

DifeCo: Differential Cooccurrence or Mutual Exclusion of Binary Genomic Alteration Data

Author: Yuanqing Yan yuanqing.yan@uth.tmc.edu

January 28, 2021

Description

Gene mutation resulting in functional dysregulation is the direct cause of most genetic diseases. In many diseases, some gene mutations tend to occur together and compensate the biological functions with each other. While for some other mutations, their functions are redundant and tend to mutually exclude with each other. This phenomenon is commonly seen in cancer biology. For example, in IDH-WT GBM, TP53 and RB1 mutations often cooccurring, while CDKN2A/B loss is mutually exclusive with TP53 mutation. Due to the disease heterogeneousness, the pattern of gene mutation cooccurrence/mutual exclusion could vary, such as patients with long vs short survival, or patients between different subtypes. The differential cooccurrence/mutual exclusion of gene mutations could be critical for disease treatment. DifeCo is an R package to evaluate the differential occurrence/mutual exclusion of gene mutation. It fits a Firth's bias-reduced logistic regression model between pairwise genes plus the additional group variable. An interaction term of independent predictors is introduced and its significance is evaluated. After the multiplicity adjustment, the pairs of gene are regarded to be statistically significant if the adjusted p value of interaction term is less than the designed cutoff. For the model with interaction term failing to reach significance, the additive model without interaction term is fit to evaluate the cooccurrence/mutual exclusion in the entire dataset. In addition to test the differential cooccurrence/mutual exclusion, DifeCo package can also be used to evaluate and visualize pairwise gene cooccurrence/mutual exclusion in two datasets (Separate mode) or single one dataset (Single mode). Which model to be used purely depends on the hypothesis as well as the nature of the data.

Installation

Installing DifeCo from GitHub

```
library(devtools)
install_github("yuanqing_yan/DifeCo")
```

Load data

Load IDH-WT gbm data. Here we focus on UT cohort and assume the patients can be splitted into two groups based on the status of PTEN. Here we have PTEN-WT and PTEN-Alt. This example is to illustrate how to use package with DC mode.

```
library(DifeCo)
data(gbm_dat)
head(gbm_dat[,1:5])
#>   cohort CDKN2A.B PTEN EGFR TP53
#> 1     UT       1    1    1    1
#> 2     UT       0    0    0    1
#> 3     UT       0    0    0    1
```

```

#> 4      UT      1      0      0      0
#> 5      UT      1      1      0      0
#> 6      UT      1      0      1      1
MutDat_UT<-gbm_dat[gbm_dat$cohort=="UT",]
#Remove cohort column
MutDat_UT<-MutDat_UT[,-1]
PTEN<-MutDat_UT$PTEN;table(PTEN)
#> PTEN
#>  0  1
#> 111 121
MutDat_WoPTEN<-subset(MutDat_UT,select=-PTEN);head(MutDat_WoPTEN[,1:5])
#>   CDKN2A.B  EGFR  TP53  NF1  MLL2
#> 1         1    1    1    0    0
#> 2         0    0    1    0    0
#> 3         0    0    1    1    0
#> 4         1    0    0    0    0
#> 5         1    0    0    0    1
#> 6         1    1    1    0    0

```

DC mode

First step is to evaluate the differential cooccurrence/mutual exclusion of genomic alterations between the patients with wild type PTEN (PTEN-WT) and PTEN function altered (PTEN-Alt). We set up the FDR cutoff equal to 0.1.

```

Result_DC<-DC.CO_Evaluation(mutation_data=MutDat_WoPTEN,
                             group=PTEN,
                             which_group_to_be_one=1,
                             mode="DC",
                             adjust.method="BH",
                             FDRCutoff=0.1)

#> [1] "Evaluating DC P, pair1:1; pair2:2"
#> [1] "Evaluating DC P, pair1:1; pair2:3"
#> [1] "Evaluating DC P, pair1:1; pair2:4"
#> [1] "Evaluating DC P, pair1:1; pair2:5"
#> [1] "Evaluating DC P, pair1:1; pair2:6"
#> [1] "Evaluating DC P, pair1:1; pair2:7"
#> [1] "Evaluating DC P, pair1:1; pair2:8"
#> [1] "Evaluating DC P, pair1:1; pair2:9"
#> [1] "Evaluating DC P, pair1:1; pair2:10"
#> [1] "Evaluating DC P, pair1:1; pair2:11"
#> [1] "Evaluating DC P, pair1:1; pair2:12"
#> [1] "Evaluating DC P, pair1:1; pair2:13"
#> [1] "Evaluating DC P, pair1:1; pair2:14"
#> [1] "Evaluating DC P, pair1:1; pair2:15"
#> [1] "Evaluating DC P, pair1:1; pair2:16"
#> [1] "Evaluating DC P, pair1:1; pair2:17"
#> [1] "Evaluating DC P, pair1:1; pair2:18"
#> [1] "Evaluating DC P, pair1:1; pair2:19"
#> [1] "Evaluating DC P, pair1:1; pair2:20"
#> [1] "Evaluating DC P, pair1:1; pair2:21"
#> [1] "Evaluating DC P, pair1:1; pair2:22"
#> [1] "Evaluating DC P, pair1:2; pair2:3"
#> [1] "Evaluating DC P, pair1:2; pair2:4"
#> [1] "Evaluating DC P, pair1:2; pair2:5"

```

[illegible]

[illegible]

[illegible]

```

#> [1] "Evaluating DC P, pair1:14; pair2:15"
#> [1] "Evaluating DC P, pair1:14; pair2:16"
#> [1] "Evaluating DC P, pair1:14; pair2:17"
#> [1] "Evaluating DC P, pair1:14; pair2:18"
#> [1] "Evaluating DC P, pair1:14; pair2:19"
#> [1] "Evaluating DC P, pair1:14; pair2:20"
#> [1] "Evaluating DC P, pair1:14; pair2:21"
#> [1] "Evaluating DC P, pair1:14; pair2:22"
#> [1] "Evaluating DC P, pair1:15; pair2:16"
#> [1] "Evaluating DC P, pair1:15; pair2:17"
#> [1] "Evaluating DC P, pair1:15; pair2:18"
#> [1] "Evaluating DC P, pair1:15; pair2:19"
#> [1] "Evaluating DC P, pair1:15; pair2:20"
#> [1] "Evaluating DC P, pair1:15; pair2:21"
#> [1] "Evaluating DC P, pair1:15; pair2:22"
#> [1] "Evaluating DC P, pair1:16; pair2:17"
#> [1] "Evaluating DC P, pair1:16; pair2:18"
#> [1] "Evaluating DC P, pair1:16; pair2:19"
#> [1] "Evaluating DC P, pair1:16; pair2:20"
#> [1] "Evaluating DC P, pair1:16; pair2:21"
#> [1] "Evaluating DC P, pair1:16; pair2:22"
#> [1] "Evaluating DC P, pair1:17; pair2:18"
#> [1] "Evaluating DC P, pair1:17; pair2:19"
#> [1] "Evaluating DC P, pair1:17; pair2:20"
#> [1] "Evaluating DC P, pair1:17; pair2:21"
#> [1] "Evaluating DC P, pair1:17; pair2:22"
#> [1] "Evaluating DC P, pair1:18; pair2:19"
#> [1] "Evaluating DC P, pair1:18; pair2:20"
#> [1] "Evaluating DC P, pair1:18; pair2:21"
#> [1] "Evaluating DC P, pair1:18; pair2:22"
#> [1] "Evaluating DC P, pair1:19; pair2:20"
#> [1] "Evaluating DC P, pair1:19; pair2:21"
#> [1] "Evaluating DC P, pair1:19; pair2:22"
#> [1] "Evaluating DC P, pair1:20; pair2:21"
#> [1] "Evaluating DC P, pair1:20; pair2:22"
#> [1] "Evaluating DC P, pair1:21; pair2:22"

```

The gene pairs with significantly differential cooccurrence/mutual exclusion can be extracted by Stat.Extraction function. In this study, the pair of PIK3CA and PIK3R1 shows significantly differential cooccurrence/mutual exclusion. In other word, the pattern of cooccurrence/mutual exclusion of PIK3CA and PIK3R1 depends on the group. In details, alterations in PIK3CA and PIK3R1 are mutually excluded in PTEN-WT group, while they cooccur in PTEN-Alt group. For the gene pairs without significant differential cooccurrence/mutual exclusion, the pattern is evaluated in the entire dataset and 13 pairs show significant.

```

sta_DC<-Stat.Extraction(obj=Result_DC)
sig_DC<-sta_DC$Stat
#Extract the gene pairs with significant differential cooccurrence/mutual exclusion
sig_DC[sig_DC$Sig.In.DC=="Yes",]
#>      Gene1 Gene2 LogOR_Group0 LogOR_Group1 RawP_Test.DC FDR_Test.DC
#> 113 PIK3CA PIK3R1   -2.098058    2.268684 0.0002720034 0.06147278
#>      LogOR_AdjustGroup RawP_AdjustGroup FDR_AdjustGroup Sig.In.DC Sig.In.CO
#> 113                NA                NA                NA        Yes        No
#Extract the gene pairs with significant cooccurrence/mutual exclusion in UT chort dataset
nrow(sig_DC[sig_DC$Sig.In.CO=="Yes",])
#> [1] 13
head(sig_DC[sig_DC$Sig.In.CO=="Yes",])

```

```

#>      Gene1  Gene2 LogOR_Group0 LogOR_Group1 RawP_Test.DC FDR_Test.DC
#> 1  CDKN2A.B  EGFR           NA           NA           NA           NA
#> 2  CDKN2A.B  TP53           NA           NA           NA           NA
#> 5      EGFR   NF1           NA           NA           NA           NA
#> 12     EGFR PDGFRA          NA           NA           NA           NA
#> 21    PDGFRA  KDR           NA           NA           NA           NA
#> 29 CDKN2A.B  RB1           NA           NA           NA           NA
#>      LogOR_AdjustGroup RawP_AdjustGroup FDR_AdjustGroup Sig.In.DC Sig.In.CO
#> 1           1.221320      1.221320      1.251297e-03      No      Yes
#> 2           -2.047585      -2.047585      2.254988e-09      No      Yes
#> 5           -2.019804      -2.019804      1.778703e-06      No      Yes
#> 12          -1.136231      -1.136231      3.282103e-02      No      Yes
#> 21           2.858522       2.858522      1.195980e-10      No      Yes
#> 29          -2.073175      -2.073175      5.094044e-06      No      Yes

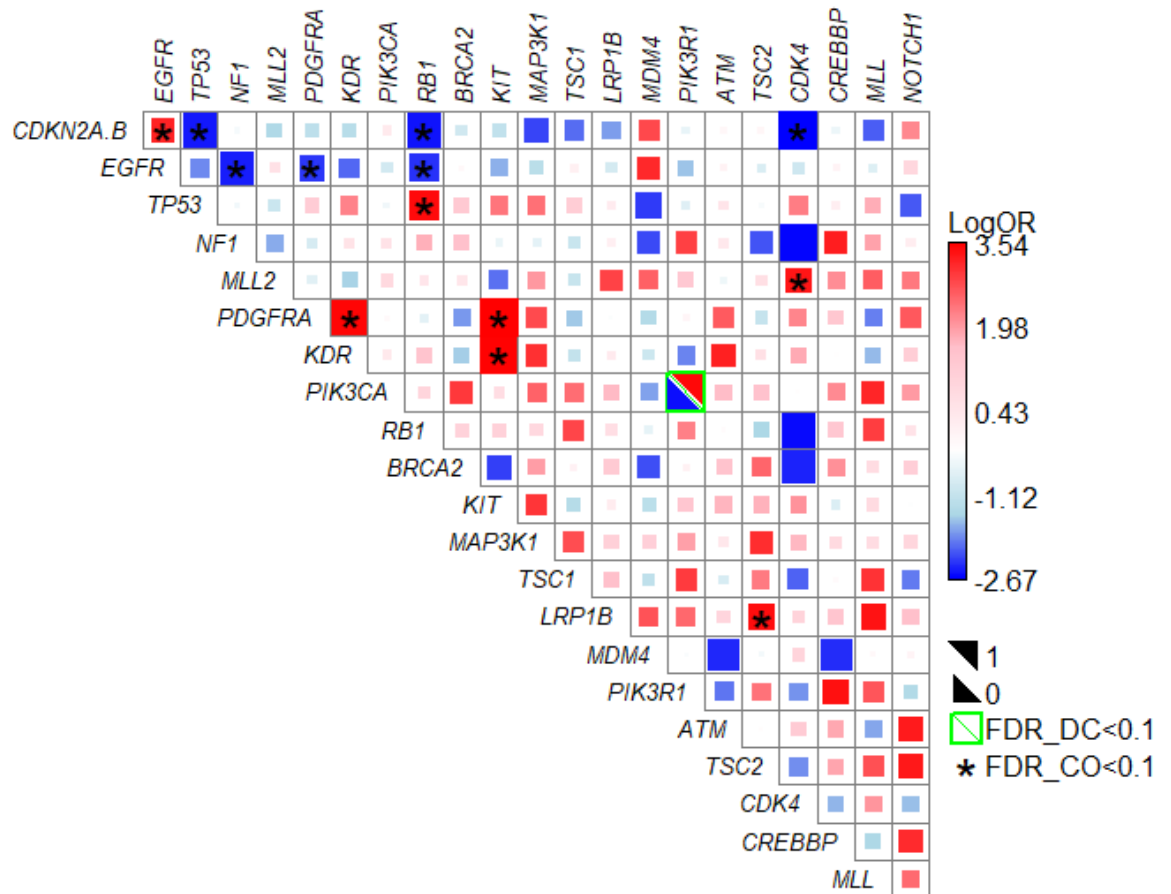
```

The following code is to visualize the result. The grid with green color shows the gene pair with significant differential cooccurrence/mutual exclusion. The cells with * are the ones with significant cooccurrence/mutual exclusion for entire UT dataset.

```

DC.CO_plot(obj=sta_DC,
           label.gene.cex=0.8)

```



Separate mode

Significance of cooccurrence/mutual exclusion is evaluated in each sub dataset in "Separate" mode. This is particularly helpful to investigate the gene pair patterns in two different datasets (One dataset for discovery purpose, while the other for validation). The example below evaluates cooccurrence/mutual exclusion in UT cohort and validate the result in TCGA cohort.

```
Cohort=gbm_dat$cohort
MutDat<-gbm_dat[,-1]
Result_Sep<-DC.CO_Evaluation(mutation_data=MutDat,
                             group=Cohort,
                             which_group_to_be_one='UT',
                             mode="Separate",
                             adjust.method="BH",
                             FDRcutoff=0.05)
```

Extract the significant gene pairs. Here, we use WhichGrp_adjP.Separate="Each" to do the multiplicity adjustment in UT and TCGA dataset separately.


```

sta_Sep<-Stat.Extraction(obj=Result_Sep,
                          WhichGrp_adjP.Separate="Each")

sig_Sep<-sta_Sep$Stat
#Extract the significant gene pairs in UT cohort
sig_Sep[sig_Sep$Sig.In.UT=="Yes",]
#>      Gene1 Gene2 LogOR.TCGA   Raw.P.TCGA FDR.Each.Grp.TCGA FDR.ALL.TCGA
#> 2   CDKN2A.B EGFR  0.5618808 3.012549e-02    4.234305e-01 3.419447e-01
#> 4   CDKN2A.B TP53 -1.3368143 3.824625e-06    1.382329e-04 1.382329e-04
#> 9      EGFR NF1 -1.6236225 1.001788e-05    3.168153e-04 3.168153e-04
#> 18     EGFR PDGFRA -0.5117044 1.329513e-01    8.008733e-01 7.312321e-01
#> 28     PDGFRA KDR  3.1957534 5.270840e-12    4.445075e-10 4.445075e-10
#> 37   CDKN2A.B RB1 -2.3387241 1.395115e-07    7.059282e-06 6.417529e-06
#> 39     EGFR RB1 -1.0438041 1.021231e-02    1.845511e-01 1.781873e-01
#> 40     TP53 RB1  2.0431515 8.505302e-07    3.586402e-05 3.310525e-05
#> 62     PDGFRA KIT  4.8182634 1.267824e-22    3.207594e-20 6.415189e-20
#> 63      KDR KIT  3.7553692 2.395803e-14    3.030691e-12 4.040922e-12
#> 172 CDKN2A.B CDK4 -2.1722233 3.384359e-09    2.140607e-07 2.140607e-07
#>      Sig.In.TCGA LogOR.UT   Raw.P.UT FDR.Each.Grp.UT FDR.ALL.UT Sig.In.UT
#> 2              No  1.235063 6.858884e-05    1.928108e-03 1.928108e-03    Yes
#> 4              Yes -2.080843 4.581517e-11    2.897809e-09 3.311782e-09    Yes
#> 9              Yes -2.087399 5.398326e-08    2.276294e-06 2.731553e-06    Yes
#> 18             No -1.178492 1.639921e-03    3.771819e-02 3.951430e-02    Yes
#> 28             Yes  2.924717 2.169950e-12    1.829991e-10 2.195990e-10    Yes
#> 37             Yes -2.083503 2.422027e-07    8.753897e-06 1.021288e-05    Yes
#> 39             No -1.473636 5.208314e-04    1.317703e-02 1.317703e-02    Yes
#> 40             Yes  1.860578 5.042269e-06    1.594617e-04 1.700925e-04    Yes
#> 62             Yes  3.634658 7.270238e-15    1.839370e-12 1.839370e-12    Yes
#> 63             Yes  3.380144 3.431016e-13    4.340236e-11 4.340236e-11    Yes
#> 172            Yes -2.747680 4.734031e-08    2.276294e-06 2.661577e-06    Yes

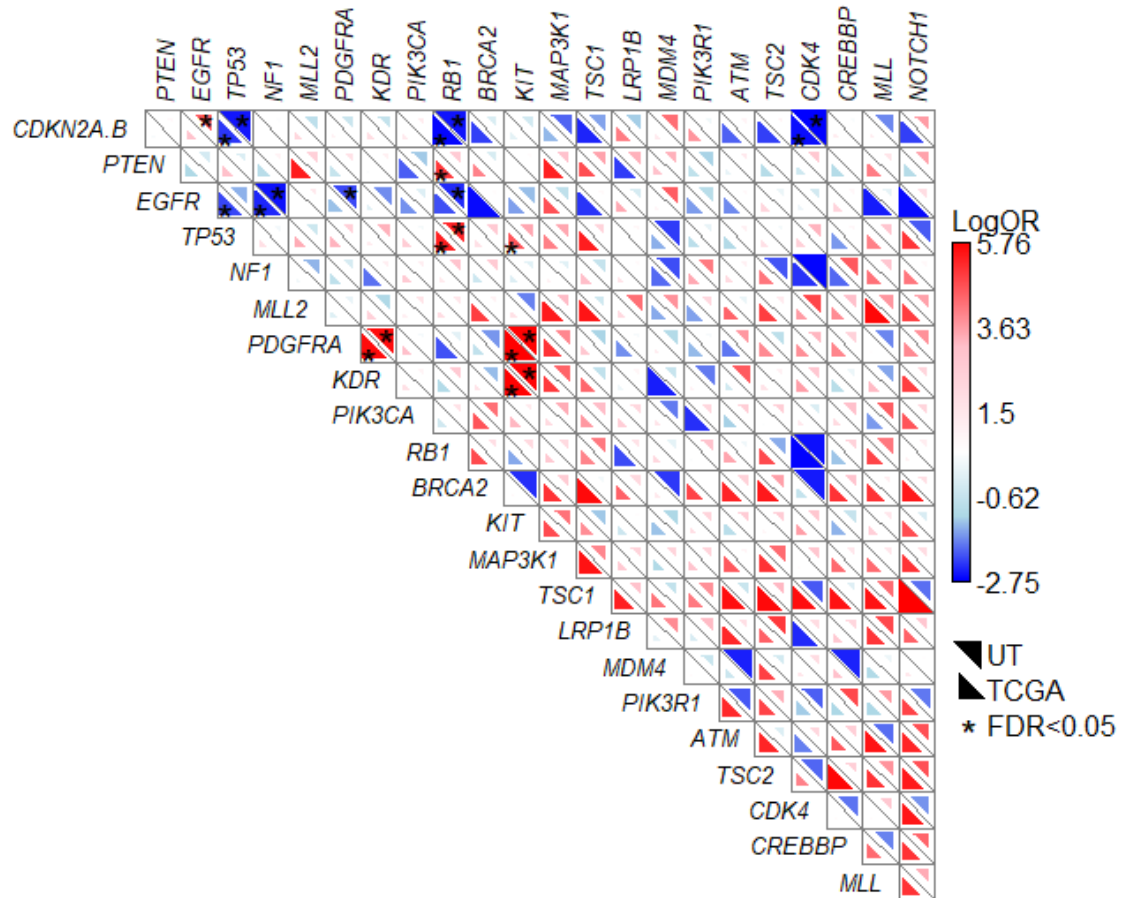
```

The result can be plotted as below.

```

DC.CO_plot(obj=sta_Sep,
            label.gene.cex=0.8,
            sig_CO.cex = 1.5)

```



Single mode

Assuming data from UT and TCGA cohort can be simply combined into one dataset, we analyze the cooccurrence and mutual exclusion of the combined dataset in "Single" mode.

```
#Remove "cohort" column
Combinedat<-gbm_dat[,-1]
Result_Sin<-DC.CO_Evaluation(mutation_data=Combinedat,
                             mode="Single",
                             adjust.method="BH",
                             FDRcutoff=0.05)
```

Extract the significant gene pairs.

```
sta_Sin<-Stat.Extraction(obj=Result_Sin)
sig_Sin<-sta_Sin$Stat
#Extract the significant gene pairs
sig_Sin[sig_Sin$Sig=="Yes",]
```

#>	Gene1	Gene2	LogOR	Raw.P	FDR	Sig
#> 2	CDKN2A.B	EGFR	0.7954841	3.897596e-05	8.964471e-04	Yes
#> 4	CDKN2A.B	TP53	-1.6164220	9.918776e-15	5.018901e-13	Yes
#> 6	EGFR	TP53	-0.9012762	7.968430e-06	2.240014e-04	Yes
#> 9	EGFR	NF1	-1.8551484	2.064501e-12	7.461697e-11	Yes
#> 18	EGFR	PDGFRA	-0.8133775	1.058038e-03	1.574609e-02	Yes
#> 28	PDGFRA	KDR	3.0072510	5.120707e-23	4.318462e-21	Yes
#> 37	CDKN2A.B	RB1	-2.1150977	3.978109e-13	1.677436e-11	Yes
#> 38	PTEN	RB1	1.0277087	3.328159e-04	6.477109e-03	Yes
#> 39	EGFR	RB1	-1.2600496	1.504132e-05	3.805453e-04	Yes
#> 40	TP53	RB1	1.9569822	7.882355e-12	2.492795e-10	Yes
#> 59	TP53	KIT	1.0154784	5.966967e-04	1.006428e-02	Yes
#> 62	PDGFRA	KIT	4.1745884	5.565107e-36	1.407972e-33	Yes
#> 63	KDR	KIT	3.4944853	4.399828e-26	5.565782e-24	Yes
#> 72	MLL2	MAP3K1	1.2856844	4.326117e-03	4.975034e-02	Yes
#> 74	KDR	MAP3K1	1.3540230	1.325282e-03	1.764717e-02	Yes
#> 168	LRP1B	TSC2	1.7650912	1.148326e-03	1.614036e-02	Yes
#> 172	CDKN2A.B	CDK4	-2.4109296	1.768317e-16	1.118460e-14	Yes
#> 176	NF1	CDK4	-2.0068708	5.691149e-04	1.006428e-02	Yes
#> 181	RB1	CDK4	-3.0559316	2.917333e-04	6.150710e-03	Yes
#> 216	MLL2	MLL	1.6306401	6.859639e-04	1.084680e-02	Yes
#> 225	LRP1B	MLL	1.7115431	1.471609e-03	1.861585e-02	Yes
#> 250	TSC2	NOTCH1	1.8519995	3.734333e-03	4.498983e-02	Yes

The result is plotted.

```
DC.CO_plot(obj=sta_Sin,
            label.gene.cex=0.8)
```

