CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation, and visualization
Supplementary material

TAOSHENG XU[1], THUC LE[2,3]

CONTENTS

[1] *Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China*
[2] *School of Information Technology and Mathematical Sciences, University of South Australia, Australia*
[3] *Centre for Cancer Biology, University of South Australia, Adelaide, Australia*

LIST OF FIGURES

1 INTRODUCTION

The *CancerSubtypes* package is designed to assist with the identification and validation of cancer subtypes based on cancer genomic datasets. The package is implemented in R and is available as a Bioconductor package at `http://bioconductor.org/packages/CancerSubtypes/`. We provide a unified framework for analyzing cancer subtypes from raw data to result visualization. The main functions include genomic data pre-processing, feature selection methods, cancer subtypes identification and results validation. The workflow and the components of the *CancerSubtypes* package are presented in Figure S1.

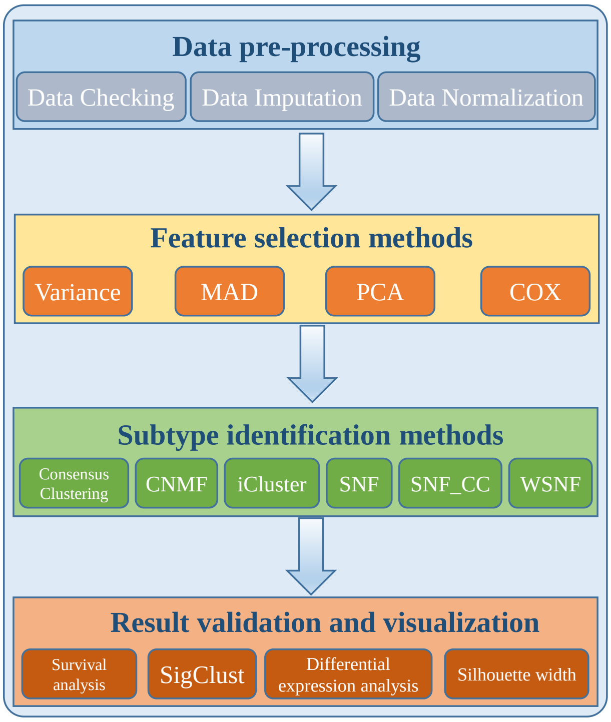

Figure S1: The workflow of *CancerSubtypes* package

The *CancerSubtypes* package has the following features:

- A framework/work flow to identify cancer subtypes and result validation and visualization.
- Unified input and output interface to perform and compare different cancer subtype discovery methods.
- 4 built-in feature selection methods for genomic dataset.
- 6 built-in algorithms for cancer subtypes identification.
- 4 built-in methods for result validation and visualization.

In the following sections, we present some typical scenarios of using the CancerSubtypes package with different purposes.

2 SCENARIO 1: A GENERAL EXAMPLE: USING CANCERSUBTYPES WITH TCGA DATA TO DISCOVER CANCER SUBTYPES

In this scenario, we present the usage of CancerSubtypes for discovering cancer subtypes with single genomic data type (gene expression data). Level 3 TCGA data can be downloaded and processed using **TCGAbiolinks** package (*Colaprico et al., TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data, 2016*) and **the TCGA Workflow** (*Silva et al., TCGA Workflow: Analyze Cancer Genomics and Epigenomics Data Using Bioconductor Packages, 2016*), or using the processed data in Bioconductor R package **RTCGA**.

2.1 Retrieve GDC online TCGA GBM gene expression data and clinical information from using *TCGAbiolinks*.

```
rm(list = ls())
##Install the latest version of TCGAbiolinks (Version:2.5.2)
devtools::install_github(repo = "BioinformaticsFMRP/TCGAbiolinks")
library("TCGAbiolinks")
library("SummarizedExperiment")
library("CancerSubtypes")
cancerType <- "GBM"
directory <- "./GDC/"
CancerProject <- paste0("TCGA-",cancerType)
DataDirectory <- paste0(directory,"GDC_",gsub("-","_",CancerProject))
FileNameData <- paste0(DataDirectory, "_","AgilentG4502A_07_1",".rda")

######GBM Gene expression Data1: AgilentG4502A_07_1########
query1 <- GDCquery(project = CancerProject,
                   data.category = "Gene␣expression",
                   data.type = "Gene␣expression␣quantification",
                   platform = "AgilentG4502A_07_1",
                   legacy = TRUE)
query_case1 <- query1$results[[1]]$cases
queryDown1 <- GDCquery(project = CancerProject,
                       data.category = "Gene␣expression",
                       data.type = "Gene␣expression␣quantification",
                       platform = "AgilentG4502A_07_1",
                       barcode = query_case1,
                       legacy = TRUE)
GDCdownload(queryDown1,directory = DataDirectory)
dataPrep1 <- GDCprepare(query = queryDown1,
                        save = TRUE,
                        directory =  DataDirectory,
                        save.filename = paste0(DataDirectory, "_",
                                     "AgilentG4502A_07_1",".rda"))
data1 <- assay(dataPrep1, 1)
##data imputation for missing measurements
data1=data.imputation(data1, fun = "microarray")

######GBM Gene expression Data2: AgilentG4502A_07_2########
query2 <- GDCquery(project = CancerProject,
                   data.category = "Gene␣expression",
                   data.type = "Gene␣expression␣quantification",
                   platform = "AgilentG4502A_07_2",
                   legacy = TRUE)
query_case2 <- query2$results[[1]]$cases
queryDown2 <- GDCquery(project = CancerProject,
                       data.category = "Gene␣expression",
                       data.type = "Gene␣expression␣quantification",
                       platform = "AgilentG4502A_07_2",
                       barcode = query_case2,
                       legacy = TRUE)
GDCdownload(queryDown2,directory = DataDirectory)
dataPrep2 <- GDCprepare(query = queryDown2,
                        save = TRUE,
                        directory =  DataDirectory,
                        save.filename = paste0(DataDirectory, "_",
                                     "AgilentG4502A_07_2",".rda"))
data2 <- assay(dataPrep2, 1)
data2=data.imputation(data2, fun = "microarray")
##combined two platform
GBM_mRNA=cbind(data1,data2)

###Extract the normal samples
index1=which(as.numeric(substr(colnames(GBM_mRNA),14,15))>9)
GBM_mRNA_Normal=GBM_mRNA[,index1]
###Extract the PRIMARY SOLID TUMOR("TP") samples
index2=which(substr(colnames(GBM_mRNA),14,15)=="01")
GBM_mRNA_Tumor=GBM_mRNA[,index2]
sampleName=substr(colnames(GBM_mRNA_Tumor),1,12)
######Remove the duplicated samples
index3=which(duplicated(sampleName))
for(i in index3)
{
  index3_3=which(sampleName==sampleName[i])
  GBM_mRNA_Tumor[,index3_3]=rowMeans(GBM_mRNA_Tumor[,index3_3])
}
GBM_mRNA_Tumor=GBM_mRNA_Tumor[,-index3]

# downloading and preparing clinical / survival data
query_clin <- GDCquery(project = CancerProject,
                       data.category = "Clinical")
clinical_case <- query_clin$results[[1]]$cases

queryDown_clin <- GDCquery(project = CancerProject,
                           data.category = "Clinical",
                           barcode = clinical_case)
GDCdownload(queryDown_clin)
clinical <- GDCprepare_clinic(queryDown_clin,
                                  clinical.info = "patient")
rownames(clinical) <- clinical$bcr_patient_barcode
GBM_clinical=clinical[,c("days_to_death",
                               "days_to_last_followup",
                               "vital_status")]

index4=which(is.na(GBM_clinical[,"days_to_death"]))
GBM_clinical[index4,"days_to_death"]=GBM_clinical[index4,
                                          "days_to_last_followup"]
status=as.vector(GBM_clinical[,"vital_status"])
status[status=="Alive"]=0
status[status=="Dead"]=1
GBM_clinical=cbind(GBM_clinical,"status"=as.numeric(status))
colnames(GBM_clinical)[1]="time"

####Exract the matched samples
intersect_samples=intersect(substr(colnames(GBM_mRNA_Tumor),1,12),
                            rownames(GBM_clinical))
index5=match(intersect_samples,substr(colnames(GBM_mRNA_Tumor),1,12))
index6=match(intersect_samples,rownames(GBM_clinical))
GBM_mRNA_Tumor=GBM_mRNA_Tumor[,index5]
GBM_clinical=GBM_clinical[index6,]
###Test the samples in gene expression dataset are matched
###with the samples in survival dataset
all(substr(colnames(GBM_mRNA_Tumor),1,12)==rownames(GBM_clinical))
save(GBM_mRNA_Normal,GBM_mRNA_Tumor,GBM_clinical,file="GBM_data.rda")
```

2.2 Apply Consensus Clustering method for cancer subtypes identification

```
####check distribution
data.checkDistribution(GBM_mRNA_Tumor)
###Feature selection by most variance
GBM_mRNA_Tumor1=FSbyVar(GBM_mRNA_Tumor, cut.type = "topk",4000)
index7=match(rownames(GBM_mRNA_Tumor1),rownames(GBM_mRNA_Normal))
GBM_mRNA_Normal1=GBM_mRNA_Normal[index7,]
##data normalization
GBM_mRNA_Tumor_norm=data.normalization(GBM_mRNA_Tumor1)
#####Concensus clustering
result1=ExecuteCC(clusterNum=3,d=GBM_mRNA_Tumor_norm,maxK=5,
                  clusterAlg="hc",distance="pearson",title="GBM")
group=result1$group
table(group)

##***************************************************
##group
##  1   2   3
##203 182 184
```

2.3 The Validation and visualization for the identified cancer subtypes result

2.3.1 *Survival analysis and Silhouette width*

```
###result validation and visualization
distanceMatrix=result1$distanceMatrix
p_value=survAnalysis(mainTitle="GBM",GBM_clinical$time,
                     GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
saveFigure(foldername="GBM",filename="GBM",image_width=10,
           image_height=10,image_res=300)

##***************************************************
##GBM Cluster= 3   Call:
#survdiff(formula = Surv(time, status) ~ group)
##n=568, 1 observation deleted due to missingness.
##         N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 202      178      165     0.948     1.545
##group=2 182      133      151     2.051     3.186
##group=3 184      125      120     0.212     0.301
## Chisq= 3.3  on 2 degrees of freedom, p= 0.196
```

The result of survival analysis is shown in Figure S2 on the following page. It is NOT significant ($p\_value = 0.196$) of the identified cancer subtypes. So the Consensus Clustering method is not competent to identify cancer subtypes in this case. In order to take an analysis example, we continue to conduct further analysis for this result by ignoring the non-significant performance.

2.3.2 *Statistical significance of clustering(Sigclust)*

```
### Statistical significance of clustering Test
sigclustTest(GBM_mRNA_Tumor_norm,group, nsim=500, nrep=1, icovest=3)
##Test result
##          Subtype 1 Subtype 2 Subtype 3
##Subtype 1         1         0         0
##Subtype 2         0         1         0
##Subtype 3         0         0         1
```

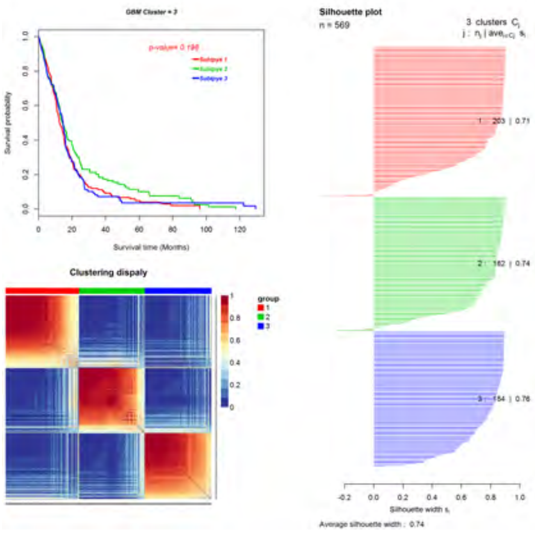

Figure S2: The result of Consensus Clustering for cancer subtypes of GBM based on gene expression data

The statistical significance test result is shown in below. It is a summary of the statistical significance of clustering p-value between the identified cancer subtypes. The sigClust summary plot between different subtypes are shown in the Figure S3.

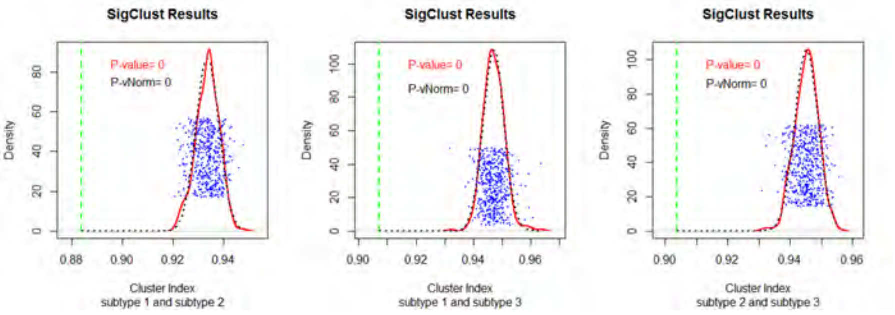

Figure S3: SigClust summary plot of simulated null distribution showing the statistical significance of clustering with respect to the simulated null distribution. The blue points, representing the simulated cluster index(CIs), are plotted with random vertical jitter for better visualization. The solid and dotted lines correspond to the estimated nonparametric density and Gaussian density fit to the simulated CIs[According to the description of Sigclust in the manuscript].

2.3.3 Differently Expression Analysis for the identified cancer subtypes

```
result2=DiffExp.limma(Tumor_Data=GBM_mRNA_Tumor1,
                      Normal_Data=GBM_mRNA_Normal1,
                      group=group,topk=NULL,RNAseq=FALSE)
## Top 6 differentially expressed genes in Subtype 1
head(result2[[1]])
##            ID      logFC   AveExpr         t       P.Value
##3854      KLK7 -4.526703 0.1773186 -41.51811 1.013013e-105
##2937     C1QL3 -4.107054 0.3597323 -27.74592  1.338856e-73
##983   KIAA1239 -5.792297 0.8571020 -27.63534  2.644676e-73
##2121     KCNV1 -4.626338 0.5141012 -27.14727  5.433236e-72
##2078     LRTM2 -3.515683 0.5212729 -24.75350  2.296553e-65
##2774    ATP2B3 -3.279332 0.5779107 -24.22061  7.563377e-64
##      adj.P.Val        B
## 4.052053e-102 228.9823
##  2.677713e-70 156.7746
##  3.526235e-70 156.1047
##  5.433236e-69 153.1297
##  1.837242e-62 138.0925
##  5.042251e-61 134.6439

##Extract top 1500 differentially expressed genes in each subtypes
Subtype1_gene=as.vector(na.omit(result2[[1]]$ID[1:1500]))
Subtype2_gene=as.vector(na.omit(result2[[2]]$ID[1:1500]))
Subtype3_gene=as.vector(na.omit(result2[[3]]$ID[1:1500]))
library(VennDiagram)
library(VennDiagram)
venn.diagram(filename="GBM_limma.png",height=3000,width=3300,
             list(Subytpe1=Subtype1_gene,Subytpe2=Subtype2_gene,
                  Subytpe3=Subtype3_gene),
             fill=c("red","green","blue"),
             alpha=c(0.5,0.5,0.5), cex=1, cat.fontface=4,
             cat.col = c("dodgerblue", "goldenrod1", "seagreen3"
                         ),cat.cex = 1,margin = 0.1,fontfamily=2)
```

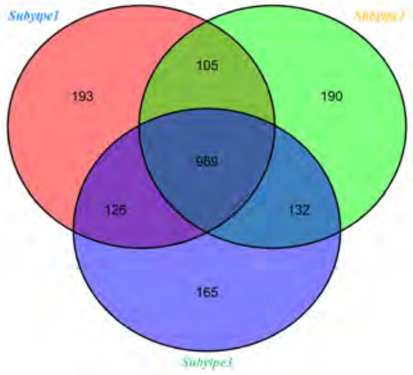

Figure S4: The Venn Diagram for the top 1500 differentially expressed genes in the three identified GBM subtypes

3 SCENARIO 2: INVESTIGATING THE IMPACT OF DIFFERENT FEATURE SELECTION METHODS IN CANCER SUBTYPE IDENTIFICATION

In this scenario, we investigate the impact of different feature selection methods in cancer subtype identification. We choose similarity fusion network(SNF) method for the cancer subtypes identification. We have processed a glioblastoma multiforme (GBM) multi-omics dataset from TCGA which includes the matched samples with gene expression, DNA methylation, miRNA expression data and survival data. These experiment datasets can be downloaded in https://github.com/xtsvm/ExperimentData/tree/master/GBM. Meanwhile, these experiment datasets are also used for next scenarios (Scenario 3 and Scenario 4).

3.1 The original SNF method without feature selection

```
library("CancerSubtypes")
load("GBM_GeneEXp.rda")
load("GBM_Methylation27.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
##origninal
GBM=list(GeneExp=GBM_GeneEXp,
         DNAmethy=GBM_Methylation27,
         miRNAExp=GBM_miRNA_8x15k)
result3 =ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result3$group
distanceMatrix=result3$distanceMatrix
p_value=survAnalysis(mainTitle="GBM␣Original",GBM_clinical$time,
                      GBM_clinical$status,group,
                      distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM Original Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##            N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 200      165    138.8    4.9629    14.611
##group=2   4        3      2.5    0.0996     0.101
##group=3  72       52     78.7    9.0814    15.427
##Chisq= 15.4  on 2 degrees of freedom, p= 0.000447
```

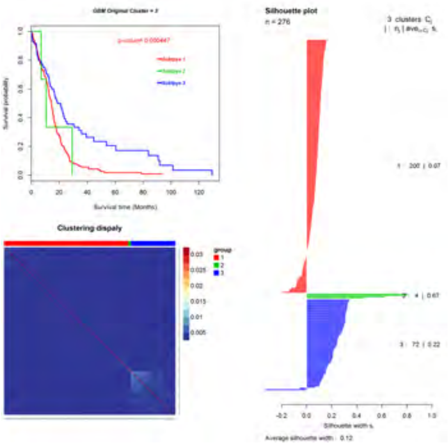

Figure S5: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM

3.2 Feature selection by most variance to select important features for cancer subtypes identification

```
par(mfrow=c(1,3))
GBM_GeneEXp_FsbyVar=FSbyVar(GBM_GeneEXp, cut.type = "cutoff", value=1)
GBM_Methylation27_FsbyVar=FSbyVar(GBM_Methylation27,cut.type = "cutoff", value=0.01)
GBM_miRNA_8x15k_FsbyVar=FSbyVar(GBM_miRNA_8x15k,cut.type = "cutoff", value=0.2)
```

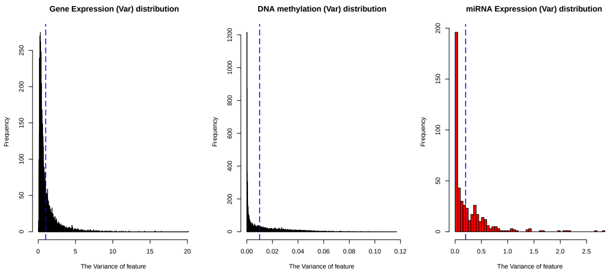

Figure S6: The variance distribution and the feature selection cutoff for different platform GBM data [The cutoff values are based on the distribution characteristic of each data]

```
GBM=list(GeneExp=GBM_GeneEXp_FsbyVar,DNAmethy=GBM_Methylation27_FsbyVar,
         miRNAExp=GBM_miRNA_8x15k_FsbyVar)
GBM=list(GeneExp=GBM_GeneEXp_FsbyVar,DNAmethy=GBM_Methylation27_FsbyVar,
         miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result4 =ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result4$group
distanceMatrix=result4$distanceMatrix
p_value=survAnalysis(mainTitle="GBM_FSbyVar",GBM_clinical$time,GBM_clinical$status,
                     group,distanceMatrix=distanceMatrix,similarity=TRUE)
## *****************************************************
##GBM FSbyVar Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 198      163   136.33     5.219    14.896
##group=2   5        4     2.65     0.689     0.701
##group=3  73       53    81.02     9.693    16.717
##Chisq= 17  on 2 degrees of freedom, p= 0.000207
```

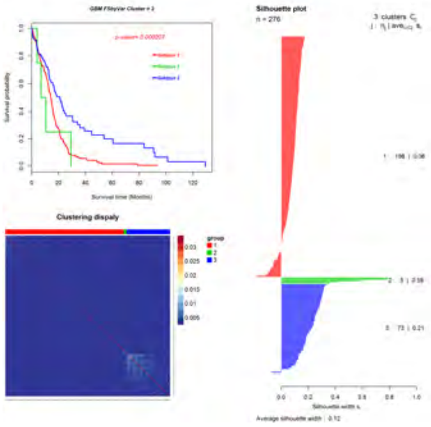

Figure S7: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM

3.3 Feature selection by most Median Absolute Deviation (MAD) to select important features for cancer subtypes identification

```
par(mfrow=c(1,3))
GBM_GeneExp_FsbyMAD=FSbyMAD(GBM_GeneExp,cut.type = "cutoff", value=0.6)
GBM_Methylation27_FsbyMAD=FSbyMAD(GBM_Methylation27,cut.type = "cutoff", value=0.05)
GBM_miRNA_8x15k_FsbyMAD=FSbyMAD(GBM_miRNA_8x15k,cut.type = "cutoff", value=0.2)
```

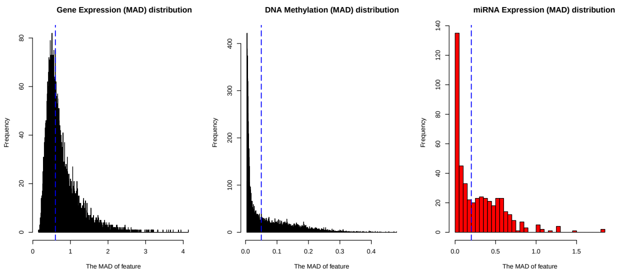

Figure S8: The MAD distribution and the feature selection cutoff for different platform GBM data [The cutoff values are based on the distribution characteristic of each data]

```
GBM=list(GeneExp=GBM_GeneExp_FsbyMAD,DNAmethy=GBM_Methylation27_FsbyMAD,
         miRNAExp=GBM_miRNA_8x15k_FsbyMAD)
result5 =ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result5$group
distanceMatrix=result5$distanceMatrix
p_value=survAnalysis(mainTitle="GBM␣FSbyMAD",GBM_clinical$time,GBM_clinical$status,
                        group,distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM FSbyMAD Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 196      161   134.67     5.146    14.326
##group=2  75       55    82.68     9.265    16.083
##group=3   5        4     2.65     0.689     0.701
##Chisq= 16.3  on 2 degrees of freedom, p= 0.000283
```

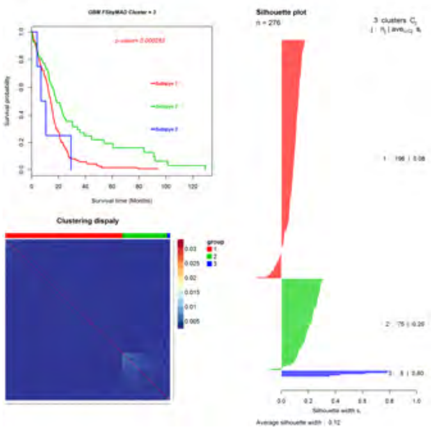

Figure S9: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM

3.4 Feature selection by COX model to select important features for cancer subtypes identification

```
library("CancerSubtypes")
load("GBM_GeneEXp.rda")
load("GBM_Methylation27.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM_GeneEXp_FsbyCox=FSbyCox(GBM_GeneEXp, GBM_clinical$time,
                            GBM_clinical$status, cutoff = 0.05)
GBM_Methylation27_FsbyCox=FSbyCox(GBM_Methylation27, GBM_clinical$time,
                                  GBM_clinical$status, cutoff = 0.05)
GBM_miRNA_8x15k_FsbyCox=FSbyCox(GBM_miRNA_8x15k, GBM_clinical$time,
                                GBM_clinical$status, cutoff = 0.05)
GBM=list(GeneExp=GBM_GeneEXp_FsbyCox,
         DNAmethy=GBM_Methylation27_FsbyCox,
         miRNAExp=GBM_miRNA_8x15k_FsbyCox)
result6 =ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result6$group
distanceMatrix=result6$distanceMatrix
p_value=survAnalysis(mainTitle="GBM_FSbyCox",GBM_clinical$time,
                     GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM FSbyCox Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 207      170   138.15   7.34489  21.95949
##group=2   3        2     2.11   0.00611   0.00621
##group=3  66       48    79.74  12.63403  22.07171
##Chisq= 22.3  on 2 degrees of freedom, p= 1.46e-05
```

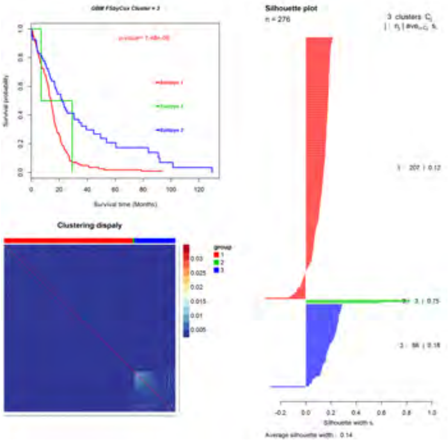

Figure S10: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM

3.5 Feature selection by PCA to select important features for cancer subtypes identification

```
library("CancerSubtypes")
load("GBM_GeneEXp.rda")
load("GBM_Methylation27.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM_GeneEXp_FsbyPCA=FSbyPCA(GBM_GeneEXp,
                            PC_percent = 0.85,scale = TRUE)
GBM_Methylation27_FsbyPCA=FSbyPCA(GBM_Methylation27,
                                  PC_percent = 0.85, scale = TRUE)
GBM_miRNA_8x15k_FsbyPCA=FSbyPCA(GBM_miRNA_8x15k,
                                PC_percent = 0.85, scale = TRUE)
GBM=list(GeneExp=GBM_GeneEXp_FsbyPCA,
         DNAmethy=GBM_Methylation27_FsbyPCA,
         miRNAExp=GBM_miRNA_8x15k_FsbyPCA)
result7 =ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result6$group
distanceMatrix=result7$distanceMatrix
p_value=survAnalysis(mainTitle="GBM_FSbyPCA",GBM_clinical$time,
                     GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM FSbyPCA Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E  (O-E)^2/V
##group=1 200      165   137.25   5.61188   16.26134
##group=2  73       53    80.64   9.47331   16.31884
##group=3   3        2     2.11   0.00611    0.00621
##Chisq= 16.5  on 2 degrees of freedom, p= 0.000265
```

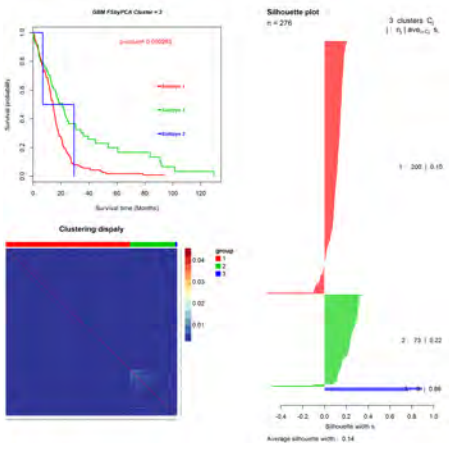

Figure S11: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM

3.6 The comparison of different feature selection methods

According to the survival analysis and Silhouette width, the feature selection by Cox model outperforms than other methods.

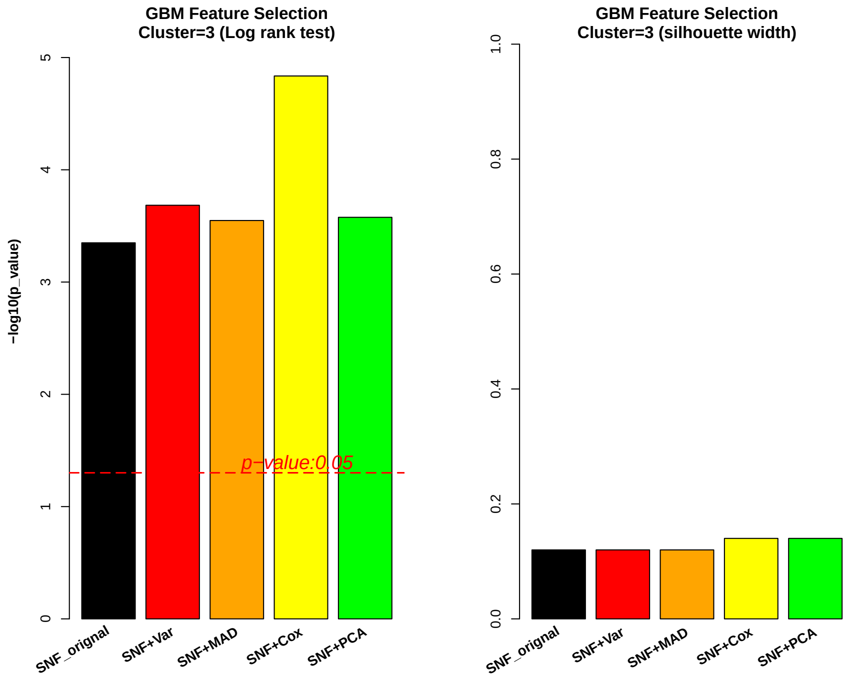

Figure S12: The barplot for the Log-rank test p-values and Silhouette width of each feature selection method

4 Scenario 3: Comparing the performance of different cancer subtype identification methods

In this scenario, we present the experiments samples for cancer subtypes identification by using different clustering methods. The CC and CNMF are designed for single-genomic datasets (e.g. gene expression datasets), while iCluster, SNF and SNF-CC focus on multi-omics data analysis. To make a fair comparison, we try to use the same input dataset for the different clustering methods comparison. The GBM gene expression and miRNA expression datasets are chosen for the experiment analysis. We concatenated the gene expression data and miRNA expression data for each patient as the input data for CC and CNMF.

4.1 Identify cancer subtypes by using Consensus Clustering(CC)

```
load("GBM_GeneEXp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
##The input dataset is multi-genomics data as a list
GBM=list(GeneExp=GBM_GeneEXp,miRNAExp=GBM_miRNA_8x15k)
result8 =ExecuteCC(clusterNum=3,d=GBM,maxK=3,clusterAlg="hc",
                   distance="pearson",title="GBM")
group=result8$group
distanceMatrix=result8$distanceMatrix
p_value=survAnalysis(mainTitle="GBM␣Consensus␣Clustering-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
##***************************************************
##GBM Consensus Clustering-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1   58       56     65.5    1.3755     2.112
##group=2 214      161    152.0    0.5320     1.848
##group=3   4        3      2.5    0.0996     0.101
##Chisq= 2.2  on 2 degrees of freedom, p= 0.339
```

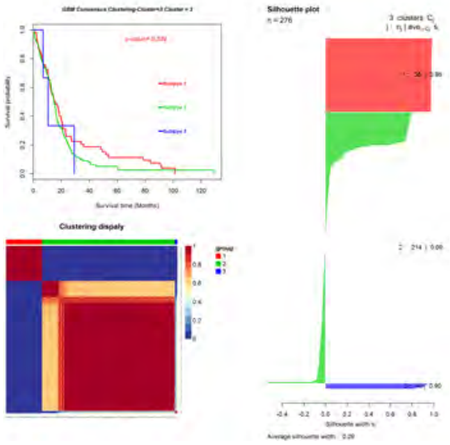

Figure S13: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM (Consensus clustering result)

4.2 Identify cancer subtypes by using Consensus Nonnegative matrix factorization(CNMF)

```
load("GBM_GeneExp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM_GeneExp_FsbyVar=FSbyVar(GBM_GeneExp, cut.type = "cutoff", value=1)
GBM_miRNA_8x15k_FsbyVar=FSbyVar(GBM_miRNA_8x15k,
                                cut.type = "cutoff", value=0.2)
##The input dataset is multi-genomics data as a list
GBM=list(GeneExp=GBM_GeneExp_FsbyVar,miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result9 =ExecuteCNMF(GBM,clusterNum=3,nrun=30)
group=result9$group
distanceMatrix=result9$distanceMatrix
p_value=survAnalysis(mainTitle="GBM␣CNMF-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM CNMF-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E  (O-E)^2/V
##group=1 137      108     88.7      4.22       7.60
##group=2  81       56     65.9      1.47       2.14
##group=3  58       56     65.5      1.38       2.11
##Chisq= 7.6  on 2 degrees of freedom, p= 0.0224
```

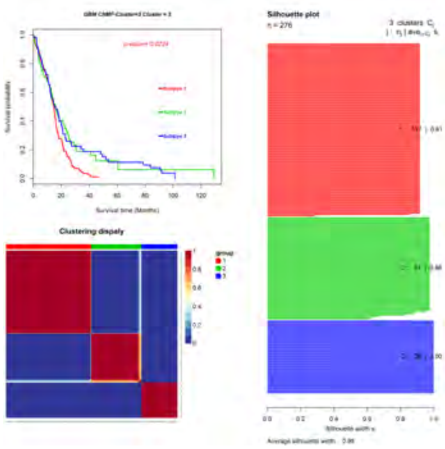

Figure S14: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM (CNMF result)

4.3 Identify cancer subtypes by using Integrative clustering of multiple genomic data(iCluster)

```
load("GBM_GeneExp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
##For iCluster algorithm, it cannot process high-dimensional data,
##otherwise it is very very time-consuming or reports a mistake.
##We choose top 2000 most variance genes and top 500 most
##variance miRNAs for analysis
GBM_GeneExp_FsbyVar=FSbyVar(GBM_GeneExp, cut.type = "topk", value=2000)
GBM_miRNA_8x15k_FsbyVar=FSbyVar(GBM_miRNA_8x15k,
                                cut.type = "topk", value=300)
GBM=list(GeneExp=GBM_GeneExp_FsbyVar,miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result10 =ExecuteiCluster(datasets=GBM, k=3, lambda=list(0.44,0.33))
group=result10$group
p_value=survAnalysis(mainTitle="GBM␣iCluster-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group)

##*****************************************************
##GBM iCluster-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1  58       56     65.5     1.376     2.112
##group=2  79       61     64.4     0.181     0.261
##group=3 139      103     90.1     1.848     3.271
##Chisq= 3.6  on 2 degrees of freedom, p= 0.164
```

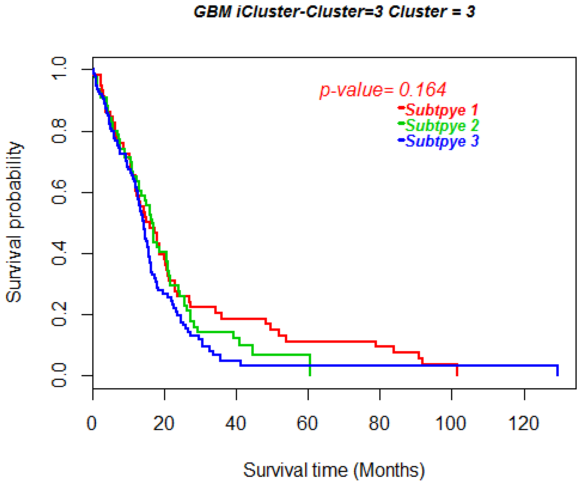

Figure S15: The Survival curves for the identified cancer subtypes of GBM (iCluster result)

4.4 Identify cancer subtypes by using Similarity Network Fusion (SNF)

```
load("GBM_GeneExp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM=list(GeneExp=GBM_GeneExp,miRNAExp=GBM_miRNA_8x15k)
result11=ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result11$group
distanceMatrix=result11$distanceMatrix
p_value=survAnalysis(mainTitle="GBM_SNF-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)

##*****************************************************
##GBM SNF-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 199      163    143.8    2.5514     7.679
##group=2  73       54     73.7    5.2456     8.226
##group=3   4        3      2.5    0.0996     0.101
##Chisq= 8.2  on 2 degrees of freedom, p= 0.0163
```

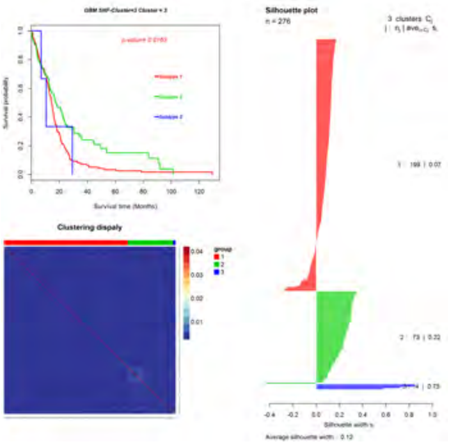

Figure S16: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM (SNF result)

4.5 Identify cancer subtypes by combining the SNF and CC

```
load("GBM_GeneExp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM=list(GeneExp=GBM_GeneExp,miRNAExp=GBM_miRNA_8x15k)
result12=ExecuteSNF.CC(GBM, clusterNum=3, K=20, alpha=0.5, t=20,
                     maxK = 5, pItem = 0.8,reps=500,
                     title = "GBM", plot = "png",
                     finalLinkage ="average")
group=result12$group
distanceMatrix=result12$distanceMatrix
p_value=survAnalysis(mainTitle="GBM_SNF.CC-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)

##*******************************************************
##GBM SNF.CC-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 182      146    130.9      1.74      4.45
##group=2  73       54     73.7      5.25      8.23
##group=3  21       20     15.5      1.34      1.45
##Chisq= 8.7  on 2 degrees of freedom, p= 0.0131
```

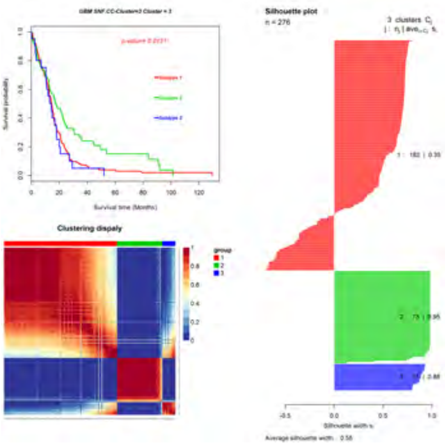

Figure S17: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM (SNF.CC result)

4.6 Identify cancer subtypes by Weighted Similarity Network Fusion

```
load("GBM_GeneEXp.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM=list(GeneExp=GBM_GeneEXp,miRNAExp=GBM_miRNA_8x15k)
###1. Use the defualt ranking in the package.
data(Ranking)
####Retrieve there feature ranking for genes
gene_Name=rownames(GBM_GeneEXp)
library(HGNChelper)
gene_Name_1=checkGeneSymbols(gene_Name)[,3]
index1=match(gene_Name,Ranking$mRNA_TF_miRNA.v21._SYMBOL)
gene_ranking=data.frame(gene_Name,Ranking[index1,],stringsAsFactors=FALSE)
index2=which(is.na(gene_ranking$ranking_default))
gene_ranking$ranking_default[index2]=min(gene_ranking$ranking_default,na.rm =TRUE)
####Retrieve there feature ranking for genes
miRNA_ID=rownames(GBM_miRNA_8x15k)
index3=match(miRNA_ID,Ranking$mRNA_TF_miRNA_ID)
miRNA_ranking=data.frame(miRNA_ID,Ranking[index3,],
                           stringsAsFactors=FALSE)
index4=which(is.na(miRNA_ranking$ranking_default))
miRNA_ranking$ranking_default[index4]=min(miRNA_ranking$ranking_default,na.rm =TRUE)
###Clustering
ranking1=list(gene_ranking$ranking_default, miRNA_ranking$ranking_default)
result13=ExecuteWSNF(datasets=GBM, feature_ranking=ranking1,beta=0.8,clusterNum=3,
                      K = 20,alpha = 0.5,t = 20, plot = TRUE)
group=result13$group
distanceMatrix=result13$distanceMatrix
p_value=survAnalysis(mainTitle="GBM␣WSNF-Cluster=3",
                     GBM_clinical$time,GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)
##*****************************************************
##GBM WSNF-Cluster=3 Cluster= 3   Call:
##survdiff(formula = Surv(time, status) ~ group)
##          N Observed Expected (O-E)^2/E (O-E)^2/V
##group=1 195      160    140.6    2.6824     7.807
##group=2  77       57     76.9    5.1577     8.344
##group=3   4        3      2.5    0.0996     0.101
## Chisq= 8.4  on 2 degrees of freedom, p= 0.0154
```

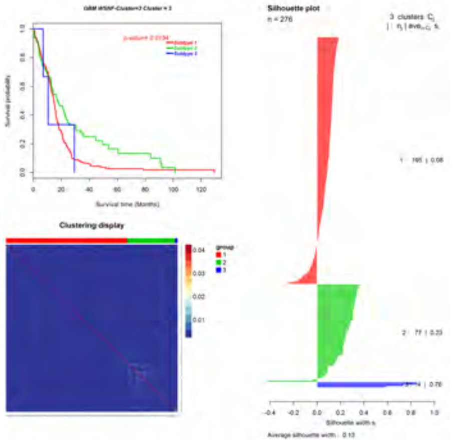

Figure S18: The Survival curves and Silhouette plots for the identified cancer subtypes of GBM (WSNF result)

4.7 The comparison of different cancer subtypes identification methods

The summary of the *Log-rank test p-value* for each cancer subtypes identification methods is shown in the Figure S19. We don't list the *Silhouette width* for comparison because the similarity matrix for each cancer subtypes identification method is in different numerical level. So the *Silhouette width* does not have a comparative meaning but can provide the important information for the insight investigation of the identified cancer subtypes. Figure S19 shows that SNF and its variants (SNF.CC and WSNF) perform the best in this dataset.

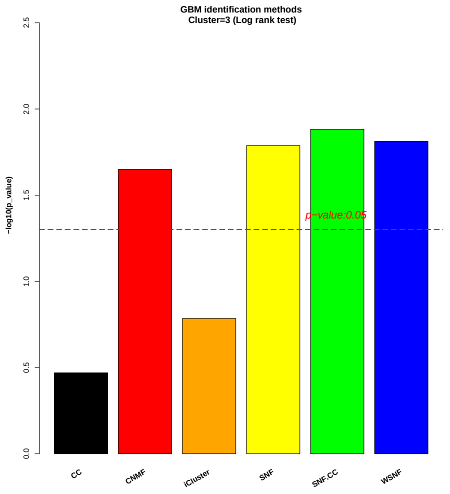

Figure S19: The barplot for the Log-rank test p-values of each cancer subtypes identification method

5 SCENARIO 4: INVESTIGATING THE IMPACT OF DIFFERENT GENOMIC DATA TYPES ALTERS THE RESULTS WITH THE SELECTED FEATURE SELECTION AND CANCER SUBTYPE IDENTIFICATION METHODS

For the experiments of cancer subtypes identification with different genomic data types and feature selection methods, we choose the Consensus Clustering(CC) and the Similarity Network Fusion for single dataset input and multiple datasets input clustering, respectively. To have a comprehensive comparison, We intend to conduct the six groups of experiments which are listed below.

(1) mRNA-Var-CC

(2) mRNA+miRNA-Var-CC

(3) mRNA+miRNA-Var-SNF

(4) mRNA+DM+miRNA-Var-SNF

(5) mRNA+miRNA-COX-SNF

(6) mRNA+DM+miRNA-COX-SNF

```
load("GBM_GeneEXp.rda")
load("GBM_Methylation27.rda")
load("GBM_miRNA_8x15k.rda")
load("GBM_clinical.rda")
GBM_GeneEXp_FsbyVar=FSbyVar(GBM_GeneEXp, cut.type = "topk", value=4000)
GBM_Methylation27_FsbyVar=FSbyVar(GBM_Methylation27,cut.type = "topk", value=4000)
GBM_miRNA_8x15k_FsbyVar=FSbyVar(GBM_miRNA_8x15k,cut.type = "topk", value=250)
GBM_GeneEXp_FsbyCox=FSbyCox(GBM_GeneEXp, GBM_clinical$time,
                            GBM_clinical$status, cutoff = 0.05)
GBM_Methylation27_FsbyCox=FSbyCox(GBM_Methylation27, GBM_clinical$time,
                                  GBM_clinical$status, cutoff = 0.05)
GBM_miRNA_8x15k_FsbyCox=FSbyCox(GBM_miRNA_8x15k, GBM_clinical$time,
                                GBM_clinical$status, cutoff = 0.05)
####1.mRNA-Var-CC
result5_1=ExecuteCC(clusterNum=3,d=GBM_GeneEXp_FsbyVar,maxK=5,
                  clusterAlg="hc",distance="pearson",title="GBM")
group=result5_1$group
distanceMatrix=result5_1$distanceMatrix
p_value5_1=survAnalysis(mainTitle="GBM_mRNA-Var-CC",GBM_clinical$time,
                     GBM_clinical$status,group,
                     distanceMatrix=distanceMatrix,similarity=TRUE)

####2. mRNA+miRNA-Var-CC
GBM=list(GeneExp=GBM_GeneEXp_FsbyVar,miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result5_2=ExecuteCC(clusterNum=3,d=GBM,maxK=5,
                    clusterAlg="hc",distance="pearson",title="GBM")
group=result5_2$group
distanceMatrix=result5_2$distanceMatrix
p_value5_2=survAnalysis(mainTitle="GBM_mRNA+miRNA-Var-CC",GBM_clinical$time,
                        GBM_clinical$status,group,
                        distanceMatrix=distanceMatrix,similarity=TRUE)
####3. mRNA+miRNA-Var-SNF
GBM=list(GeneExp=GBM_GeneEXp_FsbyVar,miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result5_3=ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result5_3$group
distanceMatrix=result5_3$distanceMatrix
```

```
p_value5_3=survAnalysis(mainTitle="GBM_mRNA+miRNA-Var-SNF",GBM_clinical$time,
                        GBM_clinical$status,group,
                        distanceMatrix=distanceMatrix,similarity=TRUE)
####4. mRNA+DM+miRNA-Var-SNF
GBM=list(GeneExp=GBM_GeneEXp_FsbyVar,
         DNAmethy=GBM_Methylation27_FsbyVar,
         miRNAExp=GBM_miRNA_8x15k_FsbyVar)
result5_4=ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result5_4$group
distanceMatrix=result5_4$distanceMatrix
p_value5_4=survAnalysis(mainTitle="GBM_mRNA+DM+miRNA-Var-SNF",GBM_clinical$time,
                        GBM_clinical$status,group,
                        distanceMatrix=distanceMatrix,similarity=TRUE)
####5. mRNA+miRNA-COX-SNF
GBM=list(GeneExp=GBM_GeneEXp_FsbyCox,
         miRNAExp=GBM_miRNA_8x15k_FsbyCox)
result5_5=ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result5_5$group
distanceMatrix=result5_5$distanceMatrix
p_value5_5=survAnalysis(mainTitle="GBM_mRNA+miRNA-COX-SNF",GBM_clinical$time,
                        GBM_clinical$status,group,
                        distanceMatrix=distanceMatrix,similarity=TRUE)
####6. mRNA+DM+miRNA-COX-SNF
GBM=list(GeneExp=GBM_GeneEXp_FsbyCox,
         DNAmethy=GBM_Methylation27_FsbyCox,
         miRNAExp=GBM_miRNA_8x15k_FsbyCox)
result5_6=ExecuteSNF(GBM, clusterNum=3, K=20, alpha=0.5, t=20)
group=result5_6$group
distanceMatrix=result5_6$distanceMatrix
p_value5_6=survAnalysis(mainTitle="GBM_mRNA+miRNA-COX-SNF",GBM_clinical$time,
                        GBM_clinical$status,group,
                        distanceMatrix=distanceMatrix,similarity=TRUE)
```

The comparison of the six groups of Log-rank test p-values is shown in Figure S20. The multi-omics with COX model for SNF could discovery cancer subtypes with the significant different survival patterns.

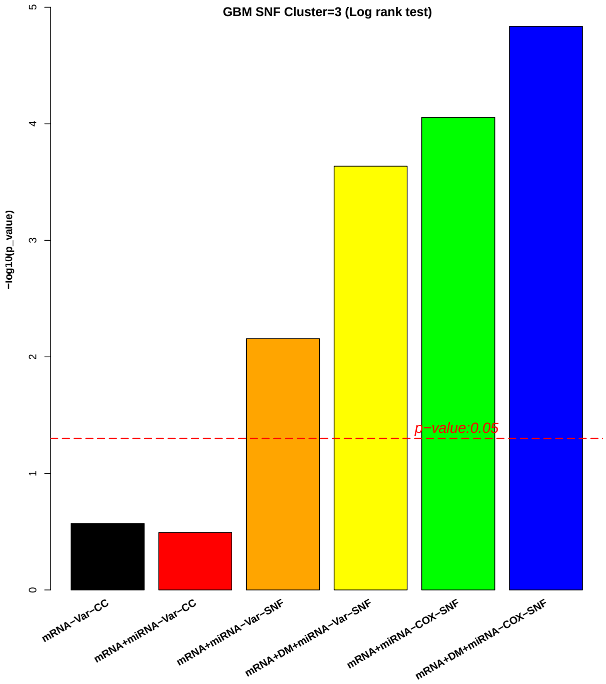

Figure S20: The barplot for the Log-rank test p-values