# Prerequisite

**OS**: Windows, Mac or Linux
**Compilers**: perl and R (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.**

```
abhinav@abhinav-Ultra-27 ~/DPA $ cd src/
abhinav@abhinav-Ultra-27 ~/DPA/src $ ls -alh
total 124K
drwxr-xr-x 4 abhinav abhinav 4.0K Apr  8 17:32 .
drwxr-xr-x 4 abhinav abhinav 4.0K Apr  8 17:29 ..
-rw-r--r-- 1 abhinav abhinav 4.7K Mar 12 17:22 check_dependencies.R
-rw-r--r-- 1 abhinav abhinav  12K Mar 12 16:57 de.R
drwxr-xr-x 2 abhinav abhinav 4.0K Nov 20 13:53 img
-rw-r--r-- 1 abhinav abhinav  11K Mar  2 18:15 iterate.pm
-rwxr----- 1 abhinav abhinav 8.8K Feb 28 14:04 iterate.R
-rwxr-xr-x 1 abhinav abhinav  43K Mar 13 16:07 processPathway.pm
-rw-r--r-- 1 abhinav abhinav  17K Oct 28 17:46 .Rhistory
drwxr-xr-x 2 abhinav abhinav 4.0K Jan  3 17:35 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 DPA.

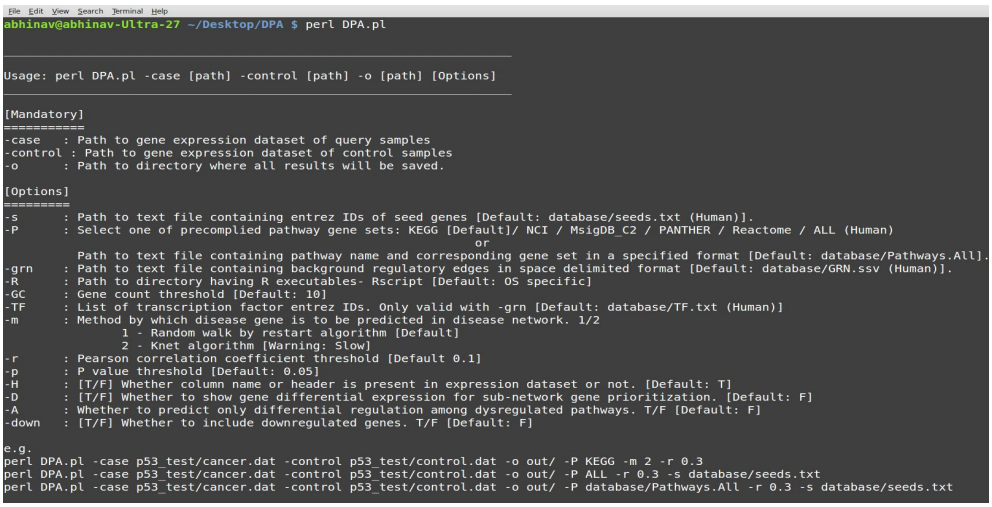
# Starting DPA

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

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

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

**$ perl DPA.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 |

DPA 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 DPA 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 DPA 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 DPA 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 DPA 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*

DPA 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 as genes as context-specific disease genes. The tool the predicts the dysregulated pathways that are enriched with context-specific disease genes that are involved in causing sub-network rewiring.

# Running the sample dataset

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

**$ perl DPA.pl -case sample/case.txt -control sample/control.txt -o p53_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 DPA.pl -case sample/case.txt -control sample/control.txt -o p53_output -P PANTHER**

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

**$ perl DPA.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/DPA/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.