**Prerequisite**

**OS**: Windows, Mac or Linux **Compilers**:perlandR(version≥ 3.1.2)

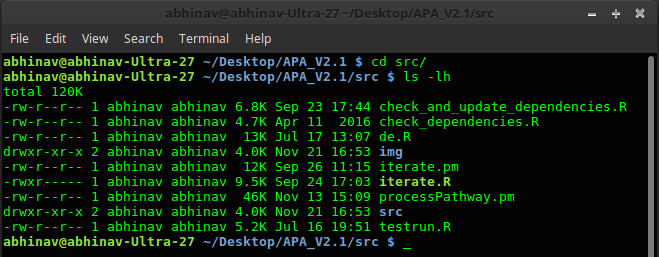
**Dependencies**: R packages- “matrixStats”, “mefa4”, “dnet”, “SANTA”,“limma”, “Biobase”

**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.**



**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/Rscrip*t 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.



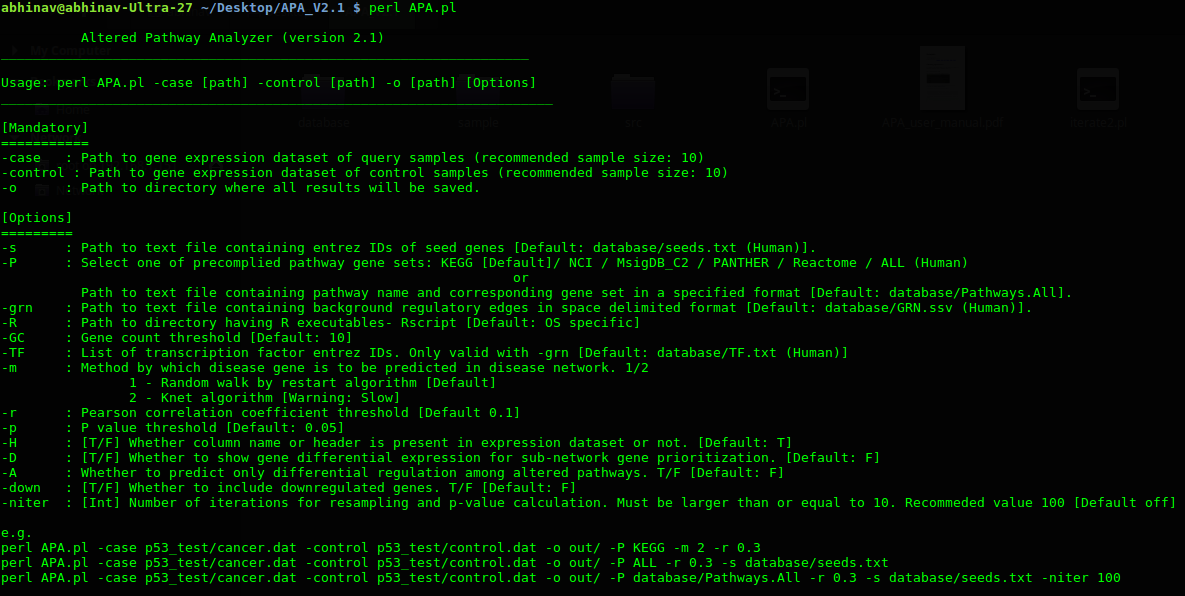
Either update dependencies manually or using above R script. If all dependencies are satisfied for desired version of R, you are now ready to use APA.

**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**

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 of corresponding gene in different samples. 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.

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.



GeneID1

GeneID2

GeneID*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**

**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

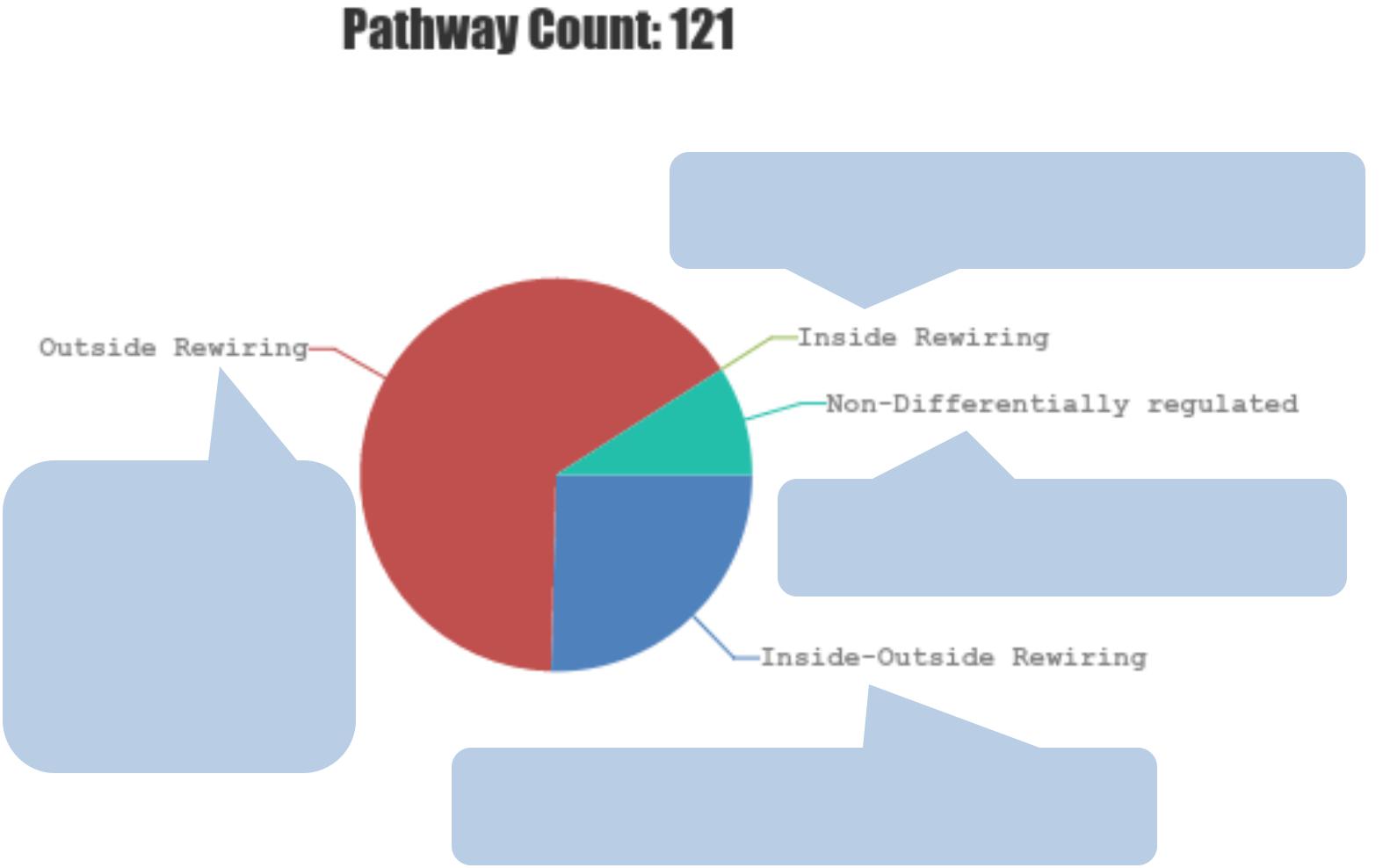
Apart from HTML results, the result table can also be accessed with text file - “result.tsv”. This is tab separated file which should be open with software like MS excel or Libre office spreadsheet.

**APA Output: Summary**



**Proportion of differentially regulated pathways with only inter-pathway TF-TG rewiring, i.e. TF is not an annotated member of pathway**

**Total number of dysregulated pathways predicted with APA .**



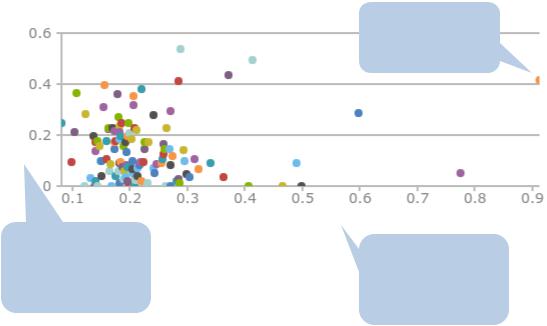
**Proportion of differentially regulated pathways with only intra-pathway TF-TG rewiring.**

**Proportion of dysregulated pathways with only gene-gene rewiring and no TF-TG rewiring.**

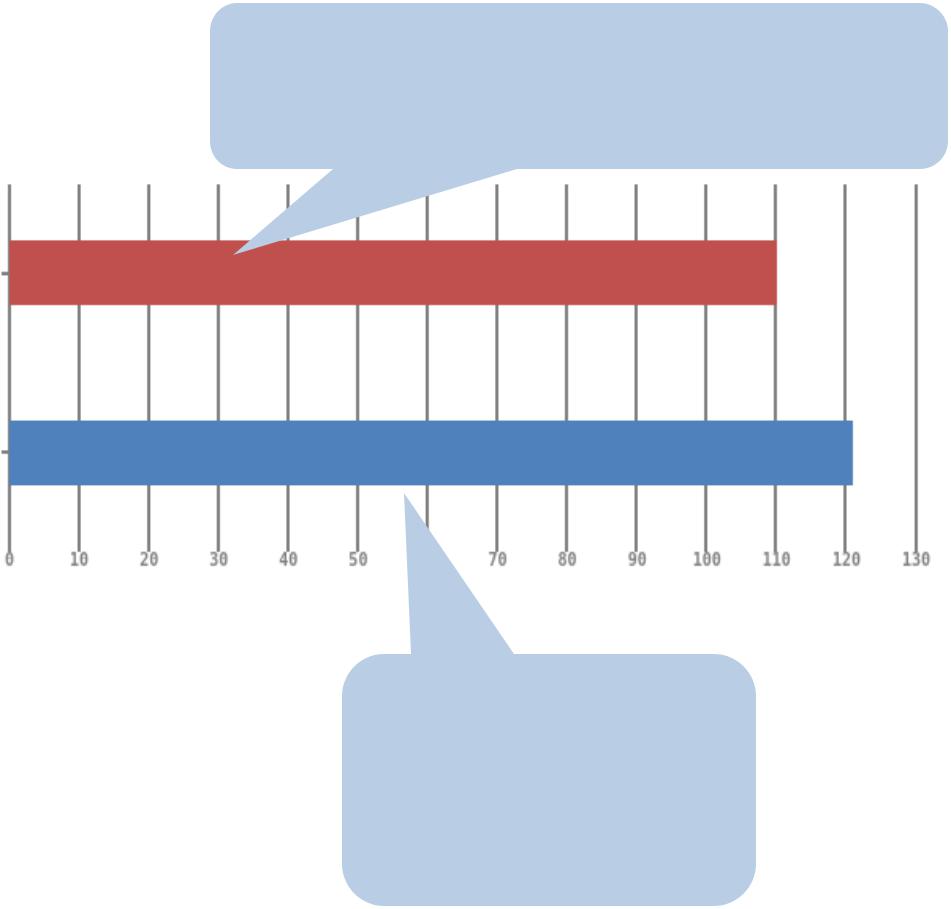
**Proportion of differentially regulated pathways with both inter- and intra-pathway TF-TG rewiring.**

|  |
| --- |
| Differential Regulation Score |

**Pathway with high gene-gene as well as regulatory sub-network rewiring**



|  |  |  |
| --- | --- | --- |
|  | Rewiring Score |  |
| **The Score obtained by** | **The Score obtained by** |  |
| **pathways due to** |  |
| **regulatory rewiring of** | **pathways due to gene-** |  |
| **pathway sub-network** | **gene sub-network** |  |
|  | **rewiring** |  |



**Total number of differentially regulated pathways with TF-TG rewiring, not necessarily other gene-gene rewiring**

Differentially regulated pathway

Total Dysregulated pathways

**Total number of dysregulated pathways with TF-TG and/or gene-gene rewiring**

**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**

**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**

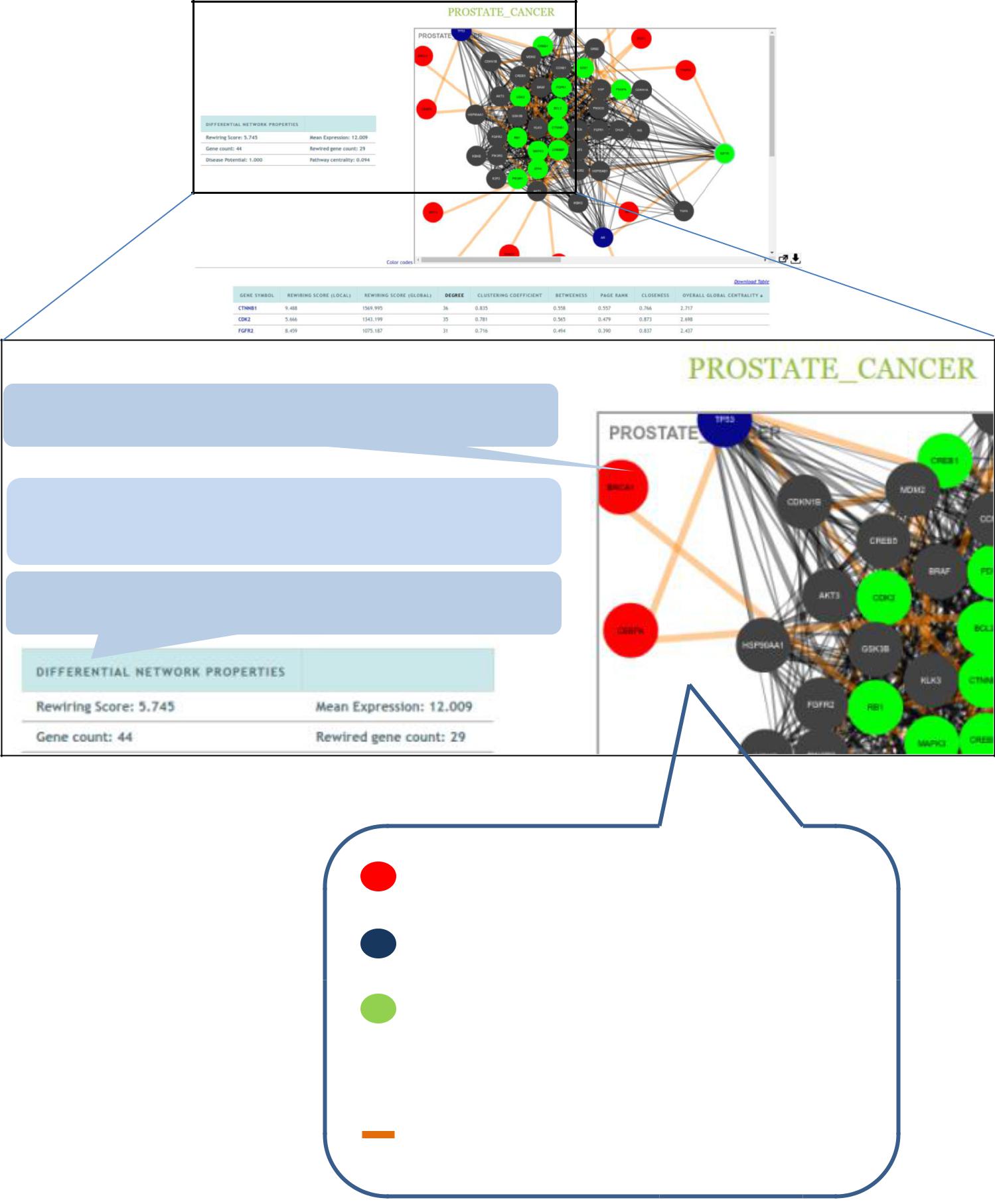
**The rank reflects the overall aberration of sub-network wiring across case-control samples. Higher the rank, more likely a pathway sub-network to include both regulatory and non-regulatory rewiring.**

**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”.**

TF that is not the annotated member of pathway gene set.

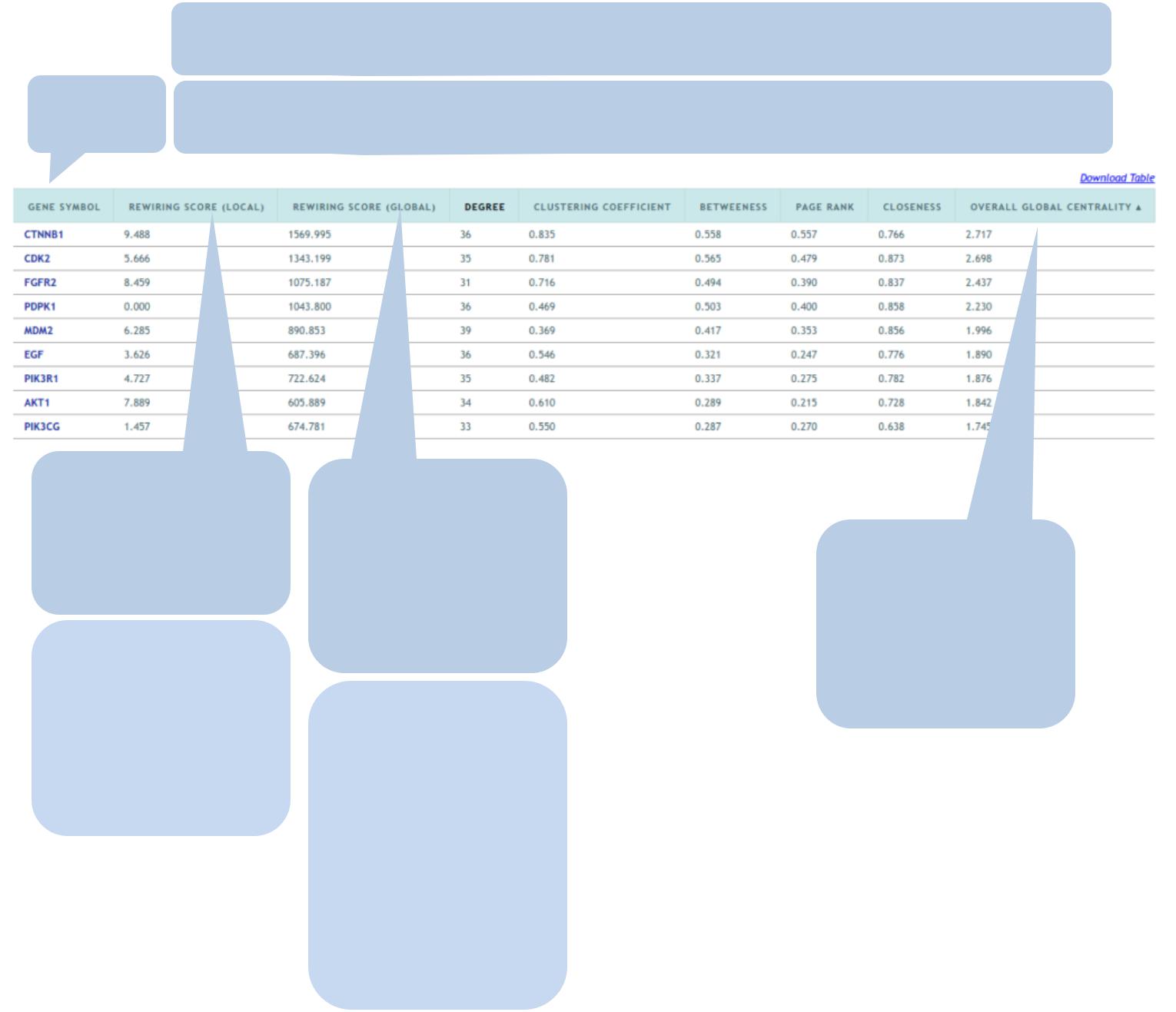
TF that is the annotated member of pathway gene set.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **The network** | **The network** |  |  |
| **centrality of gene** |  |  |
| **centrality of** |  |  |
| **within the given** | **The average** |  |
| **gene within the** |  |
| **sub-network, i.e.** |  |
| **centrality score** |  |
| **complete** |  |
| **local** |  |
| **for each pathway** |  |
| **rewired network,** |  |
|  |  |
| **The score can be** | **gene in the** |  |
| **i.e. global** |  |
| **used to identify** | **The score can be** | **respective sub-** |  |
| **the genes which** | **network.** |  |
| **used to identify** |  |
| **are more** |  |  |
| **the genes that** |  |  |
| **important in the** |  |  |
| **interacts** |  |  |
| **given pathway.** |  |  |
| **frequently with** |  |  |
|  |  |  |
|  | **all the expressed** |  |  |
|  | **genes, not** |  |  |
|  | **necessarily with** |  |  |
|  | **pathway gene** |  |  |
|  | **set.** |  |  |