

APA – Version 3

Altered Pathway Analyzer (APA) is a cross-platform and standalone tool for analyzing gene expression datasets to highlight significantly rewired pathways across case-vs-control conditions. The Tool is designed to analyze human gene expression datasets (with Entrez ID); however, the analysis can be performed on gene expression datasets of other species by using appropriate flags and input files.

APA algorithm is unique prioritization algorithm that also uses gene-regulatory network to identify transcriptionally dysregulated pathways. It, thus, uses altered Transcription Factor (TF) and Target Gene (TG) relationship for prioritizing gene circuit rewired pathways. For Human datasets, it requires Gene expression matrices (RNAseq read count or Normalized gene expression values) with genes in Entrez ID format. For other species, please refer the manual below.

Prerequisite

OS: Windows, Mac or Linux **Compilers:**

[perl](#) and [R](#) (version $\geq 3.1.2$)

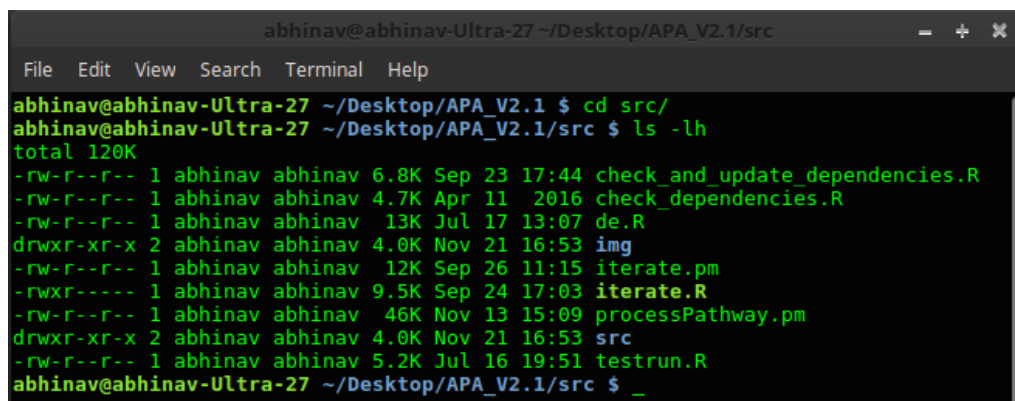
Dependencies: R packages- "[matrixStats](#)", "[mefa4](#)", "[dnet](#)", "[SANTA](#)", "[limma](#)", "[Biobase](#)"

Checking dependencies

After installing perl and R compilers along with other dependencies, user can check and update the remaining dependencies in the host OS using following steps:

Step 1: Unzip the installation package.

Step 2: Open "src/" folder in command-prompt or terminal.



```
abhinav@abhinav-Ultra-27 ~/Desktop/APA_V2.1/src
File Edit View Search Terminal Help
abhinav@abhinav-Ultra-27 ~/Desktop/APA_V2.1 $ cd src/
abhinav@abhinav-Ultra-27 ~/Desktop/APA_V2.1/src $ ls -lh
total 120K
-rw-r--r-- 1 abhinav abhinav 6.8K Sep 23 17:44 check_and_update_dependencies.R
-rw-r--r-- 1 abhinav abhinav 4.7K Apr 11 2016 check_dependencies.R
-rw-r--r-- 1 abhinav abhinav 13K Jul 17 13:07 de.R
drwxr-xr-x 2 abhinav abhinav 4.0K Nov 21 16:53 img
-rw-r--r-- 1 abhinav abhinav 12K Sep 26 11:15 iterate.pm
-rwxr----- 1 abhinav abhinav 9.5K Sep 24 17:03 iterate.R
-rw-r--r-- 1 abhinav abhinav 46K Nov 13 15:09 processPathway.pm
drwxr-xr-x 2 abhinav abhinav 4.0K Nov 21 16:53 src
-rw-r--r-- 1 abhinav abhinav 5.2K Jul 16 19:51 testrun.R
abhinav@abhinav-Ultra-27 ~/Desktop/APA_V2.1/src $ _
```

Step 3: Execute “check_and_update_dependencies.R” using following command

\$ Rscript check_and_update_dependencies.R

TRUE or

\$ [path to Rscript] check_and_update_dependencies.R TRUE

This will automatically scan the installed libraries and update the required dependencies. If you do not want to auto-install the R packages, use “FALSE” instead of “TRUE” in the above command line argument. By default, in Linux OS, path to “Rscript” is */usr/local/bin/Rscript* or */usr/bin/Rscript*. However, if multiple versions of R are installed, then it is mandatory to specify the path of R version for which all the dependencies are installed or required to be installed.

Either update dependencies manually or using above R script. If all dependencies are satisfied for desired version of R, you are now ready to use APA.

Starting APA

Right now, APA can only be executed via command line interface. To execute APA:

Step 1: Unzip the installation package, if not extracted earlier.

Step 2: Open the installation directory in command prompt or terminal and type:

\$ perl APA.pl

```
abhinav@abhinav-Lenovo-Y520-15IKBN ~/Desktop/APA_V3 $ perl APA.pl
      Altered Pathway Analyzer (version 3.0)
=====
Usage: perl APA.pl -case [path] -control [path] -o [path] [Options]
=====
[Mandatory]
-----
-case : Path to gene expression dataset of query samples (recommended sample size: 10)
-control : Path to gene expression dataset of control samples (recommended sample size: 10)
-o : Path to directory where all results will be saved.
=====
[Options]
-----
-rnaseq : The input gene expression matrices have RNAseq generated read count values.
-s : Path to text file containing entrez IDs of seed genes [Default: database/seeds.txt (Human)].
-P : Select one of precompiled pathway gene sets: KEGG [Default]/ NCI / MsigDB_C2 / PANTHER / Reactome / ALL (Human)
      or
      Path to text file containing pathway name and corresponding gene set in a specified format [Default: database/Pathways.ALL].
-grn : Path to text file containing background regulatory edges in space delimited format [Default: database/GRN.ssv (Human)].
-R : Path to directory having R executables- Rscript [Default: OS specific; not defined for IOS]
-GC : Gene count threshold [Default: 10]
-TF : List of transcription factor entrez IDs. Only valid with -grn [Default: database/TF.txt (Human)]
-m : Method by which disease gene is to be predicted in disease network. 1/2
      1 - Random walk by restart algorithm [Default]
      2 - Knet algorithm [Warning: Slow]
-r : Pearson correlation coefficient threshold [Default 0.1]
-p : P value threshold [Default: 0.05]
-H : [T/F] Whether column name or header is present in expression dataset or not. [Default: T]
-D : [T/F] Whether to show gene differential expression for sub-network gene prioritization. [Default: F]
-A : Whether to predict only differential regulation among altered pathways. T/F [Default: F]
-down : [T/F] Whether to include downregulated genes. T/F [Default: F]
-niter : [Int] Number of iterations for resampling and p-value calculation. Must be larger than or equal to 10. Recommended value 100 [Default off]
e.g.
perl APA.pl -rnaseq -case case.htseq.count -control control.htseq.count -o out
perl APA.pl -case sample/cancer.dat -control sample/control.dat -o out/ -P KEGG -m 2 -r 0.3
perl APA.pl -case sample/cancer.dat -control sample/control.dat -o out/ -P ALL -r 0.3 -s database/seeds.txt
perl APA.pl -case sample/cancer.dat -control sample/control.dat -o out/ -P database/Pathways.All -r 0.3 -s database/seeds.txt -niter 100
```

This will print command-line usage help. User is required to provide three mandatory arguments:

-case [path to input file] -control [path to input file] -o [path of output directory]

Input Files

The sample input files are provided in the installation package.

1. Case and control gene expression profiles.

Gene expression profile must be provided as normal text file in which first column must represent the gene ID and other columns represent the expression level (e.g. read count matrices from [htseq](#) or [prepareDE.py-stringtie](#)) of corresponding gene in different samples. The expression level can also be normalized expression values produced from microarray experiments. For RNAseq readcount, **-rnaseq** argument is mandatory in APA command line which allows scripts to perform [voom\(\)](#) normalizations on read count values. The file may or may not contain column headers. This file can either be *space* or *comma* or *tab* or *bar* (|) delimited. *For human dataset it is highly recommended to use Entrez gene ID only.*

Format of gene expression profile:

	Sample 1	Sample 2	Sample 3	Sample p
Entrez gene ID 1	Expression level	Expression level	Expression level	Expression level
Entrez gene ID ..	Expression level	Expression level	Expression level	Expression level
Entrez gene ID n	Expression level	Expression level	Expression level	Expression level

APA requires two such files- one for the case samples and other for the control samples. The format is general representation of gene expression datasets and many different data sources may provide pre-computed gene expression profiles. For example, with TCGA, user can obtain gene read-count values of thousands of human genes from thousands of cancer patient samples along with meta-data to distinguish case-vs-control samples. From raw fastq files, there are numerous ways of generating gene expression matrices apart from htseq and stringtie pipelines, e.g. [here](#).

2. Pathway gene sets *[Optional for human dataset]*

This text file contains sets of functionally related genes, i.e. Pathway gene sets. The simple format includes two columns separated by bar (|):

Column 1: Pathway name (e.g. Geneset 1)

Column 2: Tab separated list of gene IDs.

Pathway gene sets for humans are already available in installation package [database/pathways/]. Five different pathway databases are provided in the APA package, provided the gene expression datasets contain Entrez gene ID only.

NOTE: The gene IDs must be same with respect to IDs given in gene expression profile datasets. *If you are using pre-compiled human pathway gene sets, then it is mandatory to use Entrez gene ID in gene expression dataset.*

3. Reference gene regulatory network *[Optional for human dataset]*

The reference gene regulatory network includes the list of gene pairs, i.e. Transcription factor and its known/predicted target gene. The gene pairs will be tested for being differentially correlated across case-control samples. The text file format includes three columns:

Column 1: Gene ID [TF or TG]

Column 2: Gene ID [TF or TG]

Column 3: Edge attribute- TF-TG or TF-TF or TG-TF

Example:

Gene_ID_1	Gene_ID_2	TF-TG
Gene_ID_1	Gene_ID_3	TF-TG
Gene_ID_3	Gene_ID_5	TF-TG

Each line in the file represent an edge in network either between TF and TF (TF-TF); or TF and TG (TF-TG); or TG and TF (TG-TF). Here the third column is an edge attributes and represents the direction of relationship, i.e. TF-TG or TF-TF or TG-TF. The third column is optional; if third column is not present then each edge will be considered as TF-TG.

For humans, the reference gene regulatory is pre-compiled in APA package (see manuscript or database/GRN.ssv; works only with Entrez gene ID in gene expression dataset), however, if user wants different reference regulatory network, then its file path must be provided in APA command-line arguments with appropriate flag.

4. Seeds genes [optional]

Seed gene set is the list of known disease linked genes that will be used by APA to predict novel disease genes within close-proximity. Seed genes represent list of those gene IDs which are known to be or most likely to be dysfunction across given samples. For most of the human diseases list of such genes can be obtained from different biomarker databases. For human cancer, a list of 102 known cancer-related Entrez gene IDs is available (default), however, for non-cancerous or non-human datasets, users are advised to provide their own list of seeds as newline separated gene IDs in a text file.

```
Gene ID 1
Gene ID 2
..
Gene ID n
```

APA accepts this list of genes and using “*guild-by-association*” principal it predicts the novel set of genes that shares network proximity with given seed genes. We called these genes as context-specific disease genes. The tool then predicts dysregulated pathways that are enriched with context-specific disease genes involved in causing sub-network rewiring.

Running the sample dataset

The APA is packaged with sample dataset in which 17 control samples has wild-type p53 gene status and 33 samples have mutated p53 samples. The APA can be executed as:

```
$ perl APA.pl -case sample/case.txt -control sample/control.txt -o sam_output
```

By default KEGG [human] pathway database will be used, if user wants to change the pathway database then -P flag can be used:

```
$ perl APA.pl -case sample/case.txt -control sample/control.txt -o sam_output -P PANTHER
```

In this case, instead of KEGG, PANTHER database for human pathways will be used. For more such options, type:

```
$ perl APA.pl
```

Note: For RNAseq read count ‘-rnaseq’ argument is required in APA command line.

Output

All the output files will be available in output directory (provided with -o flag). The main file is “index.html”, which should be open with any updated browser (**recommended: google chrome**). The results are self-explanatory in which computed scores for each predicted dysregulated pathways are given, with high score represent high dysregulation and/or differential regulation (see manuscript). Sample results are available at:

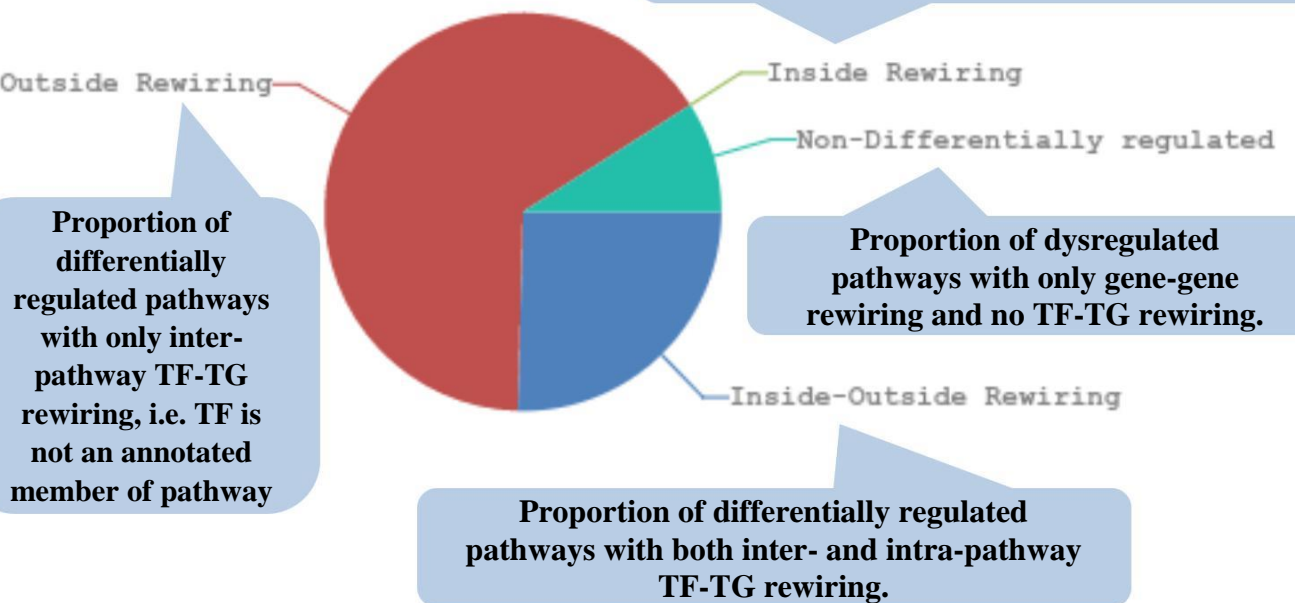
<http://bioinfo.icgeb.res.in/APA /tp53>

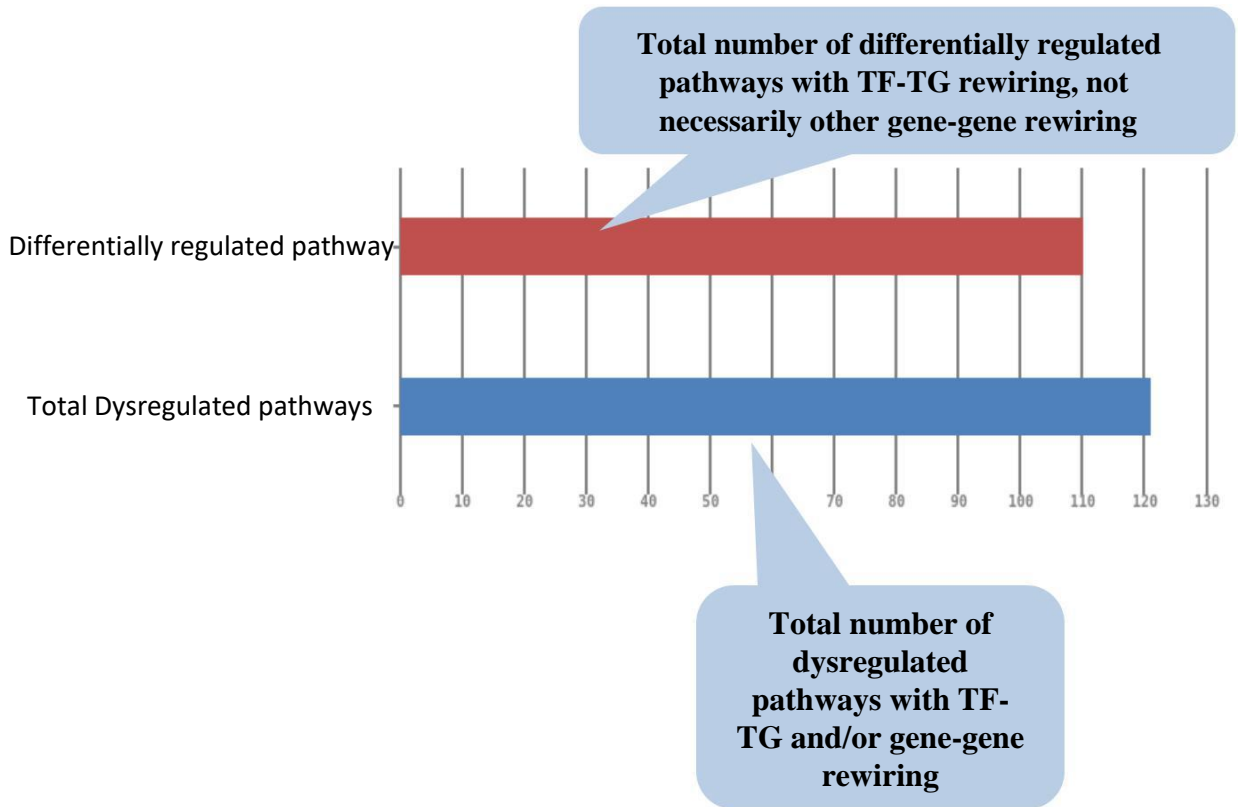
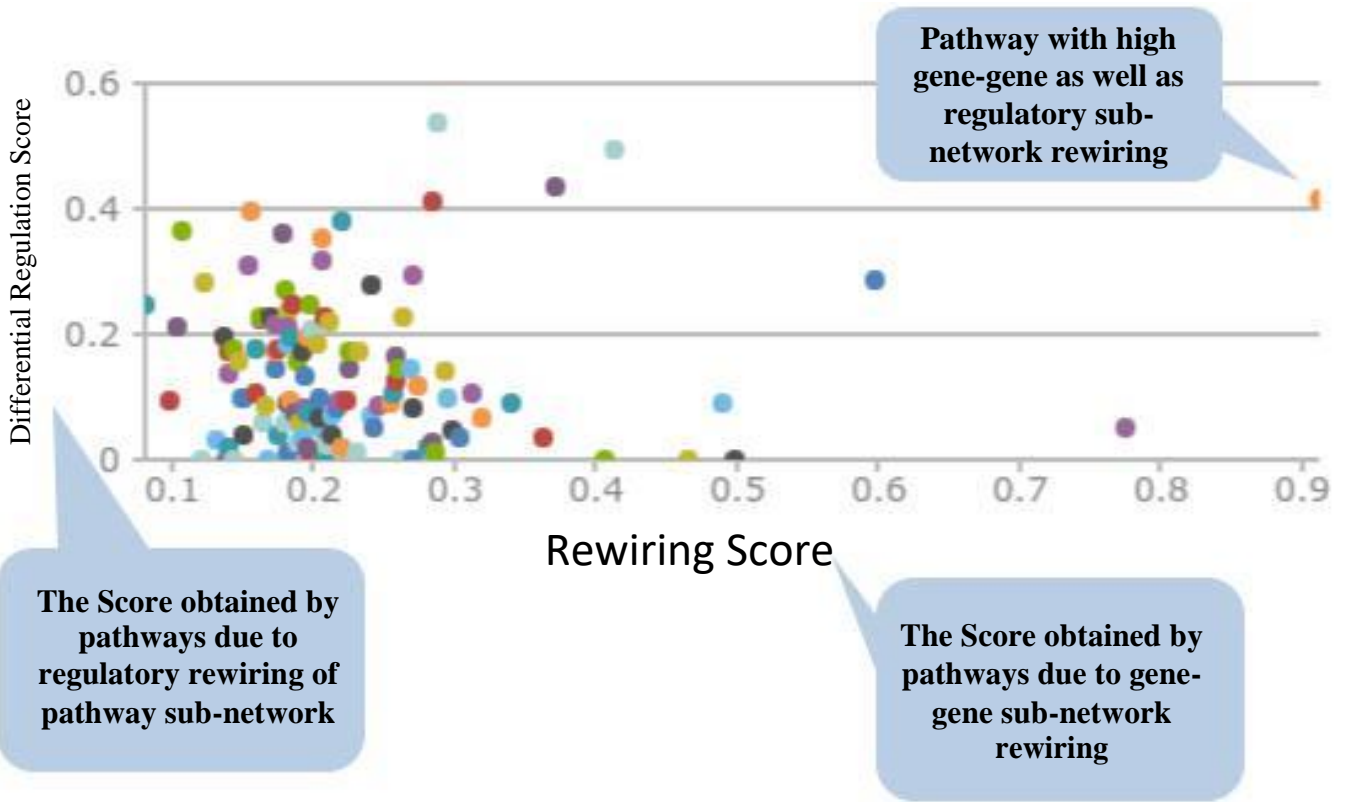
Apart from HTML results, the result table can also be accessed with text file - “result.tsv”. This is tab separated file which should be open with software like MS excel or Libre office spreadsheet.

APA Output: Summary

Total number of dysregulated pathways predicted with APA .

Pathway Count: 121





Sub-network Analysis

DR score: The score obtained by a pathway sub-network due to TF-TG (regulatory) rewiring

DY score: The score obtained by a pathway sub-network due to gene-gene (non-regulatory) rewiring

Result table. Click header for column-wise sorting

PATHWAY	RANK ▼	DY SCORE	DR SCORE	DP SCORE
PATHWAYS_IN_CANCER	1.000	0.179	0.272	0.306
JAK_STAT_SIGNALING_PATHWAY	0.496	0.185	0.246	0.102
PROSTATE_CANCER	0.468	0.154	0.311	1.000

Gene Count: 166

Pathway name

The rank obtained by the pathway with respect to other pathway. The rank measures the overall pathway dysregulation due to regulatory as well as non-regulatory rewiring

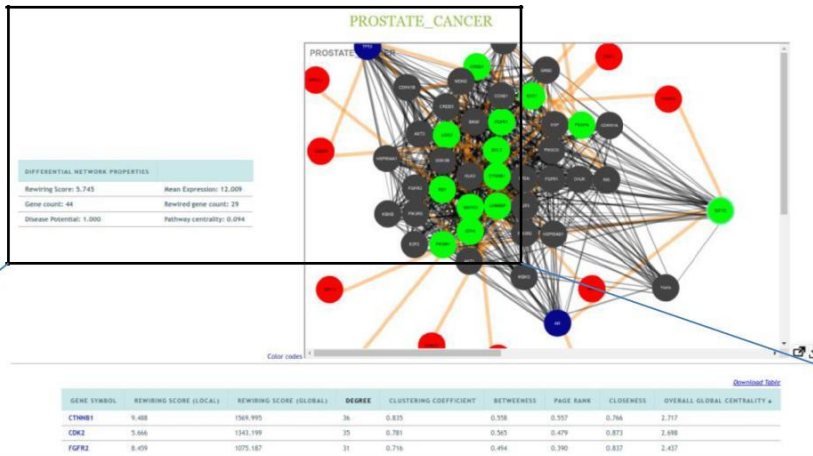
The rank reflects the overall aberration of sub-network wiring across case-control samples. Higher the rank, more likely a pathway sub-network to include both regulatory and non-regulatory rewiring.

DP score: The score reflects the enrichment of predicted disease genes in dysregulated sub-network of a given pathway.

The score enables the user to re-prioritize dysregulated pathway.

Pathways with higher DP score have higher enrichment of context specific genes (e.g. genes closely connected with mutated genes). These pathway can thus be used therapeutic targeting, if user defined seed genes are known disease genes.

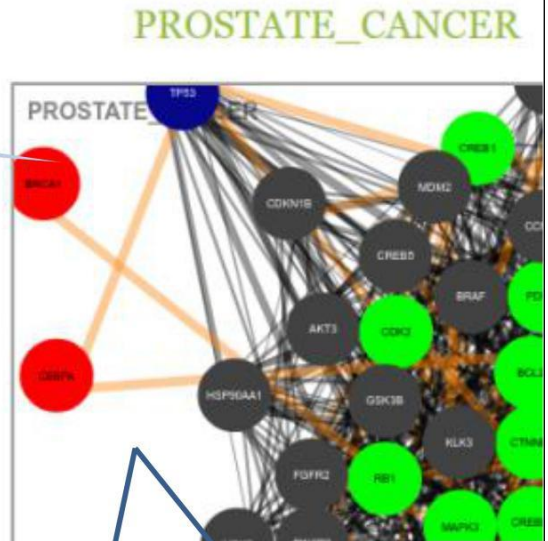
Sub-network Analysis



The cytoscape window for subnetwork visualization.

Rewiring score: The change in the strength of gene-gene connections across case-control networks.

Sub-network properties of pathway "Prostate cancer".



DIFFERENTIAL NETWORK PROPERTIES	
Rewiring Score: 5.745	Mean Expression: 12.009
Gene count: 44	Rewired gene count: 29

- TF that is not the annotated member of pathway gene set.
- TF that is the annotated member of pathway gene set.
- Up-regulated target gene that shares rewired regulatory connection with at least one TF.
- Any other gene.
- Rewired connection.

Sub-network Analysis

Sub-network properties: Properties of genes within a given sub-network.

Gene ID

These properties can be used to identify those genes which are network important as compared to other genes of the same pathway.

[Download Table](#)

GENE SYMBOL	REWIRING SCORE (LOCAL)	REWIRING SCORE (GLOBAL)	DEGREE	CLUSTERING COEFFICIENT	BETWEENNESS	PAGE RANK	CLOSENESS	OVERALL GLOBAL CENTRALITY ▲
CTNNB1	9.488	1569.995	36	0.835	0.558	0.557	0.766	2.717
CDK2	5.666	1343.199	35	0.781	0.565	0.479	0.873	2.698
FGFR2	8.459	1075.187	31	0.716	0.494	0.390	0.837	2.437
PDPK1	0.000	1043.800	36	0.469	0.503	0.400	0.858	2.230
MDM2	6.285	890.853	39	0.369	0.417	0.353	0.856	1.996
EGF	3.626	687.396	36	0.546	0.321	0.247	0.776	1.890
PIK3R1	4.727	722.624	35	0.482	0.337	0.275	0.782	1.876
AKT1	7.889	605.889	34	0.610	0.289	0.215	0.728	1.842
PIK3CG	1.457	674.781	33	0.550	0.287	0.270	0.638	1.745

The network centrality of gene within the given sub-network, i.e. local

The score can be used to identify the genes which are more important in the given pathway.

The network centrality of gene within the complete rewired network, i.e. global

The score can be used to identify the genes that interacts frequently with all the expressed genes, not necessarily with pathway gene set.

The average centrality score for each pathway gene in the respective sub-network.