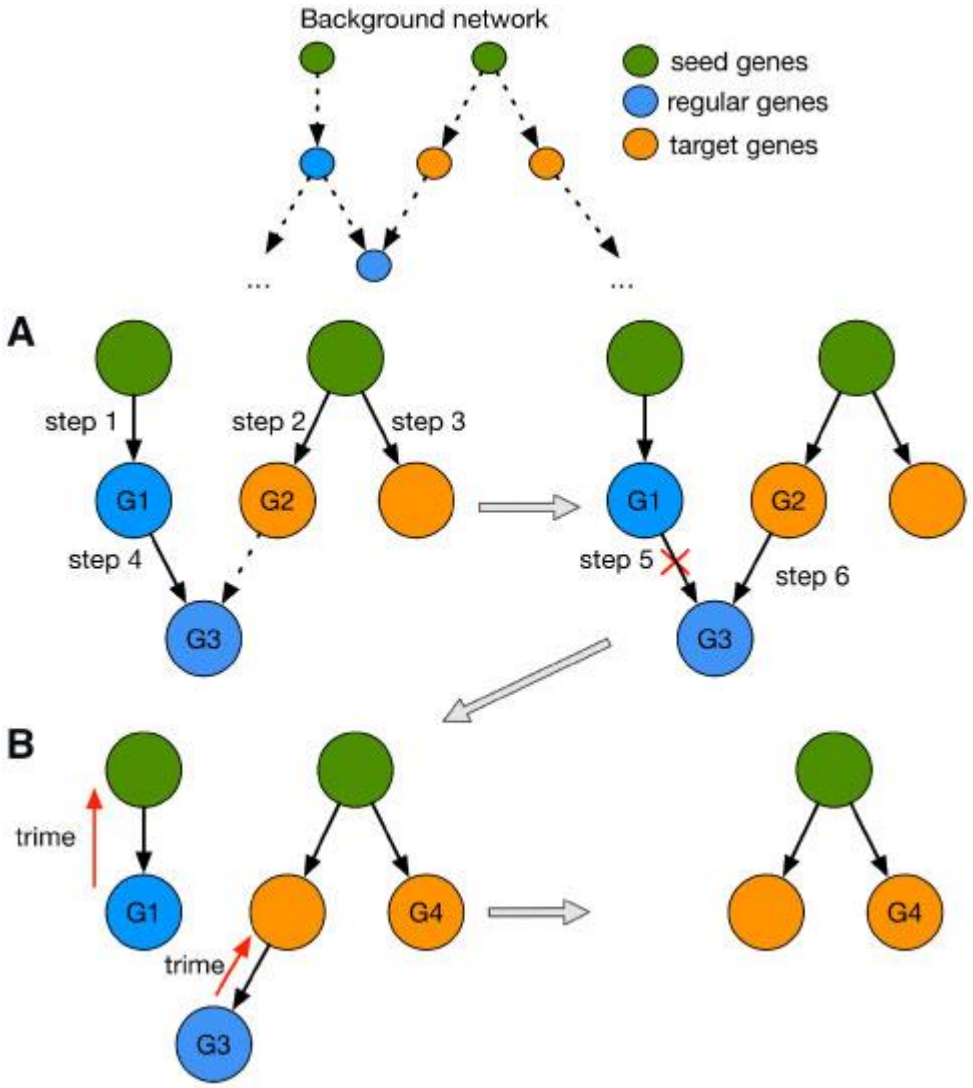
Time series	
e	Interaction (KEGG ,PPI)
Pv	pathway

$$\text{tentative penalty}(s, v) = \text{penalty}(s) + \text{penalty}(e) \cdot \text{penalty}(p_v)$$

Method

Detecting final active pathways by truncating and backtracking

A	**Penalty update**
B	Truncating rules

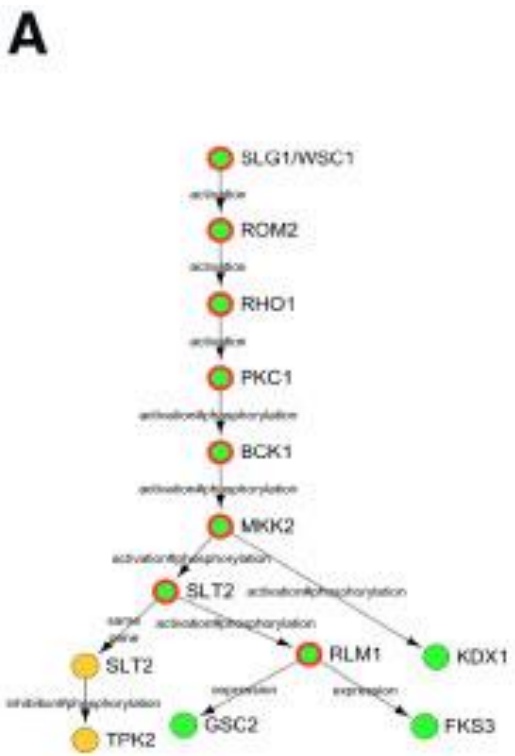


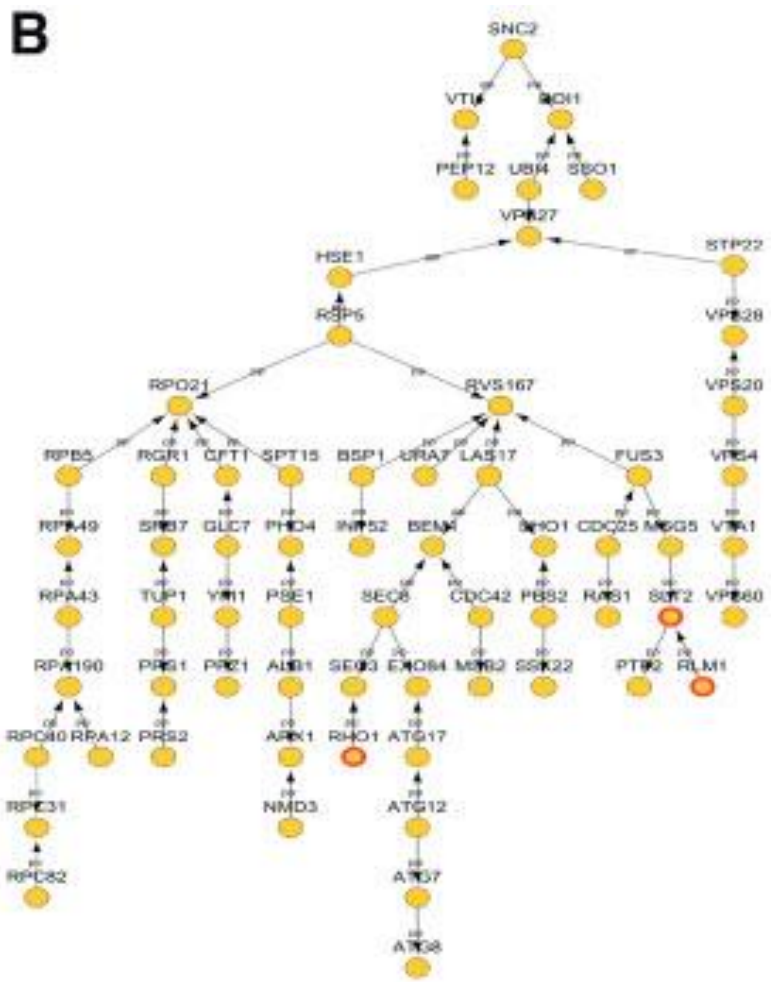
Evaluation

- Application on yeast cell wall damage stress dataset

Method	Shape
Steiner tree	B
IMPRes	A

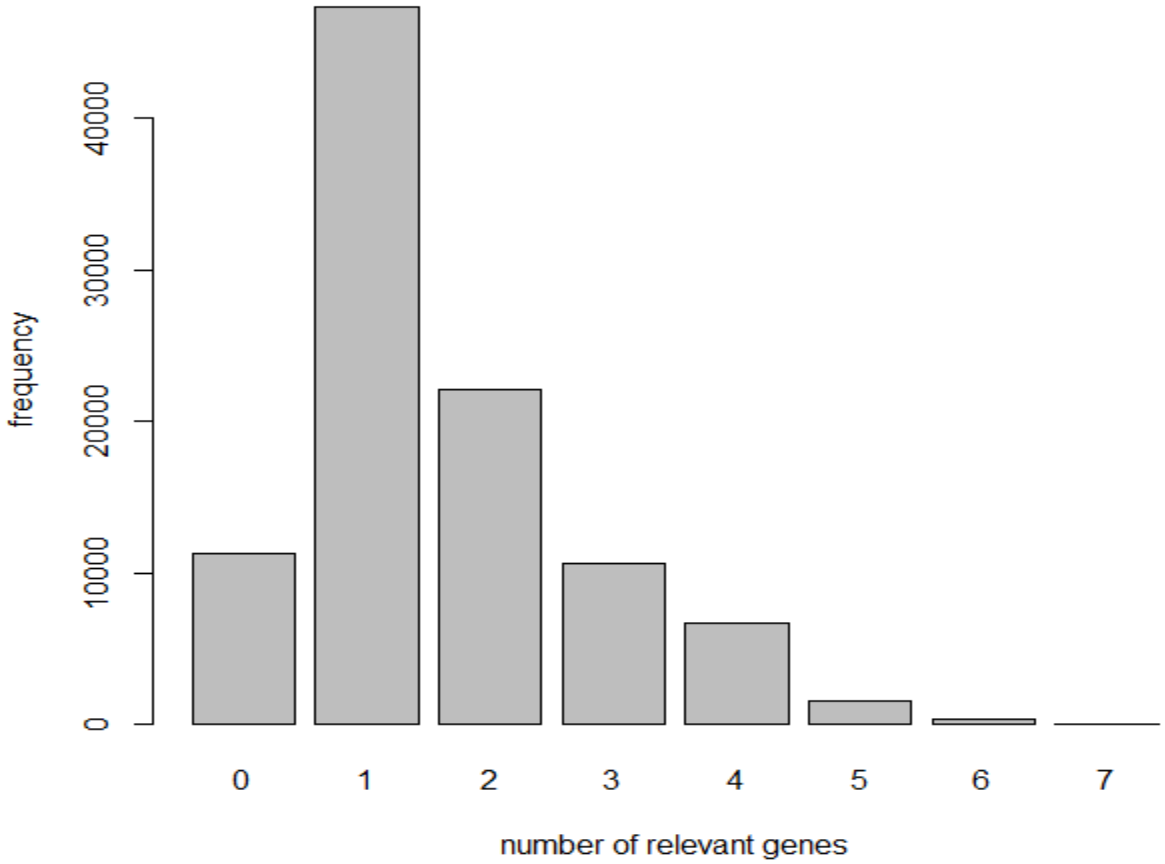
GEO	GSE31176 GSE71433





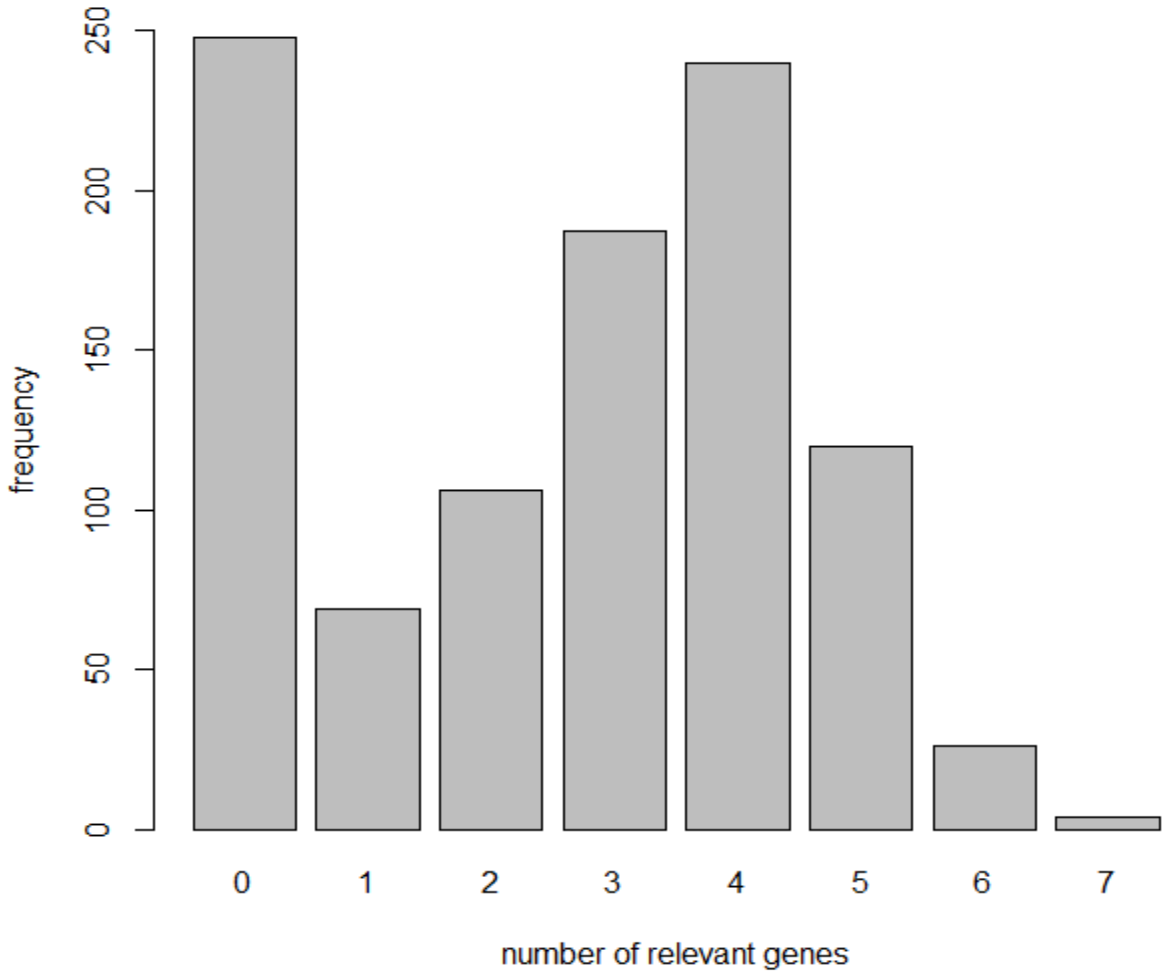
Evaluation

Is it by chance?

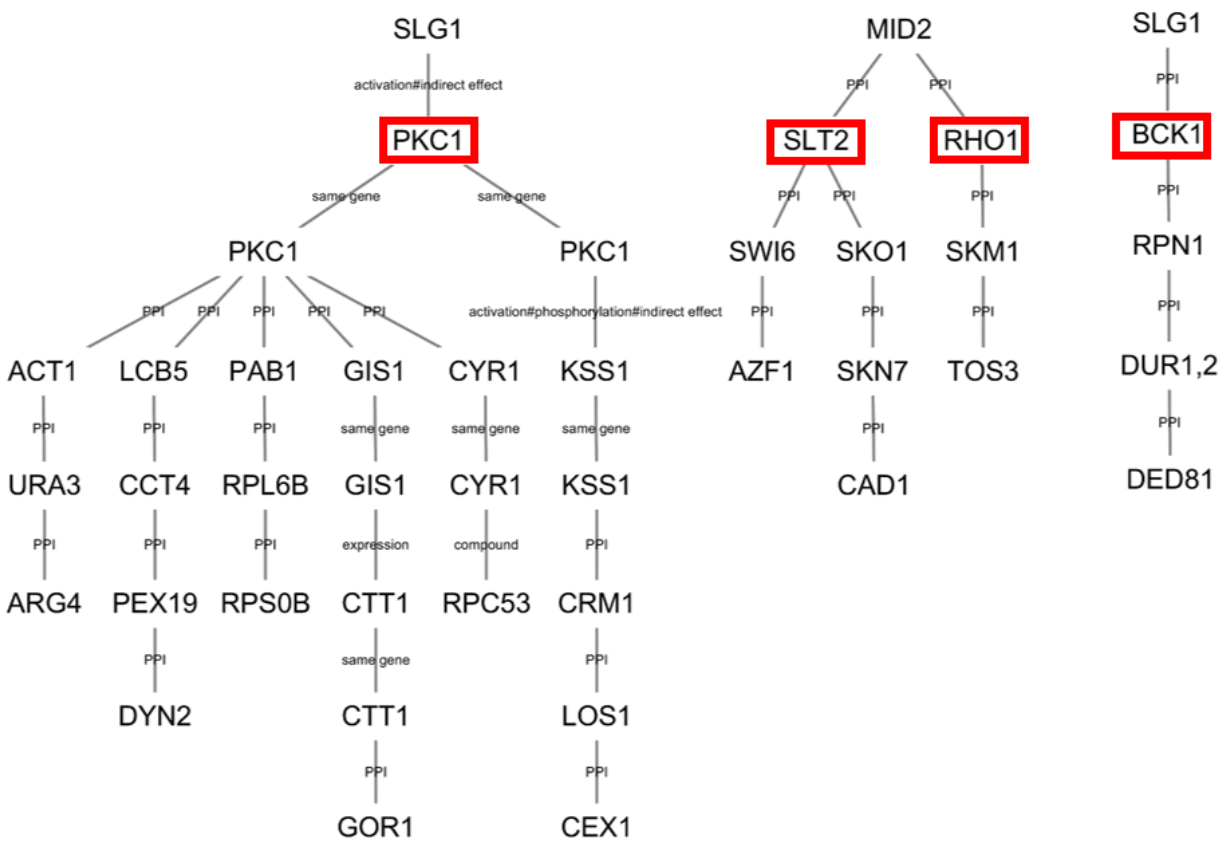


Evaluation

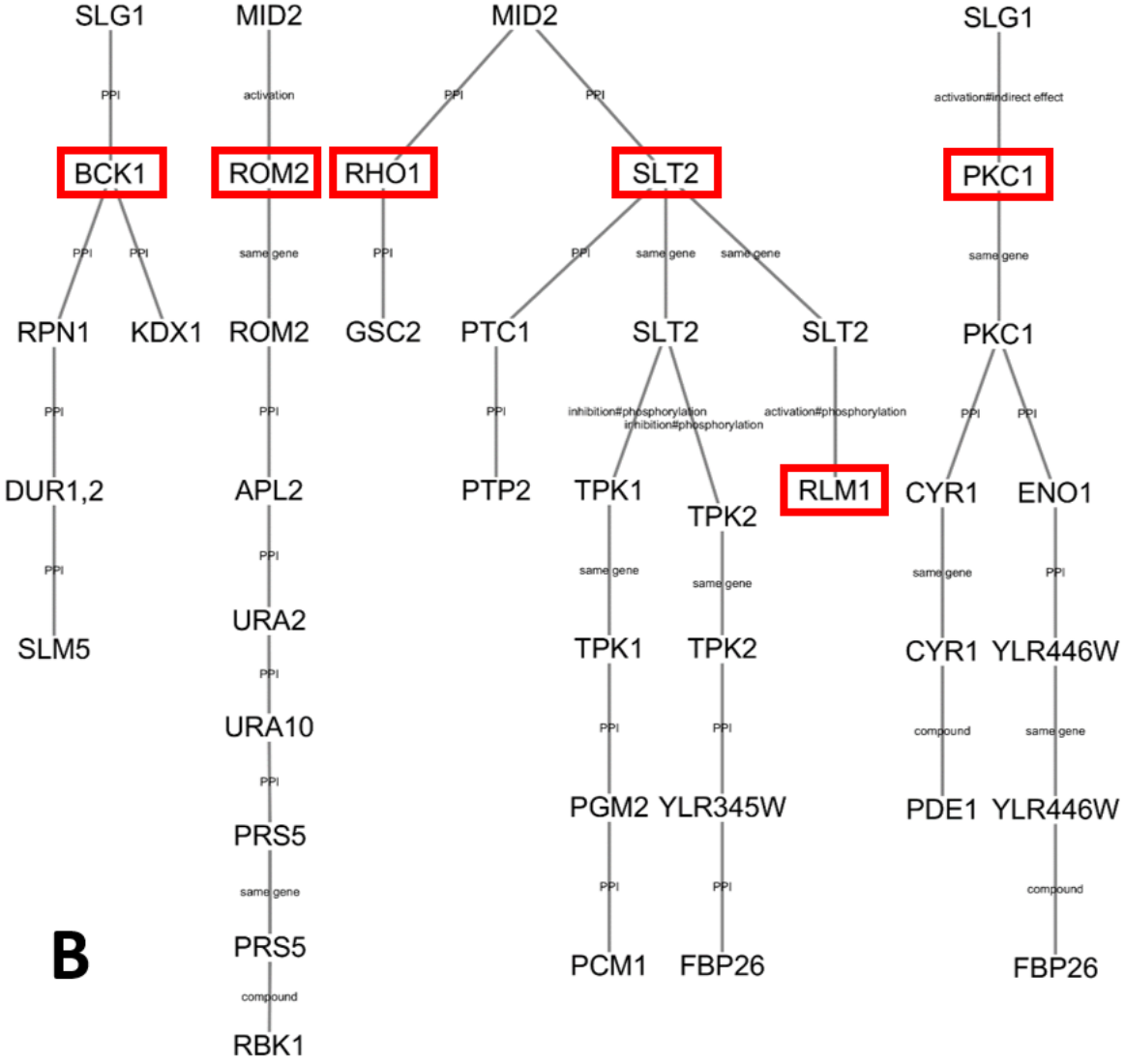
Does expression data have an impact?



Evaluation



A



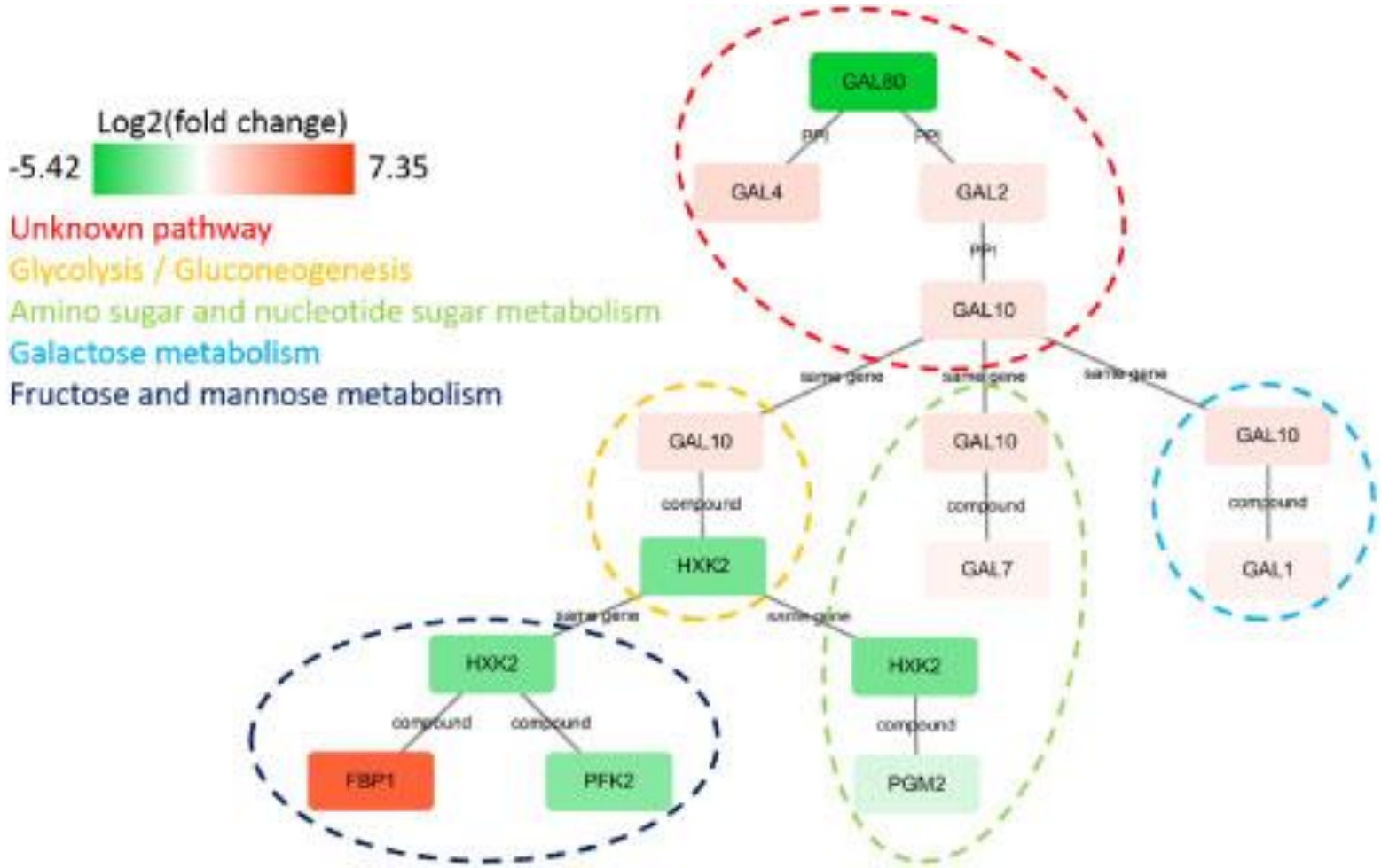
B

Evaluation

- Application on yeast perturbation dataset

Method	Small network	Large network	Network PPI
Ideker et al.(2002)	43 nodes	340 genes	362
IMPRes	10 786 nodes		54 249

Evaluation


Log2(fold change)
-5.42
7.35
Unknown pathway
Glycolysis / Gluconeogenesis
Amino sugar and nucleotide sugar metabolism
Galactose metabolism
Fructose and mannose metabolism
GAL80
GAL4
GAL2
GAL10
GAL10
GAL10
GAL10
HXK2
GAL7
GAL1
HXK2
HXK2
FBP1
PFK2
PGM2

Evaluation

- Application on yeast high osmolality stress dataset

GEO	GSE13097

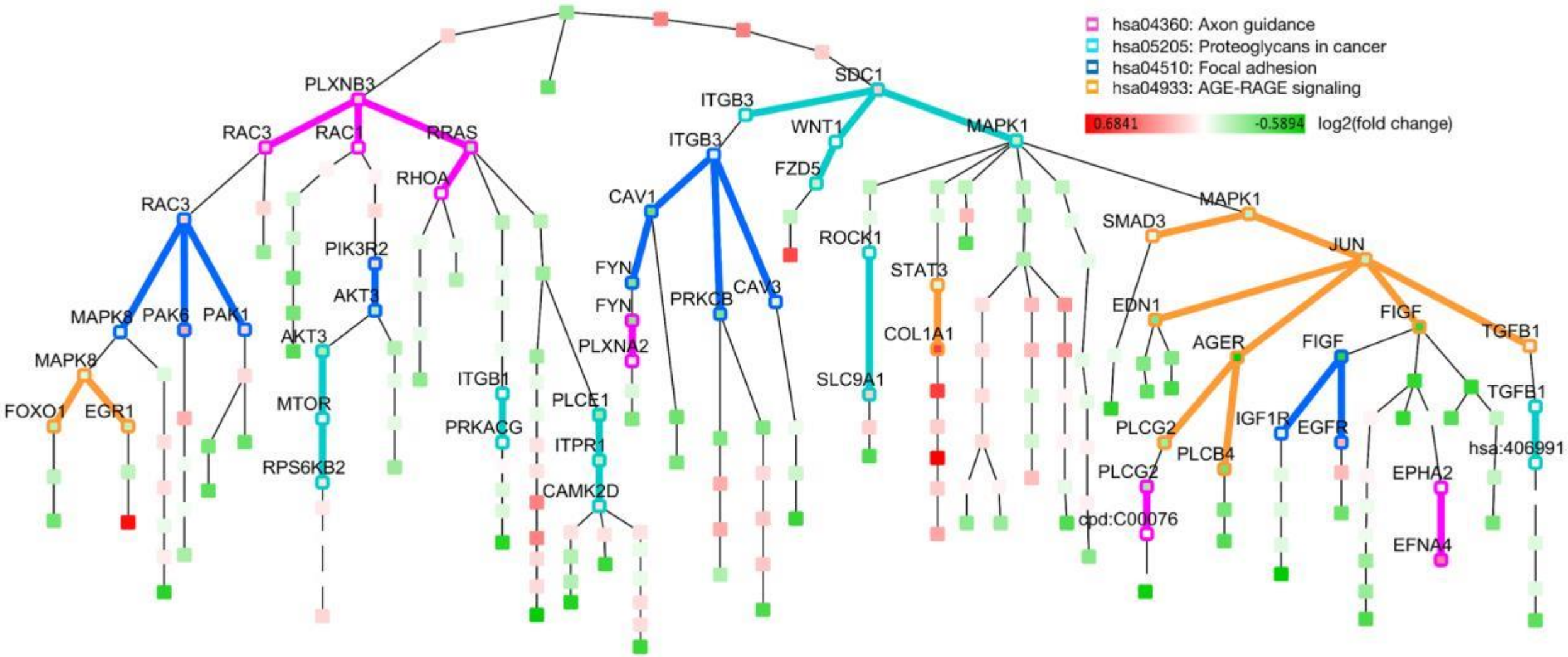
Algorithm	KEGG_ top10	KEGG_ top30	KEGG_ top40	KEGG_ top50	K+PPI_ top10	K+PPI_ top20	K+PPI_ top30	SDREM	PNM	ResponseNet
Total predictions	5	9	12	23	28	44	66	58	445	61
Predicts Hog1	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
Predicted internal	4	6	9	20	25	41	62	30	374	4
Gold standard internal	21	21	21	21	29	29	29	30	30	30
Internal overlap	4	5	5	5	3	7	7	6	9	1
Internal significance	4.5E-8	3.4E-9	6.9E-8	7.7E-6	5.1E-4	4.2E-9	8.5E-8	1.11E-8	1.61E-4	0.0227
Internal precision	100%	83%	55.6%	25%	12%	17%	11%	20%	2%	25%
Predicted TFs	1	3	3	3	3	3	4	28	71	57
Gold standard TFs	6	6	6	6	6	6	6	7	7	7
TF overlap	1	3	3	3	3	3	3	4	2	2
TF significance	0.24	8.7E-3	8.7E-3	8.7E-3	1.4E-4	1.4E-4	5.6E-4	7.7E-3	0.77	0.632
TF precision	100%	100%	100%	100%	100%	100%	75%	14%	2.8%	3%

Result

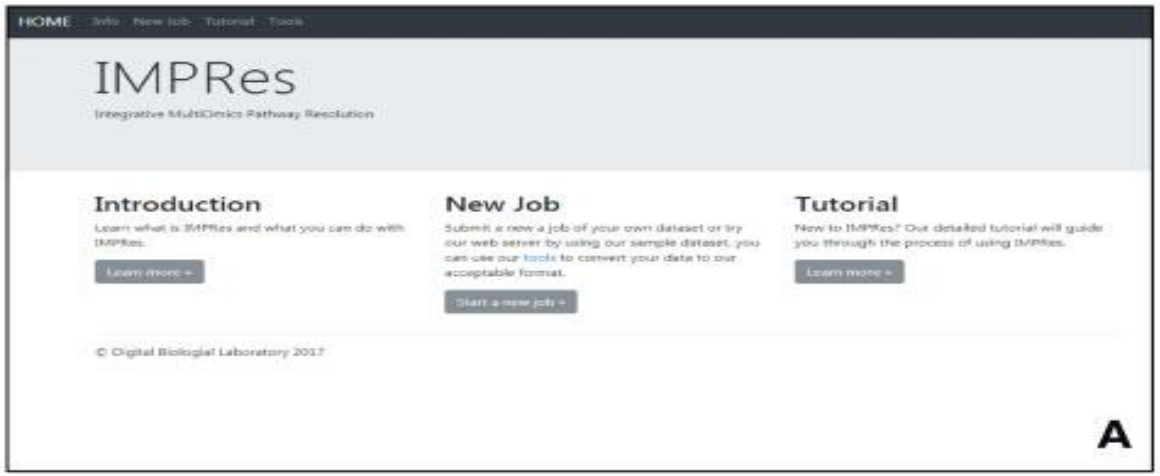
- Human lung cancer

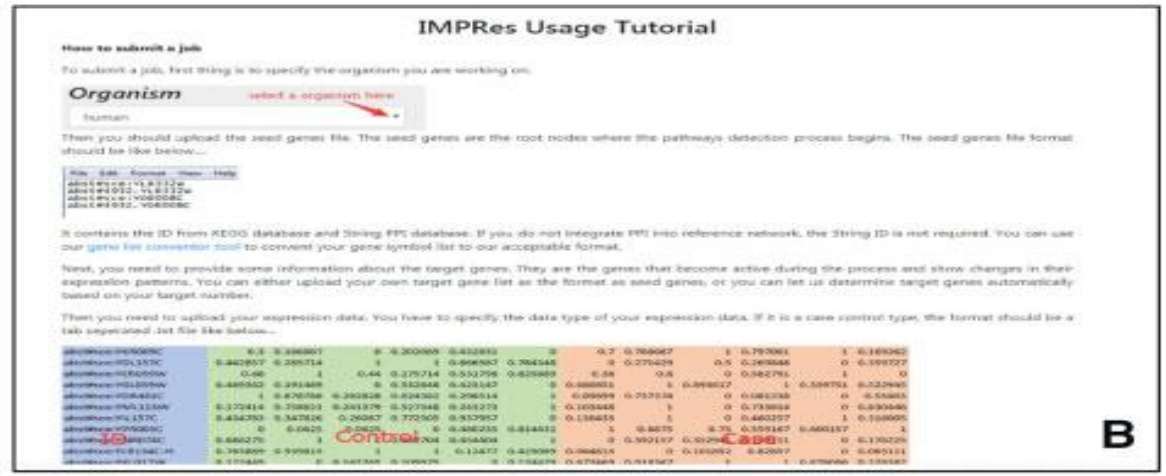
Pathways	P-value	Importance
AGE-RAGE	3.99E-9	Rap1
proteoglycans	5.43E-7	5 downregulated
focal adhesion	2.77E-5	PAK1, PAK6, EGFR, AKT3

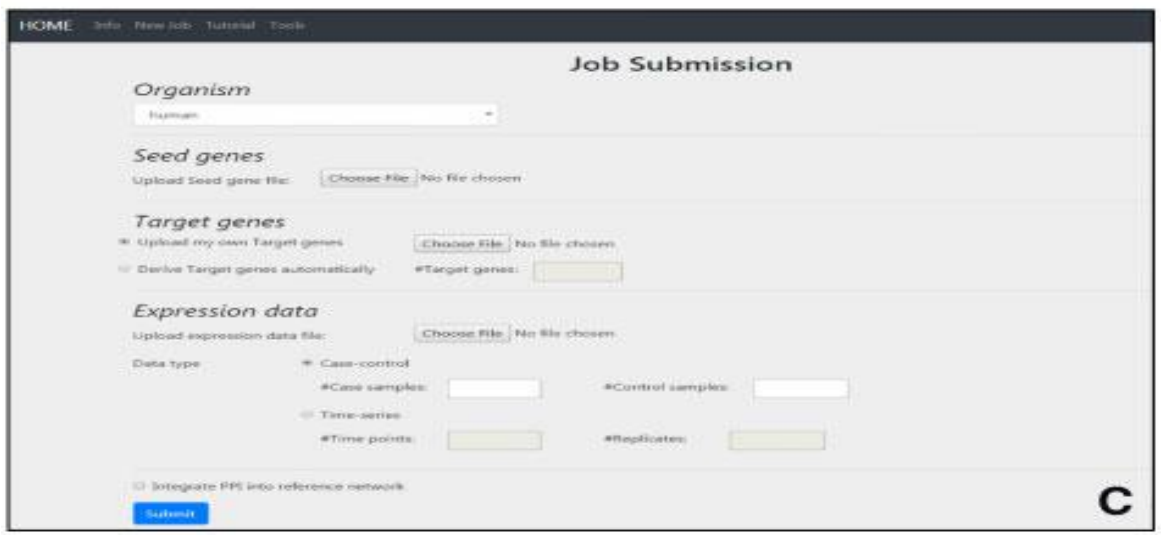
Result

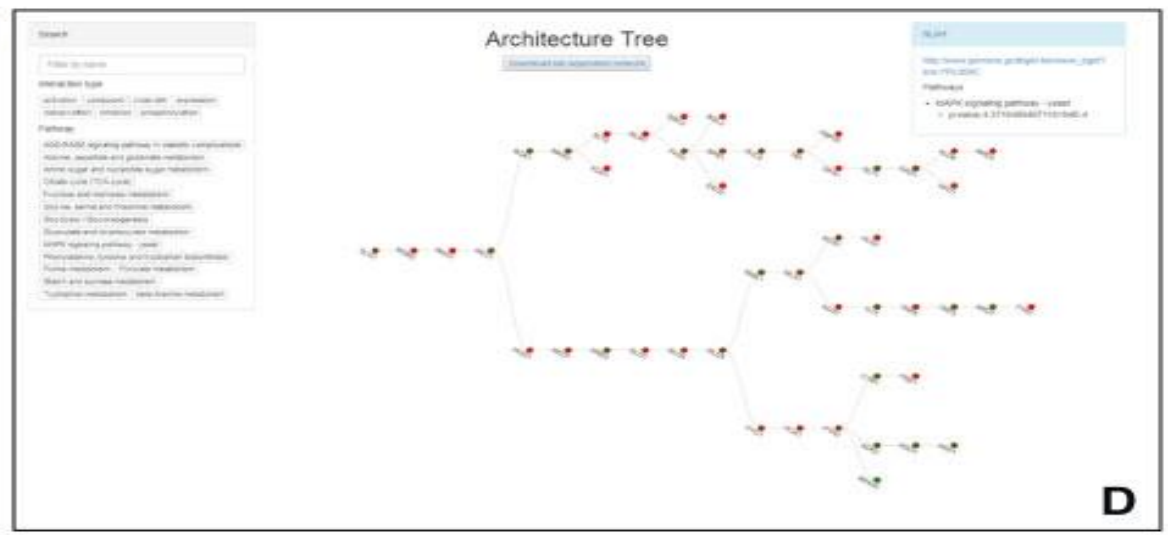

hsa04360: Axon guidance
hsa05205: Proteoglycans in cancer
hsa04510: Focal adhesion
hsa04933: AGE-RAGE signaling
0.6841
-0.5894
log2(fold change)
PLXNB3
RAC3
RAC1
RRAS
RHOA
RAC3
MAPK8
PAK6
PAK1
MAPK8
FOXO1
EGR1
PIK3R2
AKT3
AKT3
MTOR
RPS6KB2
ITGB1
PRKACG
PLCE1
ITPR1
CAMK2D
PLXNA2
FYN
FYN
CAV1
ITGB3
ITGB3
PRKCB
CAV3
SDC1
WNT1
FZD5
MAPK1
ROCK1
SLC9A1
STAT3
COL1A1
MAPK1
SMAD3
JUN
EDN1
AGER
FIGF
FIGF
TGFB1
TGFB1
hsa:406991
IGF1R
EGFR
PLCG2
PLCB4
PLCG2
cpd:C00076
EPHA2
EFNA4

Web server









https://impres.missouri.edu/impres