Vignette for NEM-Tar:a method for cancer regulatory network inference and prioritization of potential therapeutic targets

Yuchen ZHANG,Lina ZHU,Xin WANG

February 28, 2021

Contents

1 Introduction

Cancers are not single disease entities, but comprising multiple molecularly distinct subtypes, and the heterogeneity prevents precise selection of patients for optimized therapy. Dissecting cancer subtype-specific signaling pathways is crucial to pinpointing dysregulated genes for the prioritization of novel therapeutic targets. Nested effects models (NEMs) are a group of graphical models that encode subset relations between

observed downstream effects under perturbations to upstream signaling genes, providing a prototype for mapping the inner workings of the cell. In this study, we developed NEM-Tar, which extends the original NEMs to predict drug targets by incorporating causal information of (epi)genetic aberrations for signaling pathway inference. An information theory-based score, weighted information gain (WIG), was proposed to assess the impact of signaling genes on a specific downstream biological process. We will show in detail how to implement a toy example for singaling network inference as well as a real case study for prioritizing the potential therapeutic targets.

2 A brief overview of NEMs and NEM-Tar

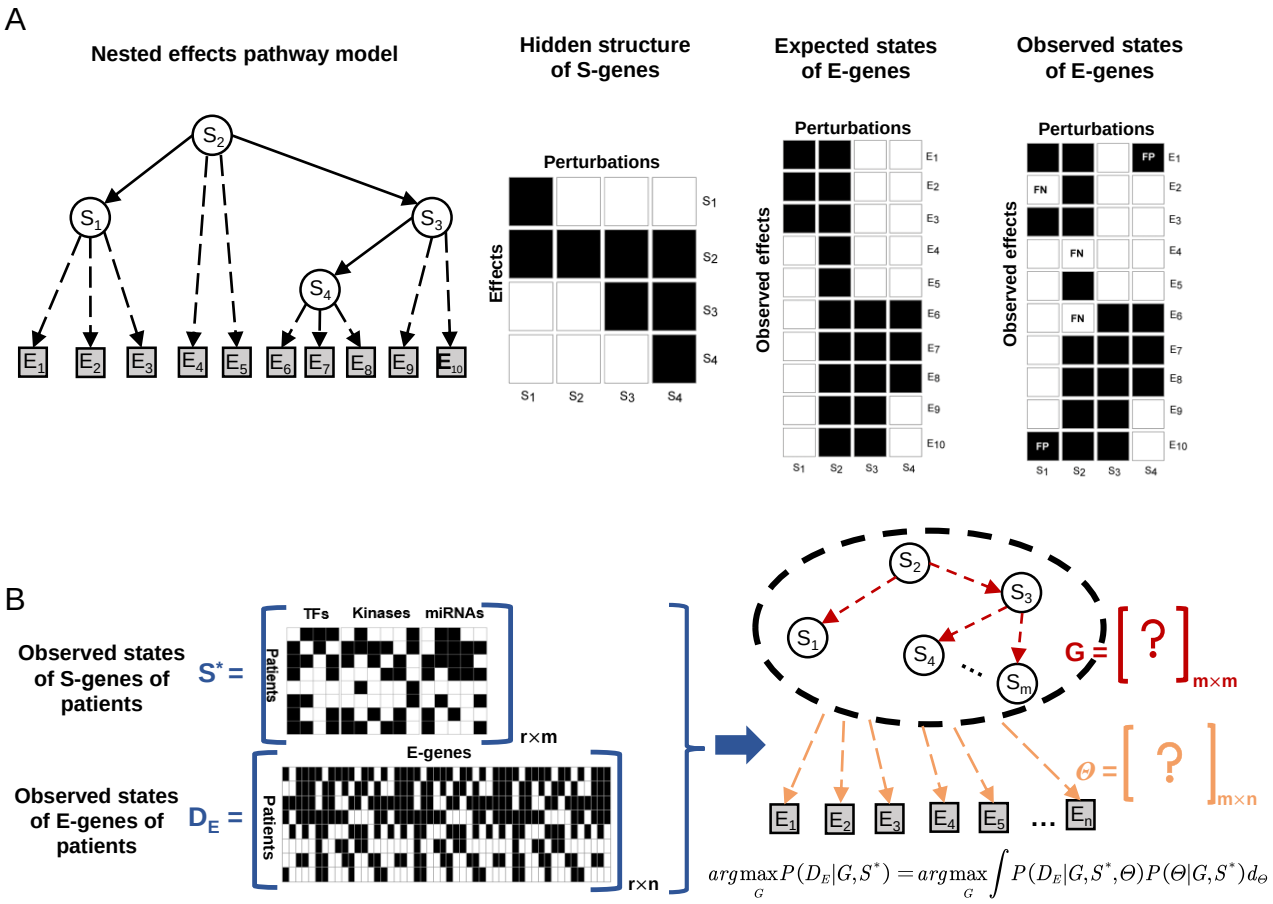


Figure 1. Comparison between the structures of (A) classic NEM and (B) our proposed NEM-Tar

As shown in Figure 1A, in classic nested effects models, the S-genes are modeled as hidden variables, and their signaling interaction graph G (solid arrows) is the target to infer. In experiments with perturbations to individual S-genes, differential expression of downstream genes could be observed and considered as effect reporter genes (E-genes). Assuming that each E-gene is directly regulated by at most one S-gene in G, the maximum a posteriori attachment Θ (dashed arrows) of effect genes to S-genes could be computed. The goal is to search for the signaling graph G, which yields the most likely probabilistic nested effects. Illustrated in Figure 1B, an extra observational dimension (the real patients) is the factor that NEM-Tar should deal with.The necessary adjustment should be conducted on the design and inference strategies of classic NEM. However, the information needs to infer is also the hidden interaction between S-genes and the attachment relationship of E-genes to S-genes.And the likelihood function of NEM-Tar is similar to that of NEMs, except the state matrix of regulators (S-genes) S* in our model.

3 Network inference on a toy example

3.1 Introduction of the in-silico data

The applicaitons on real case studies of NEM-Tar require a lot of data preprocssing and integrative selection of singaling genes(S-genes) and effect reporter genes(E-genes).For the purpose of interpreting the main work flow and contribution of NEM-Tar.At first, we will introduce the employed in-silico data.

```
library(nemTar)
```

```
## Loading required package: nem

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(dplyr)
data("example")
```

The toy example contains four different elements.The S-gene number was applied in medium size(12), thus the dimension of Sgene_hidden, storing the unobserved S-gene states was 12×12; Edata is the E-gene profiles with the dimension 804×100, S_obs is the observed profile of S-gene states 'after' perturbations,with the dimension of 100×12.

```
dim(Edata)
```

```
## [1] 804 100
```

```
dim(S_obs)
```

```
## [1] 100  12
```

```
dim(Sgene_hidden)
```

```
## [1] 12 12
```

```
para
```

```
## [1] 0.05 0.05
```

3.2 Network Inference Using Greedy hill-climbing

```
control<-nem::set.default.parameters(Sgenes=rownames(Sgene_hidden),type="mLL",para=para)
nemTar_rslt<-nem_Tar_greedy(Edata,Sgenes=control$Sgenes,S_obs,control=control)
```

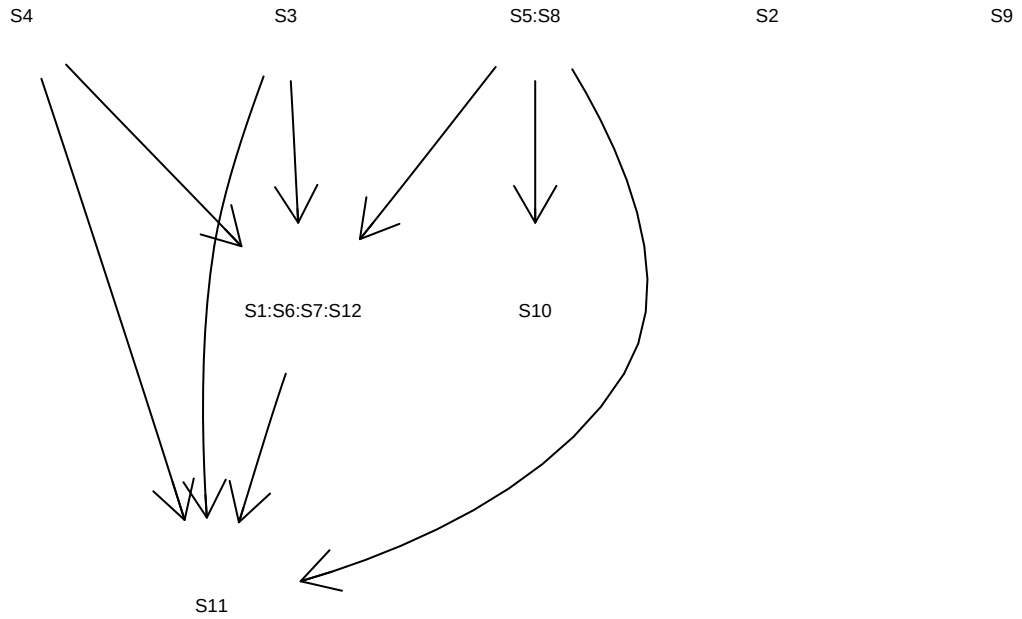
```
## Greedy hillclimber for 12 S-genes (lambda = 0 )...
##
## 132  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 132  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 130  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 129  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 128  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 123  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 117  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 115  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 108  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 66  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 53  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## --> Edge added, removed or reversed
## 49  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7 S
## log-likelihood of model =  -36204.99
```

This process takes a few seconds.

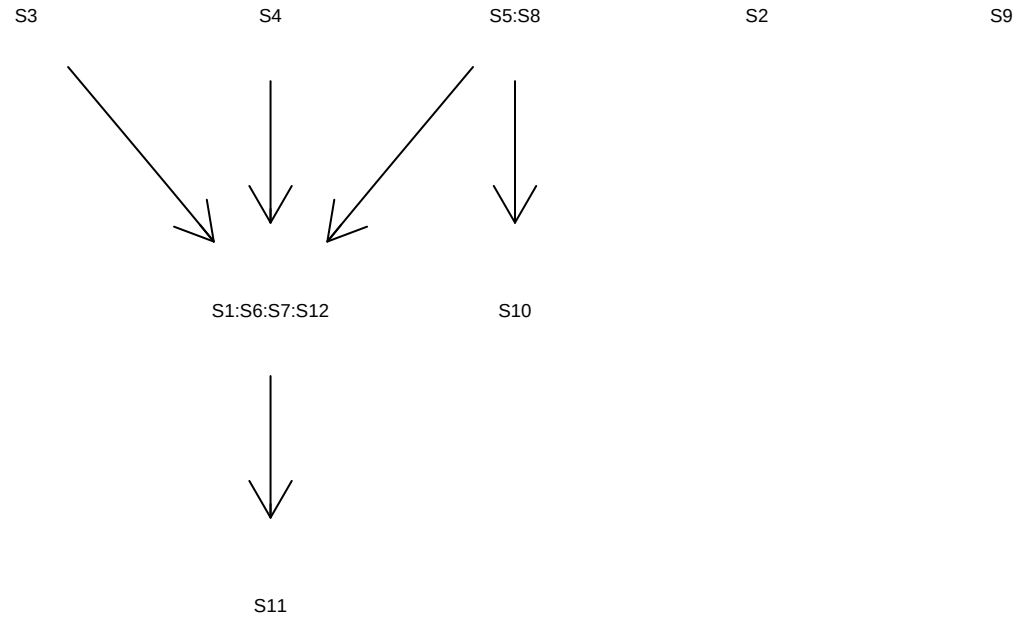
3.3 Visulization of the inferred S-gene network

To visualize the inference results, an R package RedeR is suggested.However, the function plot.nem could give a more quickly visulization.The error matrix is the difference between the inferred S-gene matrix and the generated S-gene matrix, the result that all the entries is 0 indicates that the inference is perfect.

```
plot.nem(nemTar_rslt,what="graph")
```




```
plot.nem(nemTar_rslt,what="graph",transitiveReduction=T)
```




```
adj<-as(nemTar_rslt$graph, "matrix")
error<-Sgene_hidden-adj
print(error)
```

```
##     S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12
## S1   0  0  0  0  0  0  0  0  0   0   0   0
## S2   0  0  0  0  0  0  0  0  0   0   0   0
## S3   0  0  0  0  0  0  0  0  0   0   0   0
## S4   0  0  0  0  0  0  0  0  0   0   0   0
## S5   0  0  0  0  0  0  0  0  0   0   0   0
## S6   0  0  0  0  0  0  0  0  0   0   0   0
## S7   0  0  0  0  0  0  0  0  0   0   0   0
## S8   0  0  0  0  0  0  0  0  0   0   0   0
## S9   0  0  0  0  0  0  0  0  0   0   0   0
## S10  0  0  0  0  0  0  0  0  0   0   0   0
## S11  0  0  0  0  0  0  0  0  0   0   0   0
## S12  0  0  0  0  0  0  0  0  0   0   0   0
```

4 Case study I-Inferring the Signaling Network Driving the EMT Subtype of Gastric Cancer and Prioritization of Potential Drug Targets

4.1 Introduction of the input profiles of GC

The profiles of GC contains three different elements.The S-gene number was determined as 14, thus the dimension of adjacency matrix of S-gene interacction network was 15×15; Edata_GC_ori and D are the E-gene profiles before and after the transformation to binary variable,with the dimension of 1194×38 and 824×28, respectively; Sdata_GC is the observed profile of 'natural' perturbation states of Sgenes with the dimension of 28×15.

```
# load the input profiles of GC
library(nemTar)
data("case_GC")
data("EMT_list")
```

4.2 Transforming the E-gene profiles into binary variable and network inference

```
res.disc <- nem.discretize(Edata_GC_ori,neg.control=1:8,pos.control=9:10,nfold=2,cutoff= 0.5)
```

```
## discretizing with respect to POS and NEG controls
```

```
D<-res.disc$dat
para<-res.disc$para
control<-set.default.parameters(Sgenes=Sgenes_GC,type="mLL",para=para)
nemTar_GC<-nem_Tar_greedy(D=D,Sgenes=Sgenes_GC,S_pattern=Sdata_GC,control=control)
```

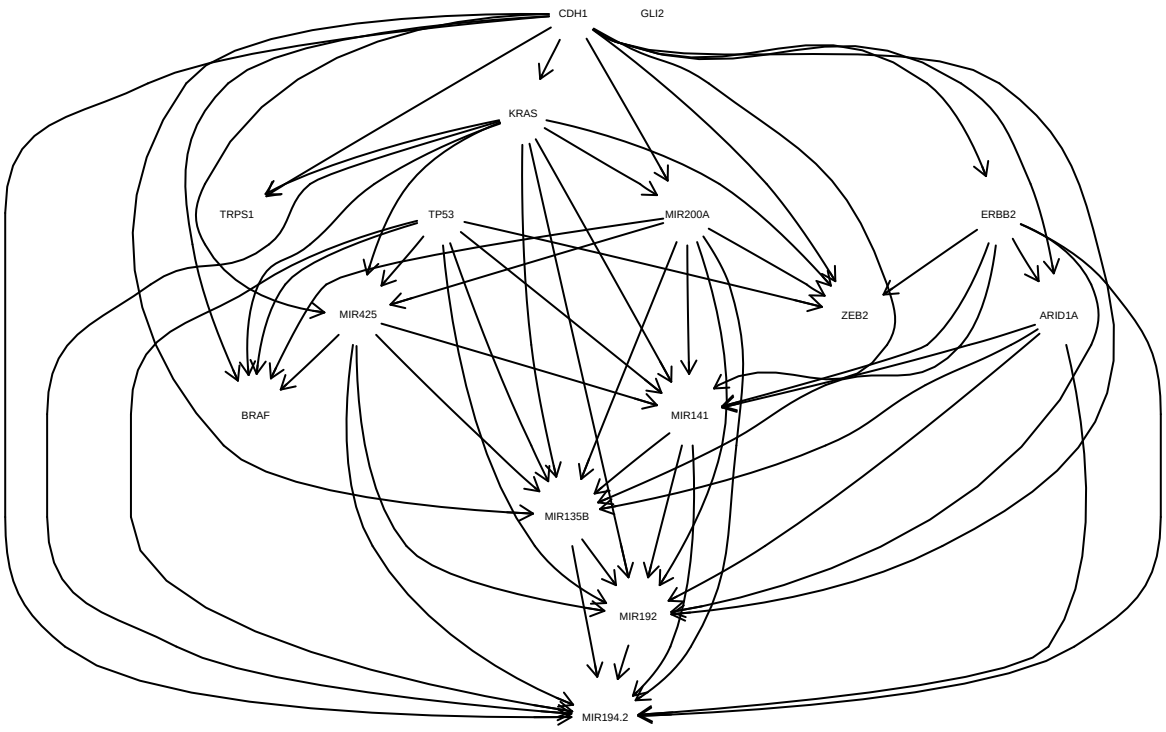
```
## Greedy hillclimber for 15 S-genes (lambda = 0 )...
##
## 210  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 210  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 209  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 207  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 205  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 202  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 196  local models to test ...
```

```
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 195  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 188  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 187  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 185  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 177  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 172  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 169  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 159  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 158  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 156  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 155  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## log-likelihood of model =  -17091.24
```

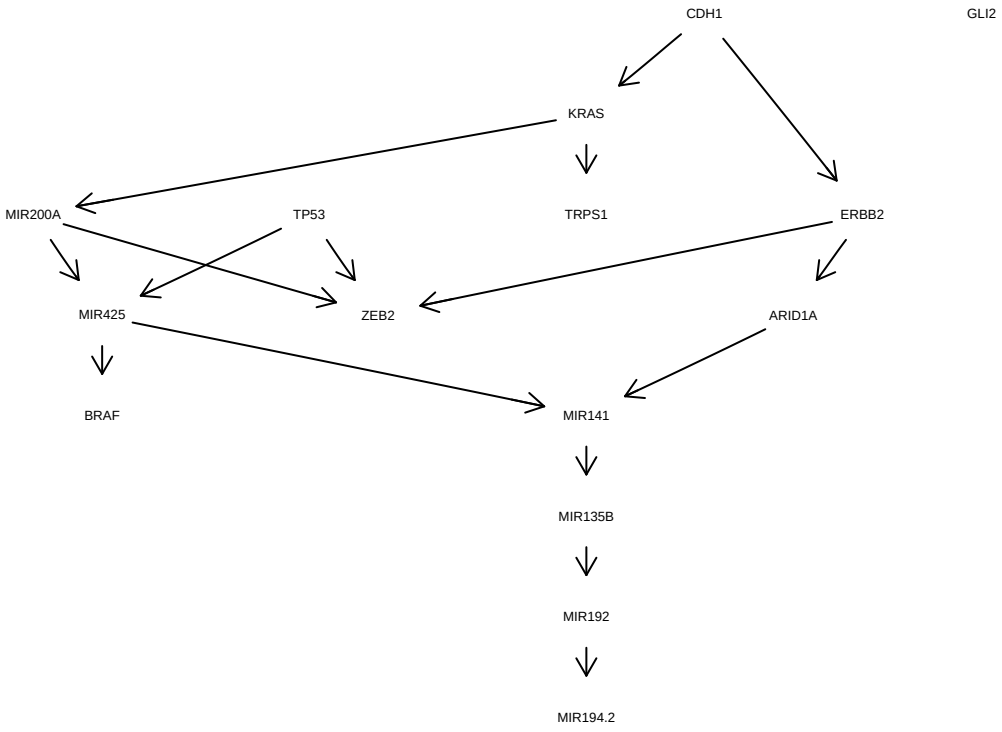
4.3 Visulization of the inferred S-gene network

4.3.1 Quick visulization

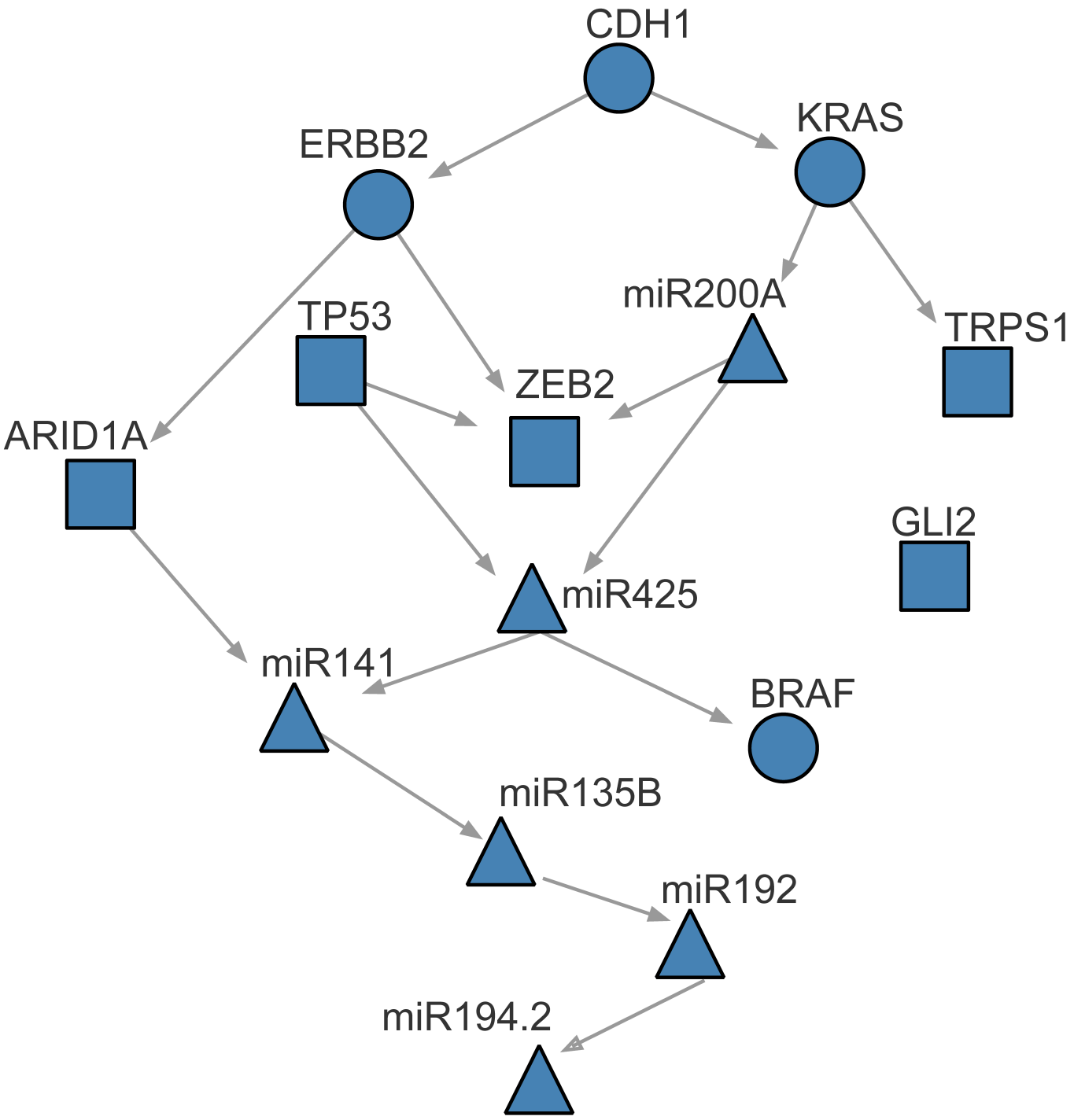
```
plot.nem(nemTar_GC,what="graph")
```




```
plot.nem(nemTar_GC,what="graph",transitiveReduction=T)
```


CDH1
GLI2
KRAS
MIR200A
TP53
TRPS1
ERBB2
MIR425
ZEB2
ARID1A
BRAF
MIR141
MIR135B
MIR192
MIR194.2

4.3.2 Employing the 'RedeR' package to visualize the network(after adjustment of the layout,color etc.)



4.4 Assessment of the influence of S-gene perturbations on EMT process

The statistical significance of the causal effect that the S-genes exert on downstream pathways(EMT pathway here) is quantified by permutation tests, i.e.,random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences.(The value 0 of the adjusted P stands for the corresponding P value is less than the resolution of current sampling times,in the following case P<1e-05.)

```
EMT_post<-path_post(nemTar_GC,EMT_list,Sgenes_GC)
WIG<-compute_WIG(EMT_post$post_affected,EMT_post$path_affected,15)
sample_WIG0<-WIG_sample(EMT_post,nemTar_GC,EMT_list,15,1e05)
sig_test<-WIGsig_test(sample_WIG0,WIG,1e05)
### results summary
sig_rslt<-data.frame(S_genes=names(EMT_post$Sig_affected)[which(lengths(EMT_post$path_affected)!=0)],
                     WIG=round(WIG$WIG,2),adjusted_P=format(round(sig_test,6),
                                                            scientific = TRUE,digits = 3))
colnames(sig_rslt)<-c("S genes","WIG","Adjusted P")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt,align="l",row.names=F,caption="Assessment of single S-gene perturbation
             on EMT in GC")
```

Table 1: Assessment of single S-gene perturbationon EMT in GC

S genes	WIG	Adjusted P
CDH1	66.15	0.00e+00
ERBB2	51.65	0.00e+00
GLI2	42.17	3.42e-01
ARID1A	35.29	2.31e-02
TP53	23.78	0.00e+00
KRAS	20.05	0.00e+00
MIR200A	15.02	4.20e-04
ZEB2	7.04	6.88e-01
MIR425	4.70	2.25e-03
BRAF	2.71	3.22e-01
MIR141	1.99	1.25e-01
MIR192	1.99	3.39e-01
MIR194.2	1.99	3.39e-01
MIR135B	1.99	2.53e-01
TRPS1	1.61	6.88e-01

Also, the weigheted information gain(WIG) of combinational perturbation of 2 S-genes could be calcualted.The kinases of double perturbation with higher WIGs could be the candidate combinational therapeutic targets.

```
Sgene_double_WIG<-WIG_double(EMT_post,15)
### results summary
sig_rslt<-data.frame(S_genes=names(Sgene_double_WIG$WIG),
                     WIG=round(Sgene_double_WIG$WIG,2))
colnames(sig_rslt)<-c("S genes","WIG")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt[match(c("CDH1/ERBB2","KRAS/CDH1","BRAF/CDH1","KRAS/ERBB2",
                              "BRAF/ERBB2","KRAS/BRAF"),rownames(sig_rslt)),],
```

```
            align="l",row.names=F,caption="Assessment of double perturbations
            (kinase only) on EMT in GC")
```

Table 2: Assessment of double perturbations (kinase only) on EMT in GC

S genes	WIG
CDH1/ERBB2	66.15
KRAS/CDH1	66.15
BRAF/CDH1	66.15
KRAS/ERBB2	59.12
BRAF/ERBB2	54.36
KRAS/BRAF	20.05

5 Case study II-Inferring the Signaling Network Driving the CMS4-mesenchymal Subtype of Colorectal Cancer(CRC) and Prioritization of Potential Drug Targets

5.1 Introduction of the input profiles of CRC

The profiles of GC contains three different elements.The S-gene number was prioritized as 15, thus the dimension of adjacency matrix of S-gene interacction network was 13×13; Edata_GC_ori and D are the E-gene profiles before and after the transformation to binary variable,with the dimension of 1337×96 and 1323×50,respectively; Sdata_CRC is the observed profile of 'natural' perturbation states of S-genes with the dimension 50×13.

```
library(nemTar)
data("case_CRC")
data("EMT_list")
```

5.2 Transforming the E-gene profiles into binary variable and network inference

```
res.disc <- nem.discretize(Edata_CRC_ori,neg.control=1:30,pos.control=31:46,nfold=2,cutoff=0.6)
```

```
## discretizing with respect to POS and NEG controls
```

```
D<-res.disc$dat
para<-res.disc$para
control<-set.default.parameters(Sgenes=Sgenes_CRC,type="mLL",para=para)
nemTar_CRC<-nem_Tar_greedy(D,Sgenes=control$Sgenes,Sdata_CRC,control=control)
```

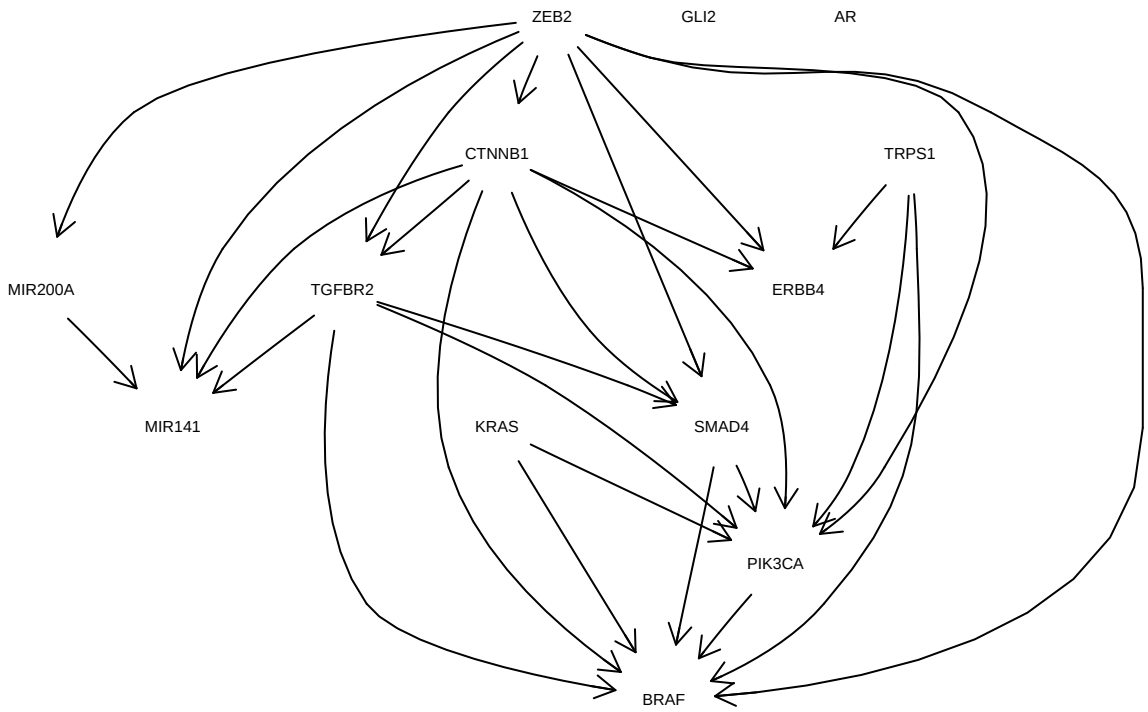
```
## Greedy hillclimber for 13 S-genes (lambda = 0 )...
##
## 156  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
```

```
## 156  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 154  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 153  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 152  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 149  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 140  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 139  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 136  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 134  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 132  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 131  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## --> Edge added, removed or reversed
## 130  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A KRAS BI
## log-likelihood of model =  -29580.53
```

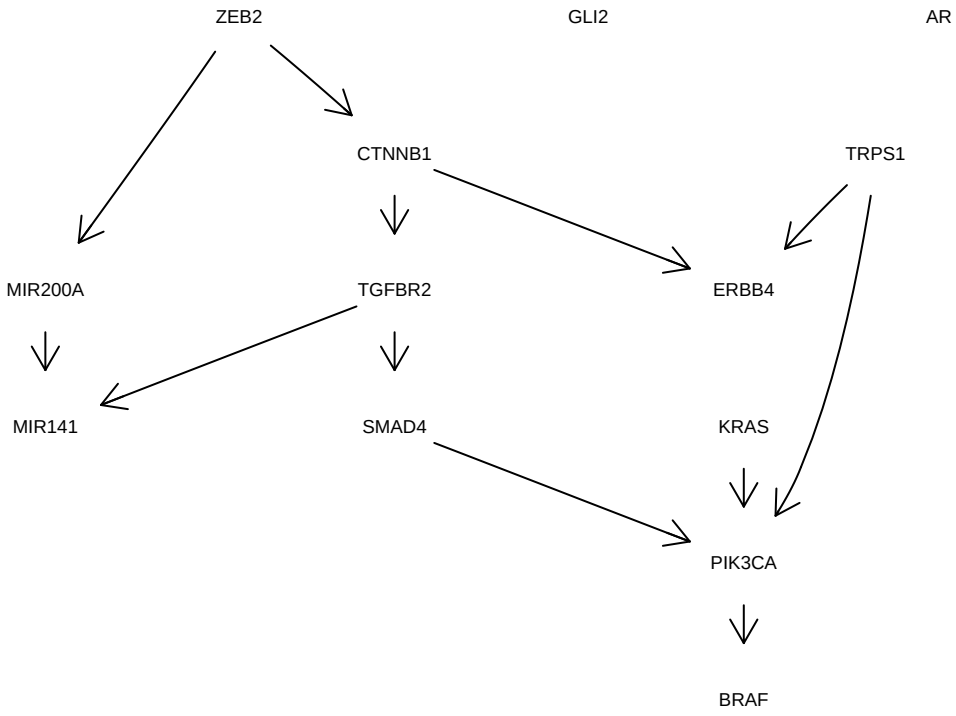
5.3 Visulization of the inferred S-gene network

5.3.1 Quick visulization

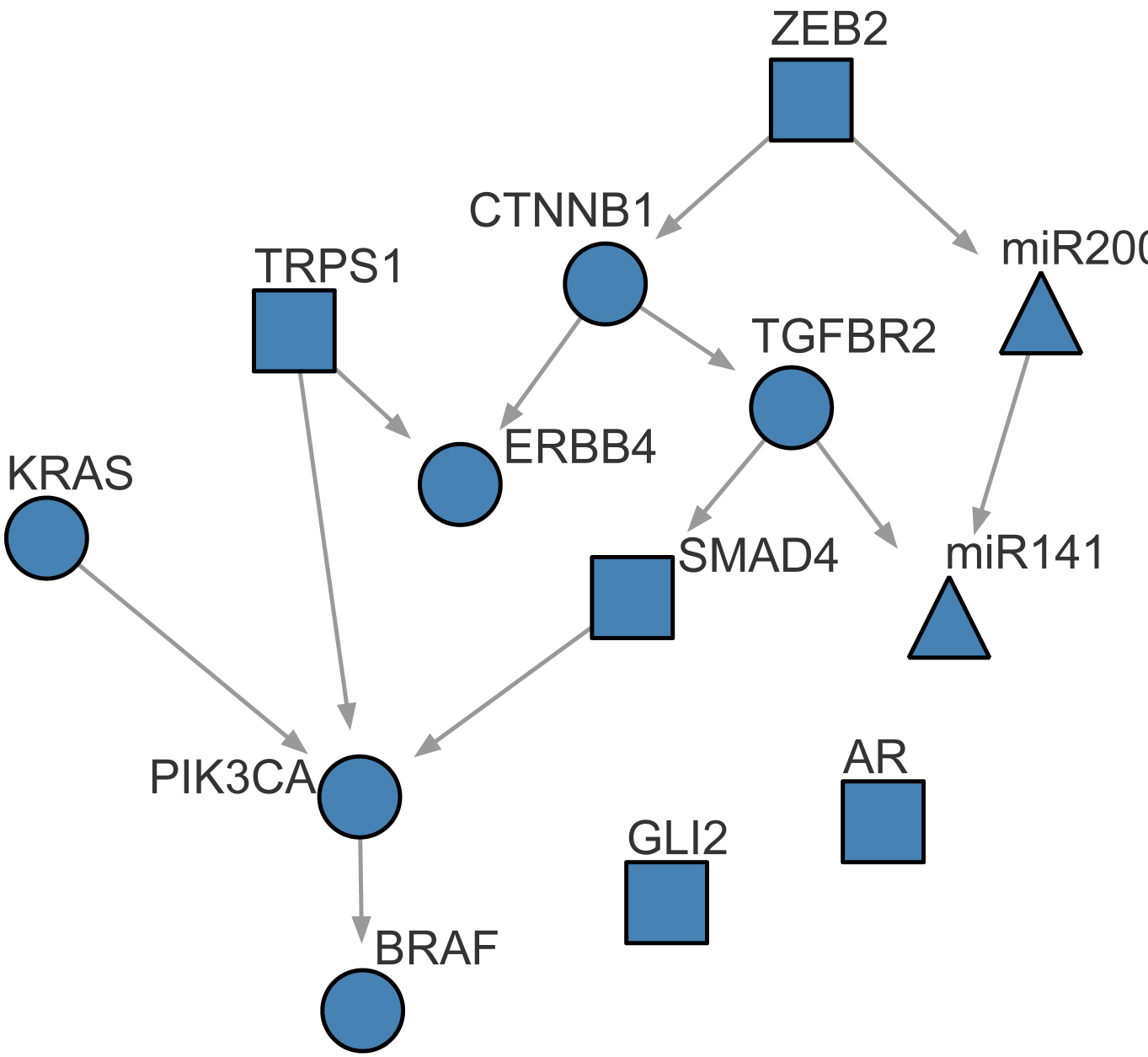
```
plot.nem(nemTar_CRC,what="graph")
```




```
plot.nem(nemTar_CRC,what="graph",transitiveReduction=T)
```


ZEB2
GLI2
AR
CTNNB1
TRPS1
MIR200A
TGFBR2
ERBB4
MIR141
SMAD4
KRAS
PIK3CA
BRAF

5.3.2 Employing the 'RedeR' package to visualize the network(after adjustment of the layout,color etc.)



5.4 Assessment of the influence of S-gene perturbations on EMT process

The statistical significance of the causal effect that the S-genes exert on downstream pathways(EMT pathway here) is quantified by permutation tests, i.e.,random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences.(The value 0 of the adjusted P stands for the corresponding P value is less than the resolution of current sampling times,in the following case P<1e-05.)

```
EMT_post<-path_post(nemTar_CRC,EMT_list,Sgenes_CRC)
WIG<-compute_WIG(EMT_post$post_affected,EMT_post$path_affected,13)
sample_WIG0<-WIG_sample(EMT_post,nemTar_CRC,EMT_list,13,1e05)
sig_test<-WIGsig_test(sample_WIG0,WIG,1e05)
### results summary
sig_rslt<-data.frame(S_genes=names(EMT_post$Sig_affected)[which(lengths(EMT_post$path_affected)!=0)],
                     WIG=round(WIG$WIG,2),adjusted_P=format(round(sig_test,6),
                                                            scientific = TRUE,digits = 3))
colnames(sig_rslt)<-c("Sgenes","WIG","Adjusted P")
# sig_rslt[which(sig_rslt[3]==0),3]<-"<1e-05"
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt,align="l",row.names=F,caption="Assessment of the impact of single S-gene
             perturbation on EMT in CRC")
```

Table 3: Assessment of the impact of single S-gene perturbation on EMT in CRC

Sgenes	WIG	Adjusted P
AR	39.38	5.15e-01
KRAS	38.61	1.51e-01
ZEB2	27.27	0.00e+00
CTNNB1	23.26	0.00e+00
GLI2	18.60	5.15e-01
TGFBR2	16.67	0.00e+00
TRPS1	13.22	1.82e-03
SMAD4	10.40	7.80e-03
PIK3CA	5.98	3.04e-02
BRAF	3.94	5.15e-01
MIR200A	3.54	1.51e-01
MIR141	3.16	5.24e-01
ERBB4	2.95	5.24e-01

Similar to study in GC, the weigheted information gain(WIG) of combinational perturbation of 2 Sgenes could be calcualted.The kinases of double perturbation with higher WIGs could be the candidate combinational therapeutic targets.

```
Sgene_double_WIG<-WIG_double(EMT_post,13)
### results summary
sig_rslt<-data.frame(S_genes=names(Sgene_double_WIG$WIG),
                     WIG=round(Sgene_double_WIG$WIG,2))
colnames(sig_rslt)<-c("S genes","WIG")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt[match(c("KRAS/CTNNB1","KRAS/TGFBR2","KRAS/ERBB4","KRAS/BRAF",
                              "KRAS/PIK3CA","BRAF/CTNNB1","PIK3CA/CTNNB1","TGFBR2/CTNNB1",
                              "ERBB4/CTNNB1","TGFBR2/ERBB4"),rownames(sig_rslt)),],
             align="l",row.names=F,caption="Assessment of double perturbations
             (kinase only) on EMT in CRC")
```

Table 4: Assessment of double perturbations (kinase only) on EMT in CRC

S genes	WIG
KRAS/CTNNB1	55.88
KRAS/TGFBR2	49.29
KRAS/ERBB4	41.56
KRAS/BRAF	38.61
KRAS/PIK3CA	38.61
BRAF/CTNNB1	23.26
PIK3CA/CTNNB1	23.26
TGFBR2/CTNNB1	23.26
ERBB4/CTNNB1	23.26
TGFBR2/ERBB4	19.62