RePhine User Guide

summary

RePhine takes in Pharmacogenomics data containing 1) gene expression data; 2) gene copy number data; 3) mutation data; 4) confounding information (such as cancer types, tissue type et.al, recommended but not necessary) for estimation of associations between TR-affected gene expression and drug response.

RePhine also takes in ChIP-seq data (Regulatory potential, RP scores) for target inference.

GitHub link: https://github.com/coexps/Rephine

Package available

We have compiled RePhine code to R package, which could be downloaded in GtHub (https://github.com/coexps/Rephine/blob/master/Rephine_0.1.0.tar.gz)

Package Install

Packages can be installed with the install.packages() function in R. For example.
Install.package (Rephine_0.1.0.tar.gz)

Data Preparation and format transformation

Pharmacogenomics data are required to be structured for processing.

Expression data: each row represents gene information and each column should be samples or cells. Sample name and gene name are required. Gene name should be official gene names as is used in ChIP-seq RP scores files

CN data: each row represents genes and each column should be samples or cells. Sample name and gene name are required.

Examples of expression data or copy number data

Gene_name	Cell1	Cell2	Cell3	Cell4
Gene1	2.4	5.4	2.5	5.4
Gene2	5.4	4.5	34.5	2.5

mutation data: mutation information should be dummy values. 1 stands for mutation and 0 stands for wildtype. each row represents genes and each column should be samples or cells.

Examples of mutation data

Gene_name	Cell1	Cell2	Cell3	Cell4
Gene1	1	0	1	0
Gene2	0	0	1	0

Note*: sample names or cell names of Pharmacogenomics data should be consistent. Inconsistent sample may lead to errors. For example, RePhine could distinguish between A-375 and A375. Samples from Pharmacogenomics data should be overlapped. Genes in expression data and copy number data should also be overlapped. All gene name of Pharmacogenomics data should be official gene names as is used in ChIP-seq RP scores files.

Drug response data: drug response data help RePhine to identify the target patterns associated with drug response. Row is the cell line and columns the drugs. RePhine could detect the association between target patterns and drug response. If higher drug response values stand for drug sensitivity, the positive correlation coefficient means TR targets are positively correlated with drug response and vice versa.

Note* The sample orders, cell line name and sample counts should be totally consistent with the samples in mutation matrix. Format should be get structured before using the RePhine.

Example of the Drug response matrix

Gene_name	Drug1	Drug2	Drug3	Drug4
Cell1	0.4	1.2	0.5	0.5
Cell2	0.5	5.3	3.2	1.3

ChIP-seq data: RePhine use ChIP-seq RP scores to infer target information. RePhine has processed Encode data and calculated RP scores in manuscript and for users. These resources are available in Github (https://github.com/coexps/Rephine).

Custom ChIP-seq data: If the users are interested in Custom TF regulations other than Encode TFs. RePhine enables users to generate custom TF RP scores from ChIP-seq peak bed files.

“RP_calculation_modified.py” is offered to calculated the RP scores from peak files by considering both peak distance to genes and peak signal strength. But RP scores of all TFs should be merged and in structured format. First columns should be gene orders or ranks. The second column should be Official gene symbols. The following columns are the corresponding RP scores from ChIP-seq data. Each row are the gene. Gene with different isoforms could be duplicated. RePhine will choose the isoforms with the highest RP scores.

Example of RP scores data format

#order	symbol	ChIP-seq1	ChIP-seq2
1	WASH7P	0.0	0.0
2	FAM138A	0.0	4.1
3	FAM138F	3.0	0.0
4	FAM138C	0.0	2.1
5	OR4F5	0.0	2.3

Optional data:

Confounder files: In the manuscript, RePhine considers cancer types as confounders besides confounding mutations. Therefore, RePhine compile user custom confounders for adjustments. Confounder files should be in dummy values. Each column should be the candidate confounders and each row should be samples.

Example of Confounding files:

Gene_name	Solid tumor	Metastasis	Gender_Male
Cell1	1	0	1
Cell2	0	1	1
Cell3	0	0	1

Note* confounding mutations will be automatically estimated by RePhine. It is not required for users to calculate. Cell name and cell information should be same as in drug response data and mutation data.

TCGA differential files: In the manuscript, RePhine weighted genes according to genes differential expression in TCGA types to remove non-cancer related gene bias. Files are available at GitHub.

How to Use RePhine

Load package

After install the package, load the package in R by library function or require function

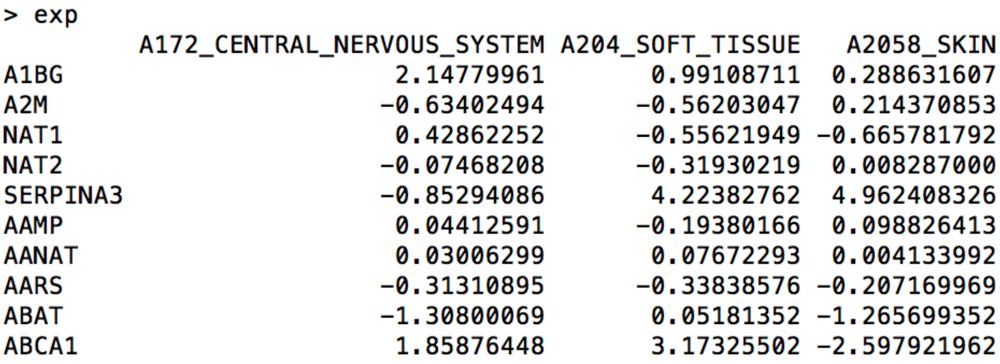
```
> library(Rephine)
```

Load demo data

If users want to use the demo data to test the workflows or pipelines, you can use the data function.

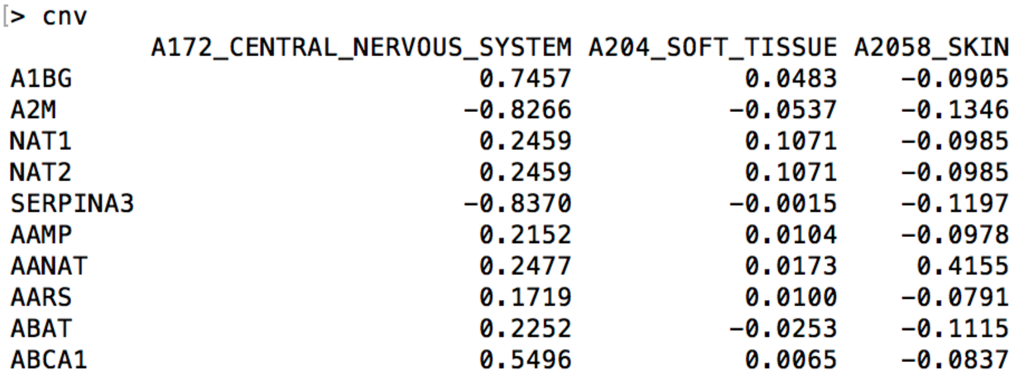
```
> data(Rephine)
```

“exp” is the demo of expression data. This data is required for RePhine analysis



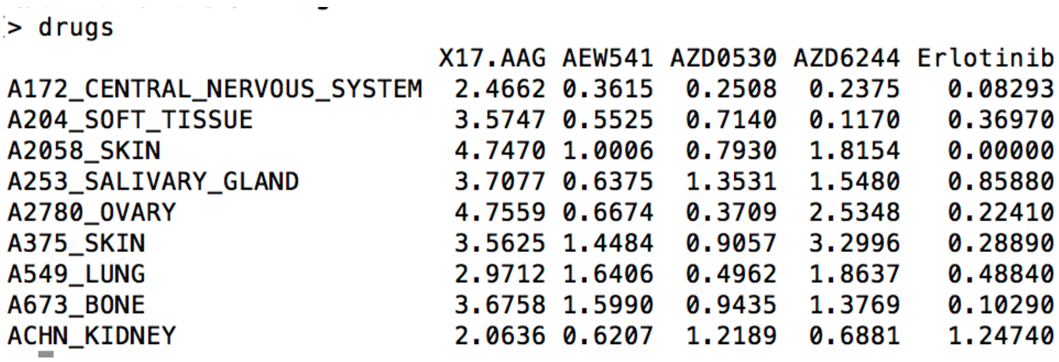
```
> exp
         A172_CENTRAL_NERVOUS_SYSTEM A204_SOFT_TISSUE   A2058_SKIN
A1BG                      2.14779961       0.99108711  0.288631607
A2M                      -0.63402494      -0.56203047  0.214370853
NAT1                      0.42862252      -0.55621949 -0.665781792
NAT2                     -0.07468208      -0.31930219  0.008287000
SERPINA3                 -0.85294086       4.22382762  4.962408326
AAMP                      0.04412591      -0.19380166  0.098826413
AANAT                     0.03006299       0.07672293  0.004133992
AARS                     -0.31310895      -0.33838576 -0.207169969
ABAT                     -1.30800069       0.05181352 -1.265699352
ABCA1                     1.85876448       3.17325502 -2.597921962
```

“cnv” is the demo of cnv data. This data is required for RePhine analysis



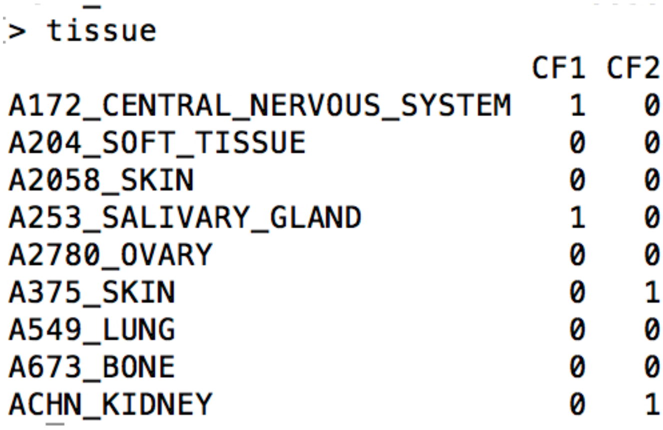
```
> cnv
         A172_CENTRAL_NERVOUS_SYSTEM A204_SOFT_TISSUE A2058_SKIN
A1BG                          0.7457           0.0483    -0.0905
A2M                          -0.8266          -0.0537    -0.1346
NAT1                          0.2459           0.1071    -0.0985
NAT2                          0.2459           0.1071    -0.0985
SERPINA3                     -0.8370          -0.0015    -0.1197
AAMP                          0.2152           0.0104    -0.0978
AANAT                         0.2477           0.0173     0.4155
AARS                          0.1719           0.0100    -0.0791
ABAT                          0.2252          -0.0253    -0.1115
ABCA1                         0.5496           0.0065    -0.0837
```

“drugs” is the demo fo drug response data. This data is required for RePhine analysis



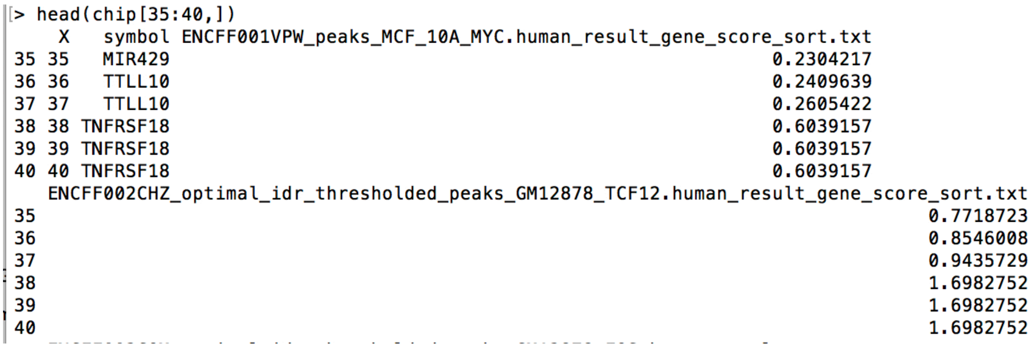
```
> drugs
                            X17.AAG AEW541 AZD0530 AZD6244 Erlotinib
A172_CENTRAL_NERVOUS_SYSTEM  2.4662 0.3615  0.2508  0.2375   0.08293
A204_SOFT_TISSUE             3.5747 0.5525  0.7140  0.1170   0.36970
A2058_SKIN                   4.7470 1.0006  0.7930  1.8154   0.00000
A253_SALIVARY_GLAND          3.7077 0.6375  1.3531  1.5480   0.85880
A2780_OVARY                  4.7559 0.6674  0.3709  2.5348   0.22410
A375_SKIN                    3.5625 1.4484  0.9057  3.2996   0.28890
A549_LUNG                    2.9712 1.6406  0.4962  1.8637   0.48840
A673_BONE                    3.6758 1.5990  0.9435  1.3769   0.10290
ACHN_KIDNEY                  2.0636 0.6207  1.2189  0.6881   1.24740
```

“tissue” is the demo of confounding information of each cell lines. this information is optional and not required.



```
> tissue
                            CF1 CF2
A172_CENTRAL_NERVOUS_SYSTEM   1   0
A204_SOFT_TISSUE              0   0
A2058_SKIN                    0   0
A253_SALIVARY_GLAND           1   0
A2780_OVARY                   0   0
A375_SKIN                     0   1
A549_LUNG                     0   0
A673_BONE                     0   0
ACHN_KIDNEY                   0   1
```

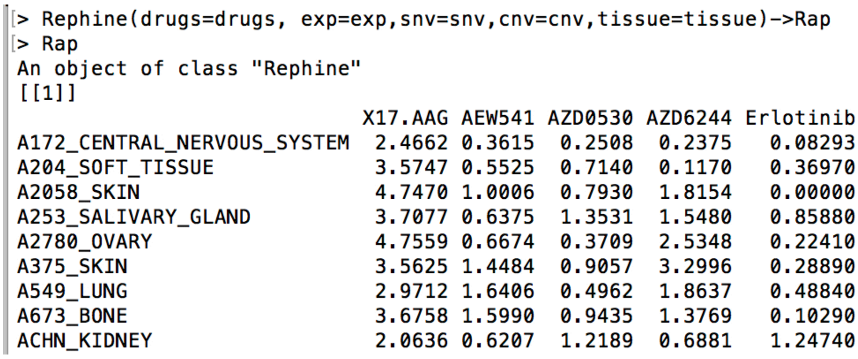
“chip” is the demo of RP scores calculated from ChIP-seq data. All RP scores could be downloaded from GitHub (https://github.com/coexps/Rephine/tree/master/source)



```
> head(chip[35:40,])
    X   symbol ENCFF001VPW_peaks_MCF_10A_MYC.human_result_gene_score_sort.txt
35 35   MIR429                                                      0.2304217
36 36   TTLL10                                                      0.2409639
37 37   TTLL10                                                      0.2605422
38 38 TNFRSF18                                                      0.6039157
39 39 TNFRSF18                                                      0.6039157
40 40 TNFRSF18                                                      0.6039157
   ENCFF002CHZ_optimal_idr_thresholded_peaks_GM12878_TCF12.human_result_gene_score_sort.txt
35                                                                                 0.7718723
36                                                                                 0.8546008
37                                                                                 0.9435729
38                                                                                 1.6982752
39                                                                                 1.6982752
40                                                                                 1.6982752
```

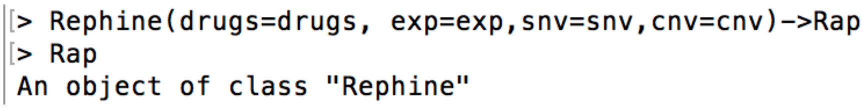
All the demo data could alternatively be download from GitHub.

First of all, load data into R use the RePhine function



```
> Rephine(drugs=drugs, exp=exp,snv=snv,cnv=cnv,tissue=tissue)->Rap
> Rap
An object of class "Rephine"
[[1]]
                           X17.AAG AEW541 AZD0530 AZD6244 Erlotinib
A172_CENTRAL_NERVOUS_SYSTEM  2.4662 0.3615  0.2508  0.2375   0.08293
A204_SOFT_TISSUE             3.5747 0.5525  0.7140  0.1170   0.36970
A2058_SKIN                   4.7470 1.0006  0.7930  1.8154   0.00000
A253_SALIVARY_GLAND          3.7077 0.6375  1.3531  1.5480   0.85880
A2780_OVARY                  4.7559 0.6674  0.3709  2.5348   0.22410
A375_SKIN                    3.5625 1.4484  0.9057  3.2996   0.28890
A549_LUNG                    2.9712 1.6406  0.4962  1.8637   0.48840
A673_BONE                    3.6758 1.5990  0.9435  1.3769   0.10290
ACHN_KIDNEY                  2.0636 0.6207  1.2189  0.6881   1.24740
```

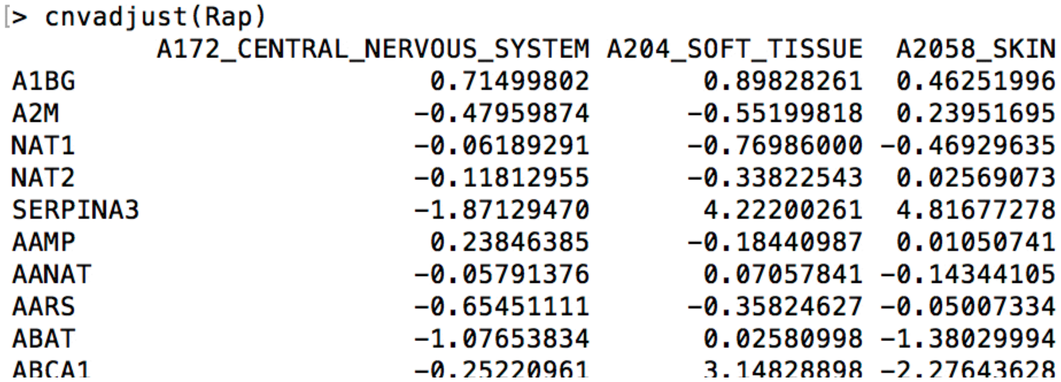
If there is no tissue information, leave it blank.



```
> Rephine(drugs=drugs, exp=exp,snv=snv,cnv=cnv)->Rap
> Rap
An object of class "Rephine"
```

Secondly, calculate the effect of TRs on gene expression by adjusting the CNVs.

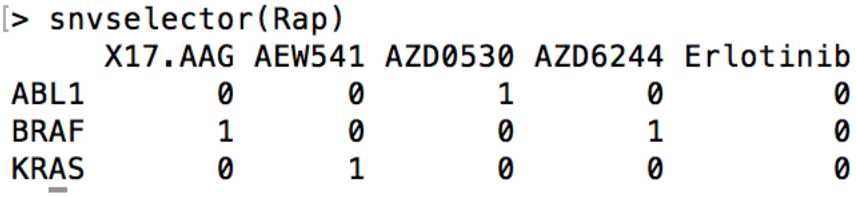
```
cnvadjust(Rap)->reg
```



```
> cnvadjust(Rap)
         A172_CENTRAL_NERVOUS_SYSTEM A204_SOFT_TISSUE  A2058_SKIN
A1BG                      0.71499802       0.89828261  0.46251996
A2M                      -0.47959874      -0.55199818  0.23951695
NAT1                     -0.06189291      -0.76986000 -0.46929635
NAT2                     -0.11812955      -0.33822543  0.02569073
SERPINA3                 -1.87129470       4.22200261  4.81677278
AAMP                      0.23846385      -0.18440987  0.01050741
AANAT                    -0.05791376       0.07057841 -0.14344105
AARS                     -0.65451111      -0.35824627 -0.05007334
ABAT                     -1.07653834       0.02580998 -1.38029994
ABCA1                    -0.25220961       3.14828898 -2.27643628
```

Thirdly, evaluate the mutations that affect the drug response independent of TR regulations

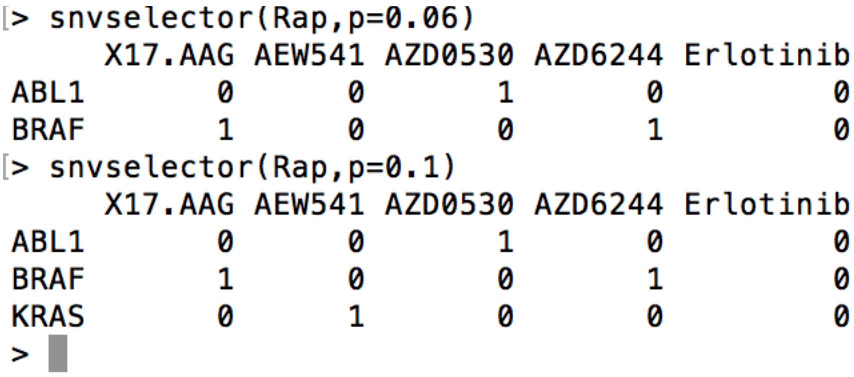
```
snvselector(Rap)->snvselect
```



```
> snvselector(Rap)
     X17.AAG AEW541 AZD0530 AZD6244 Erlotinib
ABL1       0      0       1       0         0
BRAF       1      0       0       1         0
KRAS       0      1       0       0         0
```

note* : the parameter p is the significance cutoff of likelihood ratio test. More significant

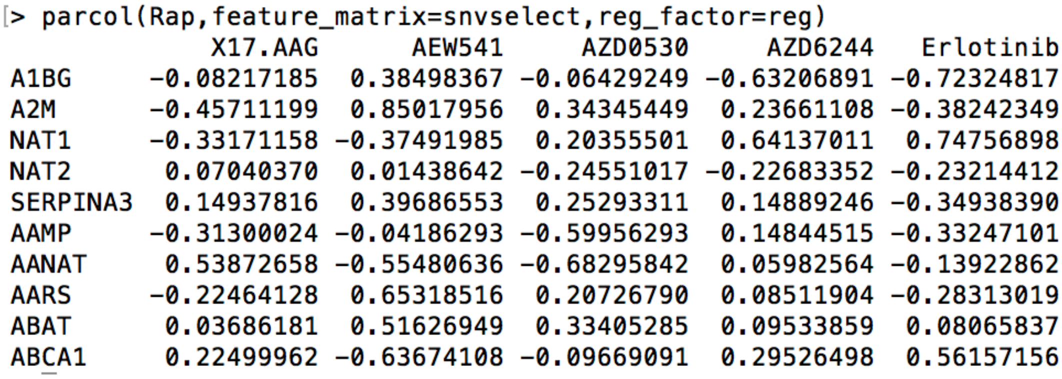
P will lead to less selected mutations.



```
> snvselector(Rap,p=0.06)
     X17.AAG AEW541 AZD0530 AZD6244 Erlotinib
ABL1       0      0       1       0         0
BRAF       1      0       0       1         0
> snvselector(Rap,p=0.1)
     X17.AAG AEW541 AZD0530 AZD6244 Erlotinib
ABL1       0      0       1       0         0
BRAF       1      0       0       1         0
KRAS       0      1       0       0         0
> 
```

Next, calculate the partial correlation between drug and adjusted expression by accounting for independent mutations and given confounders.

```
>parcol( Rap,feature_matrix=snvselect,reg_factor=reg,tissue=NULL) -> partial
```

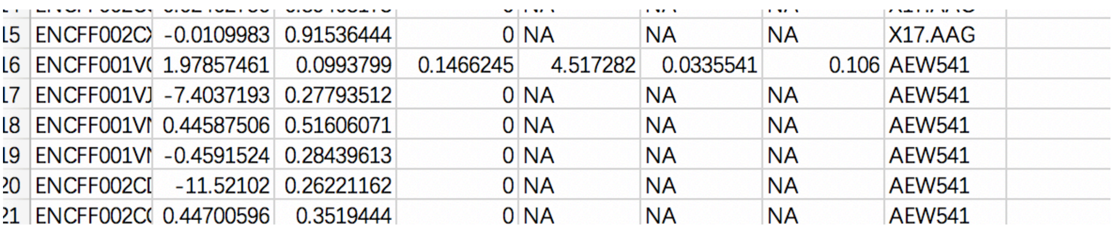


```
> parcol(Rap,feature_matrix=snvselect,reg_factor=reg)
             X17.AAG      AEW541     AZD0530     AZD6244   Erlotinib
A1BG     -0.08217185  0.38498367 -0.06429249 -0.63206891 -0.72324817
A2M      -0.45711199  0.85017956  0.34345449  0.23661108 -0.38242349
NAT1     -0.33171158 -0.37491985  0.20355501  0.64137011  0.74756898
NAT2      0.07040370  0.01438642 -0.24551017 -0.22683352 -0.23214412
SERPINA3  0.14937816  0.39686553  0.25293311  0.14889246 -0.34938390
AAMP     -0.31300024 -0.04186293 -0.59956293  0.14844515 -0.33247101
AANAT     0.53872658 -0.55480636 -0.68295842  0.05982564 -0.13922862
AARS     -0.22464128  0.65318516  0.20726790  0.08511904 -0.28313019
ABAT      0.03686181  0.51626949  0.33405285  0.09533859  0.08065837
ABCA1     0.22499962 -0.63674108 -0.09669091  0.29526498  0.56157156
```

Note, if you want to downweigh non-cancer related genes, you can use the parameter "tcga", the corresponding matrix file is available on GitHub.

At last calculate the significance of TRs by using Elastic net, Likelihood ratio test and permutation test.

```
>chipseq(chip=chip,stage_all_partial=partial)
```



15	ENCFF002C›	-0.0109983	0.91536444	0	NA	NA	NA	X17.AAG	
16	ENCFF001V(	1.97857461	0.0993799	0.1466245	4.517282	0.0335541	0.106	AEW541	
17	ENCFF001VJ	-7.4037193	0.27793512	0	NA	NA	NA	AEW541	
18	ENCFF001VI	0.44587506	0.51606071	0	NA	NA	NA	AEW541	
19	ENCFF001VI	-0.4591524	0.28439613	0	NA	NA	NA	AEW541	
20	ENCFF002CI	-11.52102	0.26221162	0	NA	NA	NA	AEW541	
21	ENCFF002C(	0.44700596	0.3519444	0	NA	NA	NA	AEW541	

The first colum showed TF information and selected ChIP-seq data, the second column is the coefficient of univariate followed by significance P-value. The forth column is the elastic net coefficient followed by significance. The seventh column is the permutation significance. if the candidate TR could not independently regulate drug response and not be selected by Elastic net, RePhine will give a NA this these sections (column 4-6). The last column is the TR influencing drugs.

Note if users has custom chip-seq data and custom correlation values, he/she can direct use the chipseq function to evaluate the TRs with significance.

For more information

RePhine package is available on GitHub. R code of RePhine workflow is also deposited to GitHub, which could be modified by users for custom data analysis.

Citation information

The RePhine manuscript is under preparation

If you have some problems, please contact wangxujun87@sjtu.edu.cn