epihet:

An R Package for Calculating and Analyzing

the Epigenetic Heterogeneity of Cancer Cells

Xiaowen Chen,Haitham Ashoor,Ryan Musich,Mingsheng Zhang,Jiahui Wang,Sheng Li

November 2018

Contents

1 Introduction

This manual introduces the epihet package and shows how the package can be used to calculate epigenetic heterogeneity of cells and visualize the results through various types of graphs. This package was designed to use output from methclone, a C++ library that analyzes the evolution of epialleles using Bisulfite Sequencing methylation data.

2 Installation

1. Download the package from Bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("epihet")
```

Or install the development version of the package from Github.

```
BiocManager::install("TheJacksonLaboratory/epihet")
```

2. Load the package into R session.

```
library(epihet)
```

3 Background

3.1 DNA Methylation & Epigenetic Heterogeneity

Methylation occurs on the DNA strand where a methyl group attaches to the cytosine of a bonded cytosine and guanine pairing. Normal methylation patterns assure the proper regulation of gene expression and stable gene siliencing. Areas of DNA that have a high percentage of methylation are considered to be hypermethylated and have been associated with the silencing of certain tumor-suppressor genes. Areas with lower percentages of methylation are called hypomethylated and are associated with cell transformation.

Recently, cell to cell variations in cancer patients have been proposed to contribute to the treatment failure as it may provide an alternative trajectory for the cancer cells to escape therapy. In addition to the genetic allelic heterogeneity, it has been reported that cancer cells may display various epigenome status within the same patients. Specifically, the epigenetic heterogeneity and dynamics measured by the phased DNA methylation patterns have been reported to associate with clinical outcome in acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), diffuse large Bcell lymphoma, and Ewing sarcoma (Sheffield et al, 2017). As the cancer cell evolves

from diagnosis to relapse, methylation patterns at a given loci can further disrupt promoter sequences and gene regulation (Pan et al, 2015). These differing methylation patterns on the same loci/gene are called epialleles and are key to determining epigenetic heterogeneity. Previous studies have shown that the methylation patterns of cancer cells on relapse can vastly differ from the patterns of the same cells on diagnosis (Li et al, 2014). This package uses epialleles consisting of four cytosine base pairs to create 16 (either a methylated or unmethylated cytosine across four unique loci) distinct methylation patterns to analyze methylation levels in the samples.

3.2 Important Variables for Analysis

To determine epigenetic heterogeneity, epihet uses the following three variables:proportion of discordant reads (PDR), epipolymorphism, and Shannon entropy values. By comparing the similarities and differences between these values at the same epiallele across multiple samples, the extent of heterogeneity between the samples can be analyzed by the user.

3.2.1 Proportion of Discordant Reads (PDR)

One important variable to analyze for heterogeneity between cells is the proportion of discordant reads (PDR). PDR at each locus is defined as a proportion of discordant reads compared to the number of total reads from that locus (Landau et al, 2014). A bisulfite sequencing read at a given locus, a concordant read is one that shows all unmethlated or all methylated sites at a given loci, such as a four methylated cytosines. A discordant read is one that shows varying states of methylated and unmethylated regions at a given loci, such as a methylated cytosine followed by three unmethylated cytosines. After finding the number of discordant reads at a given loci, a proportion can be calculated by dividing by the number of total reads from that locus. PDR values can then be used for analyzing epigenetic heterogeneity because a greater value for PDR corresponds to a greater amount of discord within the sample. With a greater level of discord, the sample is considered to be more heterogeneous in nature and could lead to adverse clinical outcomes upon treatment.

In this package, epihet calculates PDR value for each locus of one sample from the percentages of each methylation pattern. For a given locus, epihet sums the percentage of reads support all discordant methylation patterns to obtain PDR value. The resulting value is between 0 and 1.

3.2.2 Epipolymorphism

When analyzing epigenetic heterogeneity, epigenetic polymorphism, or epipolymorphism, is an important variable to calculate. Epipolymorphism is defined as “the probability that two epialleles randomly sampled from the locus differ from each other” (Landan et al, 2012). Calculating the epipolymorphism value for a given locus uses the proportions of each methylation pattern to determine how frequently the methylation patterns change across multiple reads. Epipolymorphism value ranges from 0 to 1 with larger values being associated with larger differences in methylation patterns.

In this package, epipolymorphism value for each locus is calculated based on the formula described by Landen *et al.* (Landan et al, 2012). For each locus, the proportions of each methylation pattern are squared then added together and subtracted from 1. The resulting value is the epipolymorphism value for the given locus.

3.2.3 Shannon Entropy

An important aspect of epigenetic heterogeneity is knowing how diverse a given epiallele is compared to other epialleles. This variable is best tracked through Shannon's information entropy. To calculate Shannon entropy, the proportion for each methylation pattern of a given sample must be known. Next, each proportion is multiplied by the logarithmic result of that proportion. These products are summed together and the result is negated to find the value for Shannon entropy. This result is equal to the exponential that corresponds to the "effective number" of epialleles needed to generate an equivalent Shannon entropy value based on the given methylation pattern proportions (Sherwin, 2010). A large value for Shannon entropy corresponds to a greater number of epiallele patterns needed and would be considered more diverse and, therefore, more heterogeneous. A low value for Shannon entropy corresponds to a lesser amount of epiallele patterns needed and be considered less heterogeneous based on its epigenetics. A value of 0 for Shannon entropy shows that only an epialle only contained a single methylation pattern across multiple reads.

4 Building the Comparison Matrix

In order to prepare for analysis, epihet has builtin functions that take in multiple txt files for multiple samples from the program methclone and transforms the data into a large matrix that contains the location information and PDR, epipolymorphism, and Shannon entropy values for each locus that is shared between the inputted samples. The following sections provide examples on how to use the functions in epihet to build the comparison matrix and prepare the data for analysis.

To begin using epihet, the user should already have fed bisulfite sequencing data to methclone, which outputs compressed text file containing epiallele patterns and the percentage of reads supporting each of epiallele patterns at the genomic locus in the sample. Included in epihet are example files used for the sample code in this vignette, which include only the result for epialleles on chromosome 22 for two normal samples (N1, N2) and two AML patients with the CEBPA_sil (isocitrate dehydrogenase 2) mutation (D2238, D2668). The following lines of code can be run to obtain the files and their corresponding ID names:

```
files = c(system.file("extdata","D-2238.chr22.region.methClone_out.gz",package = "epihet"),
          system.file("extdata","D-2668.chr22.region.methClone_out.gz",package = "epihet"),
          system.file("extdata","N-1.chr22.region.methClone_out.gz",package = "epihet"),
          system.file("extdata","N-2.chr22.region.methClone_out.gz",package = "epihet"))
ids = epihet::splitn(basename(files),"[.]",1)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##  method         from
##  [.quosures     rlang
##  c.quosures     rlang
##  print.quosures rlang
##
## Registered S3 method overwritten by 'enrichplot':
##  method               from
##  fortify.enrichResult DOSE
```

4.1 Creating a List of GenomicRanges Objects

To use epihet to calculate epigenetic heterogeneity, users need to assign the compressed text file suffix to methClone_out.gz. Then, makeGR function in epihet reads in the compressed text files and creates a GenomicRanges object for each file. The GenomicRanges objects are returned in a list. The makeGR function is the first step for epihet's pipeline as the result is needed for input in the comparison matrix function, compMatrix. The function takes a vector of file paths, 'files', and a vector of sample names that correspond to the files, 'ids'. The following code creates a list of GenomicRanges objects using epihet's sample files:

```
GR.List = epihet::makeGR(files = files, ids = ids,
                         cores = 1, sve = FALSE)

## Taking input= as a system command ('gzip -dc /tmp/RtmpEjyfPd/Rinst2ebd14b17422/epihet/extdata/D-2
and a variable has been used in the expression passed to `input=`.  Please use fread(cmd=...).
There is a security concern if you are creating an app, and the app could have a malicious
user, and the app is not running in a secure envionment; e.g.  the app is running as root.
Please read item 5 in the NEWS file for v1.11.6 for more information and for the option
to suppress this message.
## Taking input= as a system command ('gzip -dc /tmp/RtmpEjyfPd/Rinst2ebd14b17422/epihet/extdata/D-2
and a variable has been used in the expression passed to `input=`.  Please use fread(cmd=...).
There is a security concern if you are creating an app, and the app could have a malicious
user, and the app is not running in a secure envionment; e.g.  the app is running as root.
Please read item 5 in the NEWS file for v1.11.6 for more information and for the option
to suppress this message.
## Taking input= as a system command ('gzip -dc /tmp/RtmpEjyfPd/Rinst2ebd14b17422/epihet/extdata/N-1
and a variable has been used in the expression passed to `input=`.  Please use fread(cmd=...).
```

```
There is a security concern if you are creating an app, and the app could have a malicious
user, and the app is not running in a secure envionment; e.g.  the app is running as root.
Please read item 5 in the NEWS file for v1.11.6 for more information and for the option
to suppress this message.
## Taking input= as a system command ('gzip -dc /tmp/RtmpEjyfPd/Rinst2ebd14b17422/epihet/extdata/N-2
and a variable has been used in the expression passed to 'input='.  Please use fread(cmd=...).
There is a security concern if you are creating an app, and the app could have a malicious
user, and the app is not running in a secure envionment; e.g.  the app is running as root.
Please read item 5 in the NEWS file for v1.11.6 for more information and for the option
to suppress this message.
```

One row in each GenomicRanges object in the list cotains the information of one locus,including chromosome number, range of the strand, and strand type, as well as six additional columns that includes the locus ID, number of reads, average methylation percentage of the locus, and PDR, epipolymorphism, and Shannon entropy values of the locus. If makeGR is being used to process many files, the variable 'cores' can be changed to specify the number of cores to use for parallel execution. If the resulting GenomicRanges list must be saved for later use, the variable 'sve' can be set to TRUE and the result will be saved to a .rda file.

4.2 Generate Comparison Matrix

The list of GenomicRanges objects can be used to generate the comparison matrix that is used for the analysis functions included in epihet. The comparison matrix is created using the 'compMatrix' function which locates epialleles that are shared by a certain percentage of the samples and organizing the data into sections for read number, average methylation levels, PDR, epipolymorphism, and Shannon entropy values at these matching loci. The following code creates a comparison matrix from the GenomicRanges list created in the previous section and finds epialleles that are present in 100% of the samples (p = 1):

```
comp.Matrix = epihet::compMatrix(epi.gr = GR.List, outprefix = NULL,
                                 readNumber = 60, p = 1,
                                 cores = 1, sve = FALSE)
```

The parameter 'p' can range from 0 to 1. The comparison matrix comp.Matrix contains epigenetic heterogeneity values of the locus shared by the 100p percentile samples. For example, the loci in comp.Matrix are shared by at least half of the samples, then 'p' should be set to 0.50. If the resulting comparison matrix needs to be saved, 'sve' can be set to TRUE and 'outprefix' can be specified to add a prefix to the resulting .rda file. If large files are being used, the matrix can be created using multiple cores to speed up the execution as specified by the 'cores' variable.

4.3 Create Single GenomicRanges Object

Instead of creating a list of GenomicRanges objects, a single GenomicRanges object can be created using the 'readGR' function. The function takes a vector of files, 'files', and a vector of IDs, 'ids', that correspond to the files. The index of the file, 'n', to be used for creating the GenomicRanges object is also needed. The following code generates the GenomicRanges object for the third file in the vector:

```
GR.Object = epihet::readGR(files = files, ids = ids, n = 3)

## Taking input= as a system command ('gzip -dc /tmp/RtmpEjyfPd/Rinst2ebd14b17422/epihet/extdata/N-1
and a variable has been used in the expression passed to 'input='.  Please use fread(cmd=...).
There is a security concern if you are creating an app, and the app could have a malicious
user, and the app is not running in a secure envionment; e.g.  the app is running as root.
Please read item 5 in the NEWS file for v1.11.6 for more information and for the option
to suppress this message.
```

4.4 Simple Summary for Two Samples

If only a quick comparison is needed between two samples, epihet's summarize function can be used to provide a simple overview of how the values of one sample correlate to the values of another sample. The summarize function works by taking in two GenomicRanges objects and the values of each object that will be compared. The possible values for the summarize function are 'pdr', 'epipoly', and 'shannon' that correspond to the data values of the same name. Two different cutoffs specify read coverage of a locus which is included in the summary. The output of summarize is a dataframe that contains the mean of the first and second value of common loci between the two samples with a number of reads greater than both cutoffs. The correlation between value one and value two at these loci are also calculated as well as the number of common loci at both read cutoffs. The following code generates a summary of PDR and epipolymorphism values for the first and second sample (D2238, D2668) in the GR.List created earlier with read coverage cutoffs of 10 and 60:

```
summary = epihet::summarize(gr1 = GR.List[[1]], gr2 = GR.List[[2]],
                            value1 = 'pdr', value2 = 'epipoly',
                            cutoff1 = 10, cutoff2 = 60)

## Warning in mean.default(sub1$values.epi):  argument is not numeric or logical:  returning
NA
## Warning in mean.default(sub2$values.epi):  argument is not numeric or logical:  returning
NA
```

The result of this code shows that the PDR and epipolymorphism values have a high positive correlation at both cutoffs, but the correlation increases as the number of reads increases. The result also shows that there are about 900 common loci between the samples when increasing the number of reads from 10 to 60.

5 Analyzing the Data

Once the comparison matrix has been generated for the sample data, the epigenetic heterogeneity can be analyzed by using epihet's builtin analysis functions. With epihet, the data in the comparison matrix can be used to create boxplots, heat maps, and PCA, tSNE, and MA plots. For most of the functions used for analysis in epihet, annotation information can be added as a parameter that will annotate or group the data based on the cancer type or subtype information provided by the user. The annotation information must be a dataframe with row names as the samples, data entries as the corresponding group annotations. For our example, the subtype groupings for the samples will be used as annotations and can be created by the following code:

```
subtype = data.frame(Type= c(rep('CEBPA_sil', 2), rep('Normal', 2)),
                     row.names = names(GR.List),stringsAsFactors = FALSE)
```

5.1 Creating Boxplots

A simple way to analyze how data is distributed across samples are comparative boxplot. By creating boxplot, one can easily analyze the spread, median, range, and find any potential outliers in the data. The boxplot function in epihet, called epiBox, is used to create a boxplot of a specific value, such as 'pdr', 'epipoly', or 'shannon', for each grouping of samples as inputted by the user. The following call to epiBox compares epipolymorphism values across the samples and creates the figure seen below:

```
epihet::epiBox(compare.matrix = comp.Matrix, value = 'epipoly',
               type = subtype, box.colors = NULL, add.points = FALSE,
               points.colors = NULL, pdf.height = 10, pdf.width = 10,
               sve = TRUE)
```

As the figure 1 shows, the CEBPA_sil mutation samples have a range in epipolymorphism values ranging from 0.105 to 0.115, while the normal samples have a very small range with all average epipolymorphism values being around 0.06. The figure also shows a relatively large difference between the median epipolymorphism values between the two subtypes with CEBPA_sil mutation at 0.1075 and normal at 0.06. Overall, this figure shows that the CEBPA_sil mutation samples have greater epipolymorphism values than the normal samples. It means that the epigenetic heterogeneity of the CEBPA_sil samples is greater than the normal samples due to the varying methylation status.

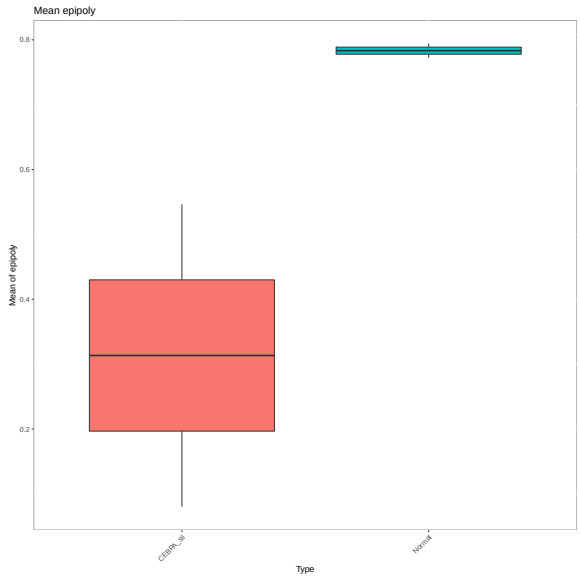

Figure 1: Boxplot of CEBPA_sil and Normal sample by epipolymorphism value.

Some other important features of the epiBox function are adding the individual data points for each sample to the boxplots through the 'add.points' parameter, customizing colors of both the boxes and points by adding a vector of colors to both 'box.colors' and 'points.colors', and saving the resulting figure as a .pdf file using 'sve', 'pdf.height', and 'pdf.width'.

5.2 Generating a Heat Map

We can cluster the samples based on the epigenetic heterogeneity using the most variable genetic loci. The function epiMap used pheatmap function (default value) to create a heatmap plot based the top userinputted percent of loci with the highest standard deviation across all samples. In the exampple, the function use epipolymorphism values to cluster the sample based on the top 5% of loci with highest standard deviation. user can create the appropriate input for the parameter annotate.colors to color the samples by subtype information:

```
pmap = epihet::epiMap(compare.matrix = comp.Matrix,
                      value = 'epipoly',annotate = subtype,
                      clustering_distance_rows = "euclidean",
                      clustering_distance_cols = "euclidean",
                      clustering_method = "complete",annotate.colors = NA,
                      color= colorRampPalette(c("blue","white","red"))(1000),
                      loci.percent = 1, show.rows = FALSE,
                      show.columns = TRUE, font.size = 15,
                      pdf.height = 10, pdf.width = 10, sve = TRUE)
```

Other features of epiMap include customizing the size of the font in the image through 'font.size', showing row or

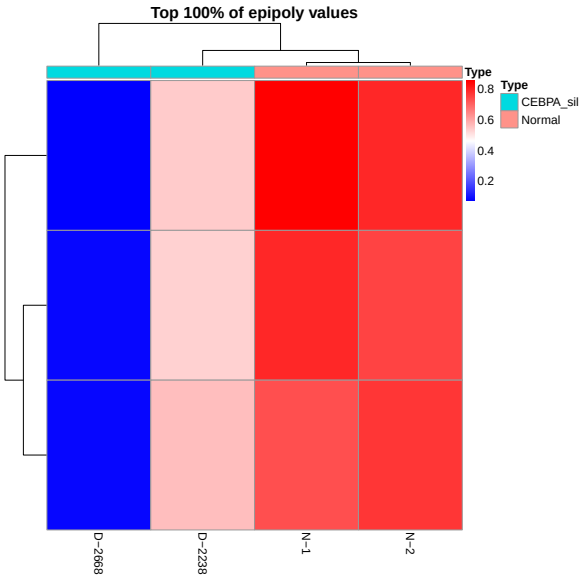

Figure 2: Heatmap of CEBPA_sil and Normal sample by epipolymorphism value.

column names using 'show.rows' and 'show.columns', and the ability to save the image as a .pdf file through 'sve', 'pdf.height', and 'pdf.width'. The colors of the annotations can also be customized by creating a vector of colors and storing them in a list. The following code creates the appropriate input for the 'annotate.colors' parameter to color CEBPA_sil mutation samples in orange and normal samples in forestgreen:

```
box.colors=c("orange","forestgreen")
names(box.colors)=c("CEBPA_sil","Normal")
annotate.colors = list(Type=box.colors)
```

5.3 Graphing a PCA Plot

An important analysis for epigenetic heterogeneity is examining how the investigated samples are grouped based on PDR, epipolymorphism, and Shannon entropy values. This can be accomplished through a principle component analysis (PCA) plot for the comparison matrix. A PCA plot uses an orthogonal transformation to change the data values in the matrix to coordinates based on variance. The epiPCA function creates a PCA plot for either PDR, epipolymorphism, or Shannon entropy values and colors the points by an inputted annotation. The following code creates the PCA plot seen below for epipolymorphism values and colors the points on the plot based on subtype groupings:

```
library(ggfortify)

## Loading required package:  ggplot2

epihet::epiPCA(compare.matrix = comp.Matrix, value = 'epipoly',
```

```
                type = subtype, points.colors = NULL,
                frames = FALSE, frames.colors = NULL,
                probability = FALSE, pdf.height = 10,
                pdf.width = 10, sve = TRUE)
```

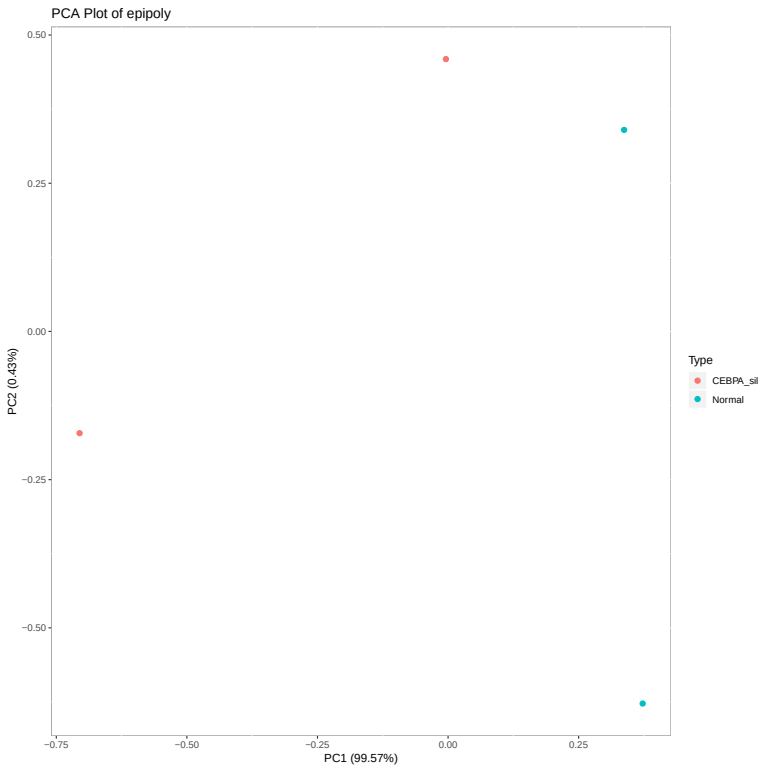

Figure 3: PCA Plot of CEBPA_sil and Normal sample by epipolymorphism value.

The PCA plot shows CEBPA_sil and normal samples can separate each other very well. The normal samples are clustered close together, which means the variance between the normal samples for epipolymorphism values is small. The CEBPA_sil mutation samples are spaced out along the yaxis, which means the variance based on their epipolymorphism values is large.

Other features of epiPCA include adding frames or probability ellipses to better define the groupings of annotations with 'frames' or 'probability', customizing the colors of the points and frames by adding a vector of colors to 'points.colors' and 'frames.colors', and saving the plot as a .pdf file using 'sve', 'pdf.height', and 'pdf.width'.

5.4 Graphing a tSNE Plot

Another plot used to analyze how the samples are group based on a given value is a tdistributed stochastic neighbor embedding (tSNE) plot. Similar to a PCA plot, a tSNE plot is used for analyzing patterns in data by grouping points together, however, a tSNE plot uses multiple dimensions for these groupings. The results are placed on a humanreadable plot in 2dimensions. The function in epihet used for tSNE plot creation is epiTSNE. By using either PDR, epipolymorphism, or Shannon entropy values, epiTSNE can create a tSNE plot and color them by an

inputted annotation. The following code creates the tSNE plot below using epipolymorphism values and colors the points based on their subtype information:

```
set.seed(42)
epihet::epiTSNE(compare.matrix = comp.Matrix, value = 'epipoly',
                type = subtype, points.colors = NULL, theta = 0.5,
                perplexity = 1, max_iter = 1000, pdf.height = 10,
                pdf.width = 10, sve = TRUE)
```

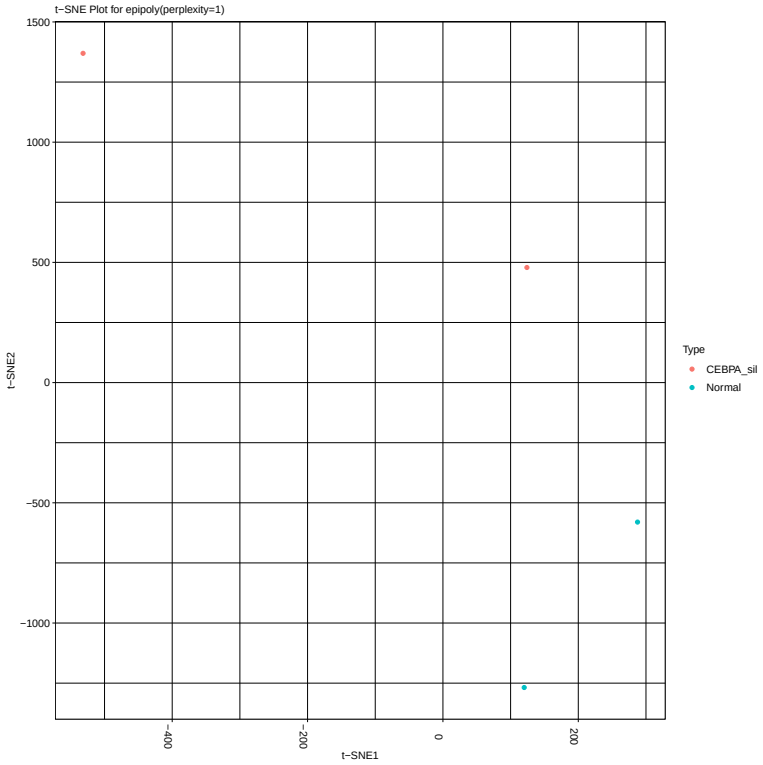

Figure 4: tSNE Plot of CEBPA_sil and Normal sample.

The tSNE plot shows the CEBPA_sil and normal samples cluster two different groups. As these two grouping are at opposite ends of the tSNE plot, one can assume that the epipolymorphism values for the CEBPA_sil mutation samples differ widely in comparison to the normal samples. This corresponds to a large difference in the number of methylation patterns found in the two subtypes.

Other features of epiTSNE include customizing the colors of the points by adding a vector of colors to 'points.colors', customizing the parameters of the Rtsne function used to generate the data for the tSNE plot through 'theta', 'perplexity', and 'max_iter', and the option to save the resulting plot as a .pdf file using 'sve', 'pdf.height', and 'pdf.width'.

5.5 Identifying Differential Epigenetic Heterogeneity locus

The diffHet functions is the main function to identify differential epigenetic heterogeneity (DEH) locus. Depending on the measures of epigenetic heterogeneity user investigates, it will either use ttests or permutation test to calculate p values. p values will be adjusted using multiple test correction. When you identify the loci with differential PDR or epipolymorphism comparing test versus control samples, the function will use ttest. Otherwise, the function will employ EntropyEXPLORER R package to perform the permutation test. And the function also return the mean epigenetic heterogeneity for each group and the mean epigenetic heterogeneity difference between test and control samples. The users can use the mean epigenetic heterogeneity difference and adjusted p values to identify DEH loci. The following code creates the dataframe for all the loci containing differential epipolymorphism between normal and CEBPA_sil mutation samples for epipolymorphism values with a heterogeneity difference cutoff of 0.20:

```
samples=data.frame(Sample=colnames(comp.Matrix)[1:(length(comp.Matrix)-2)],
                   Genotype=c(rep ("CEBPA_sil", 2), rep ("Normal", 2)),
                   stringsAsFactors = FALSE)
rownames(samples)=samples$Sample
seed = sample(1:1e+06, 1)
set.seed(seed)
diff.het.matrix = epihet::diffHet(compare.matrix = comp.Matrix,
                                  value = 'epipoly', group1 = 'CEBPA_sil',
                                  group2 = 'Normal', subtype = samples,
                                  het.dif.cutoff = 0.20,
                                  permutations = 1000,
                                  p.adjust.method = 'fdr', cores = 1)

## [1] "Finish p value calculation"
```

After calculating the heterogeneity difference and adjusted pvalues for each pvalues for each locus, an MA plot can be created. For each locus, the average of the means heterogeneity for two groups versus the heterogeneity difference was plotted. DEH loci (significant pvalues lower than the inputted cutoff) are highlighted in red. The following code creates the MA plot below using the above calculated differential heterogeneity matrix and an adjusted pvalue cutoff of 0.05:

```
data(diffhetmatrix,package="epihet")
epihet::epiMA(pval.matrix = diff.het.matrix, padjust.cutoff = 0.05,
              pch = ".", sve = TRUE)
```

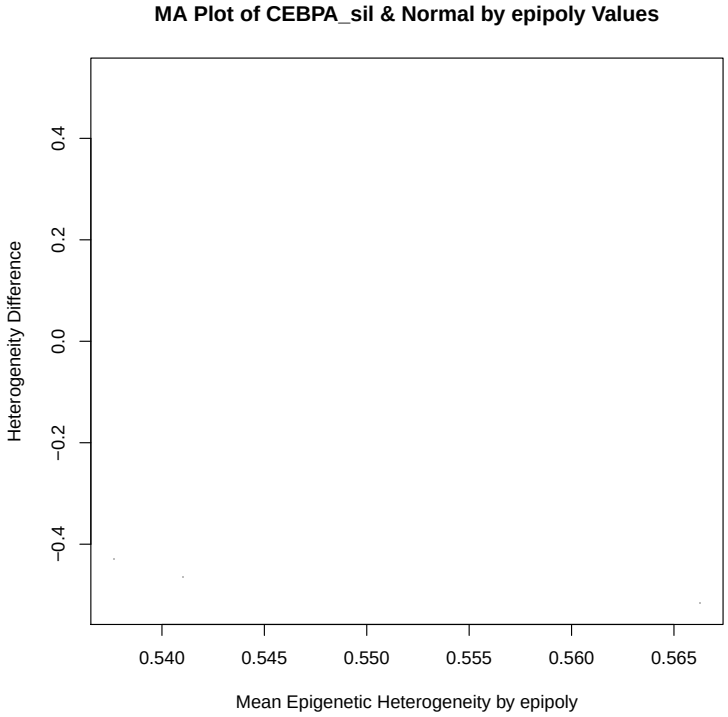

Figure 5: MA Plot of CEBPA_sil and Normal sample.

Through analysis of the MA plot, the majority of coordinates with significant adjusted pvalues are found in the negative portion of the heterogeneity difference axis. This means that the average for epipolymorphism values for CEBPA_sil mutation samples was larger than the average for epipolymorphism values for normal samples. It means the CEBPA_sil mutation samples have greater epigenetic heterogeneity compared with normal samples.
Other features of the diffHet function include specifying a cutoff for the heterogeneity difference through 'het.dif.cutoff', changing the method used to calculate adjusted pvalues using 'p.adjust.method', and the ability to use multiple cores for parallel execution using 'cores'. If Shannon entropy values are to be examined, diffHet uses the EntropyExplorer function to calculate the appropriate pvalue for each locus. The variable for permutations in EntropyExplorer can be modified using 'permutations'.
Other features for epiMA include specifying the adjusted pvalue cutoff to find significant values using 'padjust.cutoff', changing the individual point designs for the plot using 'pch', and the ability to save the plot as a .pdf file using 'sve'.

6 Constructing coepigenetci heterogeneity network

6.1 Network construction and module identification

We can construct coepigenetic heterogeneity network based on the DEH loci using WGCNA R package. For altorithm details, please refer to the tutorial of WGCNA. Here,there are two menthods to construct network,which

was decided by the parameter node.type. One method calculates the coepigenetic heterogeneity between any two DEH loci. Another one is to identify genes with genome region annotated by DEH loci and calculates the coepigenetic heterogeneity between any two genes. Genome region can be promoter,CpG islands,CpG shores and so on. The epigenetic heterogeneity of one gene was measured with the average epigenetic heterogeneity of loci associated with the gene. At this case,annotation files in BED format are need for annotating your DEH loci. You can download annotation from UCSC table browser for your genome of interest. Then you should save the BED file as the Granges object in R, and input it into the parameter annotation.obj.We also provide gene promoter files for Refseq genes. Additionally,users can also supply the clinical traits of patients in dataframe to the parameter datTraits, such as age,gender,survival time, to identify clinically significant modules. Then network can be obtained as follows:

```
library(GenomicRanges)
```

```
## Loading required package:  stats4
## Loading required package:  BiocGenerics
## Loading required package:  parallel
##
## Attaching package:  'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##    clusterExport, clusterMap, parApply, parCapply, parLapply,
##    parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##    IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##    Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##    as.data.frame, basename, cbind, colnames, dirname, do.call,
##    duplicated, eval, evalq, get, grep, grepl, intersect,
##    is.unsorted, lapply, mapply, match, mget, order, paste, pmax,
##    pmax.int, pmin, pmin.int, rank, rbind, rownames, sapply,
##    setdiff, sort, table, tapply, union, unique, unsplit, which,
##    which.max, which.min
## Loading required package:  S4Vectors
```

```
##
## Attaching package:  'S4Vectors'
## The following object is masked from 'package:base':
##
##    expand.grid
## Loading required package:  IRanges
## Loading required package:  GenomeInfoDb

library(doParallel)

## Loading required package:  foreach
## Loading required package:  iterators

registerDoParallel(cores=1)
data(sharedmatrix,package="epihet")
data(DEH,package = "epihet")
data(datTraits,package = "epihet")
data(promoter,package = "epihet")
classes=data.frame(Sample=
                  c(colnames(sharedmatrix)[1:(length(sharedmatrix)-2)],
                  paste("N",1:14,sep = "-")),group=c(rep("CEBPA_sil",6),
                  rep("Normal",14)),stringsAsFactors = FALSE)
rownames(classes)=classes$Sample
epi.network=epihet::epiNetwork(node.type = "gene",DEH,sharedmatrix,
                               value = "epipoly",group="CEBPA_sil",
                               subtype=classes,datTraits = datTraits,
                               promoter,networktype = "signed",
                               method = "pearson",prefix="epipoly",
                               mergeCutHeight = 0.25,minModuleSize = 30)
```

This function is used to generate the coepigenetic heterogeneity network and modules. It will return a list containing epigenetic heterogeneity matrix of patients, module information and genes of each module.At the same time, the function save Topological Overlap Matrices(TOM) as RData format.And, it creates a clustering dendrogram of loci/genes showing module information through assigning different colors.

The function also create a bar plot showing the number of genes associated with DEH loci in each module in the Figure 7.

When users provided the external clinical traits,the clinically significant modules were identified.The result was

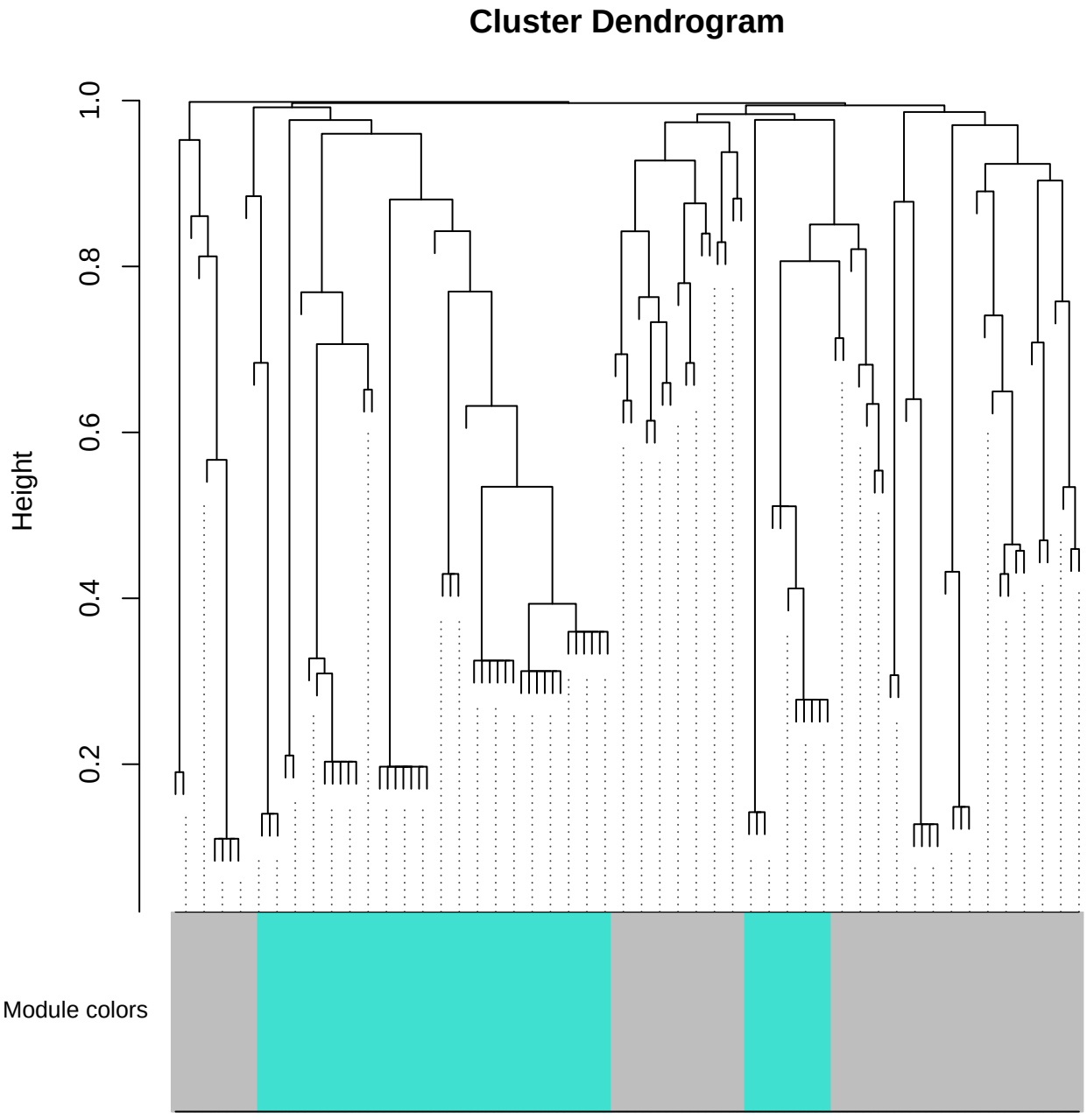

Figure 6: Clustering dendrogram of genes.

visualized by the heatmap plot in Figure 8. Each row in the table corresponds to a module,and each column to a trait. Numbers in the table report the correlations of the correspoinding module eigengenes and traits, with the p values printed below the correlations in parentheses. The color legend denotes correlation.

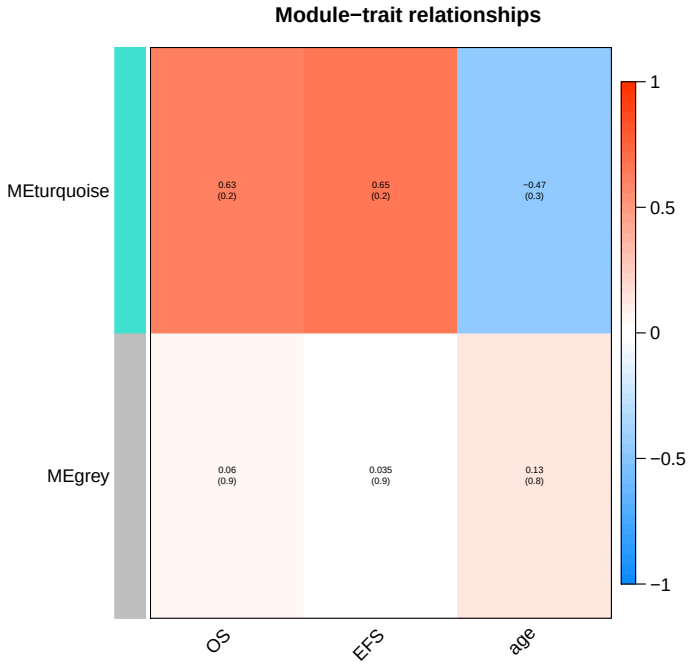

Figure 7: The bar plot showing the number of genes in each module.

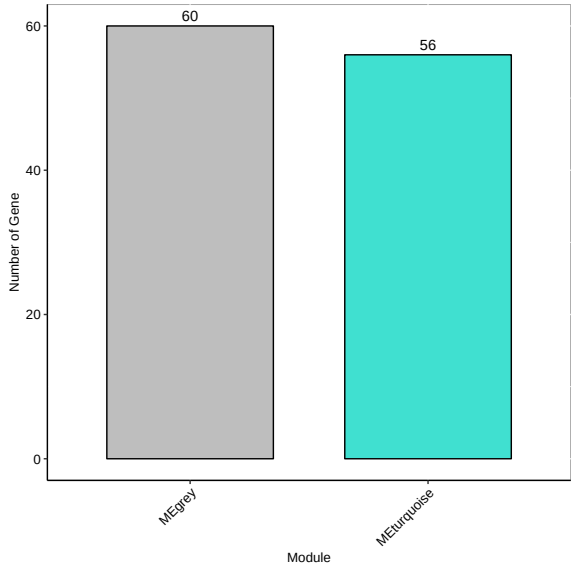

Figure 8: Heatmap showing the relationship between module eigengenes and clinical traits.

6.2 Module Visualization

We can also visualize the modules and perform the network topoloty analysis using the following function.

```
load("epipoly-block.1.RData")
module.topology=epihet::moduleVisual(TOM,
                                     value.matrix=epi.network$epimatrix,
                                     moduleColors=epi.network$module$color,
                                     mymodule="turquoise",cutoff=0.02,
                                     prefix='CEBPA_sil_epipoly',sve = TRUE)
```

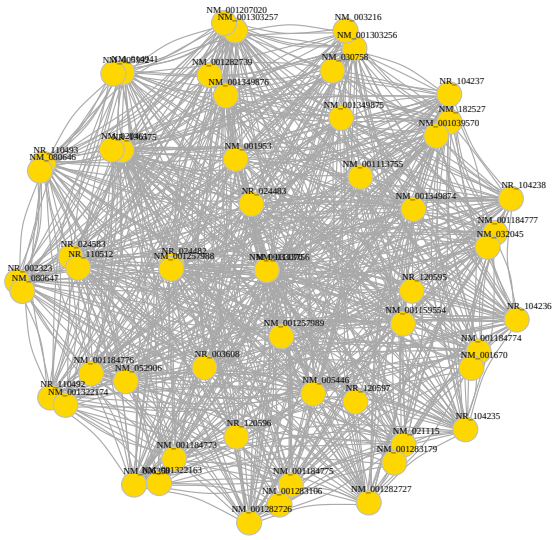

Figure 9: The lightgreen module.

Here,TOM file is loaded into R workspace, which was created by the above function epiNetwork.This function can also generate an edge file and a node file as the txt format. Users can specify network layout and style using Cytoscape.

6.3 Module Annotation

The epihet package contains a function to perform pathway enrichment analysis using a simple,single step. To run the function, ReactomePA needs to be installed before running the code. Here, ReactomePA needs entrez ID, so we firstly transform refseq ID to entrez ID using function bitr() in R package clusterProfiler.

```
library(clusterProfiler)

## clusterProfiler v3.12.0  For help:  https://guangchuangyu.github.io/software/clusterProfiler
##
## If you use clusterProfiler in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He.  clusterProfiler:  an R package
for comparing biological themes among gene clusters.  OMICS: A Journal of Integrative Biology.
2012, 16(5):284-287.

gene=unique(epi.network$module$gene)
entrez=bitr(gene,fromType = "REFSEQ",toType = "ENTREZID",
            OrgDb = "org.Hs.eg.db")

## Loading required package:  org.Hs.eg.db
## Loading required package:  AnnotationDbi
## Loading required package:  Biobase
## Welcome to Bioconductor
##
##    Vignettes contain introductory material; view with
##    'browseVignettes()'.  To cite Bioconductor, see
##    'citation("Biobase")', and for packages 'citation("pkgname")'.
##
## 'select()' returned 1:1 mapping between keys and columns

genelist=epi.network$module
head(genelist)
genelist=merge(genelist,entrez,by.x="gene",by.y="REFSEQ")
genelist=unique(genelist[,c(4,2,3)])
head(genelist)
pathway = epihet::epiPathway(genelist,cutoff = 0.05,
                             prefix="CEBPA_sil",pdf.height = 10,
                             pdf.width = 10)
```

The function returns a dataframe containing pathways significantly enriched by genes of one module and a barplot enrichment map for visualization.

Furthermore, in this analysis we would like to investigate modules that are associated with transcription variance between cancer and normal sample. so we identify the modules significantly enriched by differentially expressed genes (DEGs)

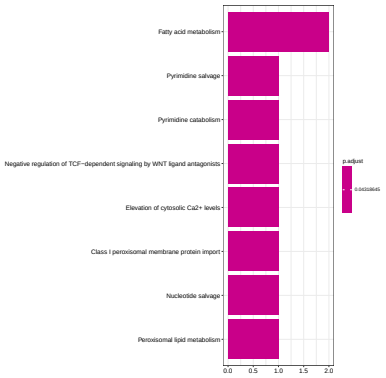

Figure 10: Bar plot of pathway enrichment results.

```
data(DEG,package = "epihet")
data(background,package = "epihet")
module.annotation=epihet::moduleAnno(DEG$refseq,background$gene,
                                     module.gene=epi.network$module,
                                     cutoff=0.1,adjust.method = "fdr",
                                     prefix='epipoly',pdf.height = 3,
                                     pdf.width = 3, sve = TRUE)
```

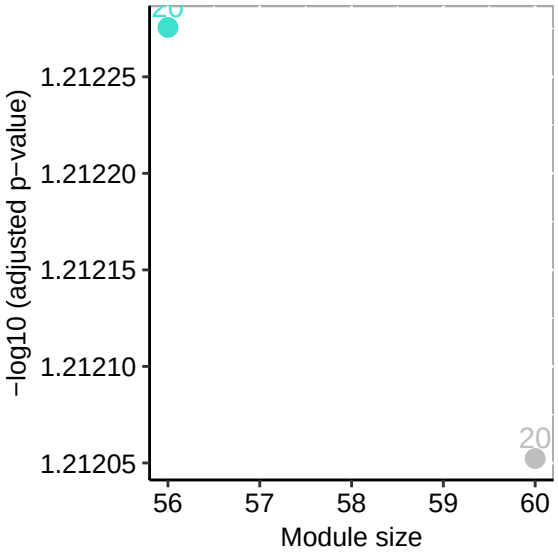

Figure 11: Scatter plot showing the modules enriched by DEG. Xaxis is the number of genes in the module. Yaxis is the log10(p value). p value is obtained from hypergeometric test. The label is the number of DEGs in the module.

6.4 Module comparison

Finally,we compare the similarity between modules using the Jaccard similarity score for different cancers or different sutypes of one cancer.

```
data(modulesil,package = "epihet")
data(moduledm,package = "epihet")
sim.score=epihet::moduleSim(module.subtype1=modulesil,
                            module.subtype2=moduledm,
                            pdf.height = 3,pdf.width = 3,
                            sve = TRUE)
```

Here,you just need to input TOM files of the two cancer types or subtypes you interested. This function returns a matrix showing the Jaccard similarity score between any two modules from different cancers or different subtypes of one cancer and a pheatmap plot to visualization.

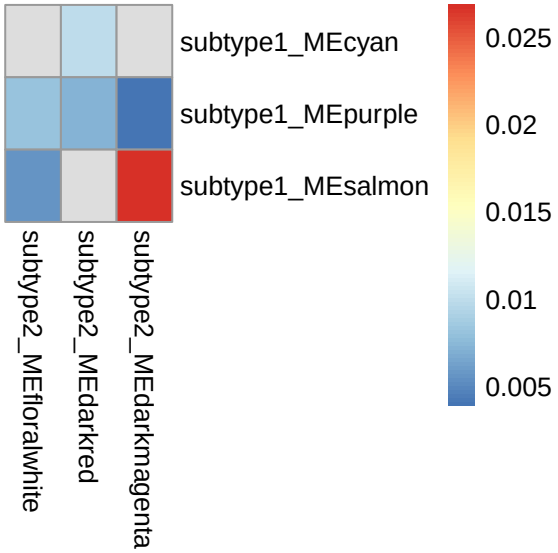

Figure 12: Heatmap plot showing the Jaccard score between any two modules from cancer types/subtypes you interested.

7 SessionInfo

```
  sessionInfo()
```

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.2 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.9-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.9-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.8.2     AnnotationDbi_1.46.0   Biobase_2.44.0
##  [4] clusterProfiler_3.12.0 doParallel_1.0.14      iterators_1.0.10
##  [7] foreach_1.4.4          GenomicRanges_1.36.0   GenomeInfoDb_1.20.0
## [10] IRanges_2.18.0         S4Vectors_0.22.0       BiocGenerics_0.30.0
## [13] ggfortify_0.4.6        ggplot2_3.1.1          knitr_1.22
##
## loaded via a namespace (and not attached):
##   [1] fgsea_1.10.0           Rtsne_0.15             colorspace_1.4-1
##   [4] ggridges_0.5.1         dynamicTreeCut_1.63-1  htmlTable_1.13.1
##   [7] qvalue_2.16.0          XVector_0.24.0         base64enc_0.1-3
##  [10] rstudioapi_0.10        farver_1.1.0           urltools_1.7.3
```

```
##  [13] ggrepel_0.8.0          bit64_0.9-7            mvtnorm_1.0-10
##  [16] xml2_1.2.0             codetools_0.2-16       splines_3.6.0
##  [19] robustbase_0.93-4      impute_1.58.0          GOSemSim_2.10.0
##  [22] polyclip_1.10-0        Formula_1.2-3          jsonlite_1.6
##  [25] WGCNA_1.67             cluster_2.0.9          GO.db_3.8.2
##  [28] pheatmap_1.0.12        graph_1.62.0           ggforce_0.2.2
##  [31] graphite_1.30.0        rrcov_1.4-7            compiler_3.6.0
##  [34] httr_1.4.0             rvcheck_0.1.3          backports_1.1.4
##  [37] assertthat_0.2.1       Matrix_1.2-17          lazyeval_0.2.2
##  [40] tweenr_1.0.1           htmltools_0.3.6        acepack_1.4.1
##  [43] prettyunits_1.0.2      tools_3.6.0            igraph_1.2.4.1
##  [46] gtable_0.3.0           glue_1.3.1             GenomeInfoDbData_1.2.1
##  [49] reshape2_1.4.3         DO.db_2.9              dplyr_0.8.0.1
##  [52] rappdirs_0.3.1         fastmatch_1.1-0        Rcpp_1.0.1
##  [55] enrichplot_1.4.0       preprocessCore_1.46.0  ggraph_1.0.2
##  [58] xfun_0.6               fastcluster_1.1.25     stringr_1.4.0
##  [61] epihet_1.0.0           DOSE_3.10.0            DEoptimR_1.0-8
##  [64] europepmc_0.3          zlibbioc_1.30.0        MASS_7.3-51.4
##  [67] scales_1.0.0           reactome.db_1.68.0     hms_0.4.2
##  [70] RColorBrewer_1.1-2     memoise_1.1.0          gridExtra_2.3
##  [73] UpSetR_1.3.3           triebeard_0.3.0        rpart_4.1-15
##  [76] latticeExtra_0.6-28    stringi_1.4.3          RSQLite_2.1.1
##  [79] highr_0.8              pcaPP_1.9-73           ReactomePA_1.28.0
##  [82] checkmate_1.9.1        BiocParallel_1.18.0    rlang_0.3.4
##  [85] pkgconfig_2.0.2        bitops_1.0-6           matrixStats_0.54.0
##  [88] evaluate_0.13          lattice_0.20-38        purrr_0.3.2
##  [91] labeling_0.3           htmlwidgets_1.3        robust_0.4-18
##  [94] cowplot_0.9.4          bit_1.1-14             tidyselect_0.2.5
##  [97] plyr_1.8.4             magrittr_1.5           R6_2.4.0
## [100] fit.models_0.5-14      Hmisc_4.2-0            DBI_1.0.0
## [103] withr_2.1.2            EntropyExplorer_1.1    pillar_1.3.1
## [106] foreign_0.8-71         survival_2.44-1.1      RCurl_1.95-4.12
## [109] nnet_7.3-12            tibble_2.1.1           crayon_1.3.4
## [112] viridis_0.5.1          progress_1.2.0         grid_3.6.0
## [115] data.table_1.12.2      blob_1.1.1             digest_0.6.18
```

```
## [118] tidyr_0.8.3           gridGraphics_0.3-0     munsell_0.5.0
## [121] viridisLite_0.3.0     ggplotify_0.0.3
```

8 References

1. H. Pagès, M. Lawrence and P. Aboyoun (2017). S4Vectors: S4 implementation of vectors and lists. R package version 0.12.2.

2. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. SpringerVerlag New York, 2009.

3. Jesse H. Krijthe (2015). Rtsne: TDistributed Stochastic Neighbor Embedding using a BarnesHut Implementation, URL: https://github.com/jkrijthe/Rtsne

4. JJ Allaire, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman and Ruben Arslan (2017). rmarkdown: Dynamic Documents for R. R package version 1.4. https://CRAN.Rproject.org/package=rmarkdown

5. Kai Wang, Charles A. Phillips, Arnold M. Saxton and Michael A. Langston (2015). EntropyExplorer: Tools for Exploring Differential Shannon Entropy, Differential Coefficient of Variation and Differential Expression. R package version 1.1.
   https://CRAN.Rproject.org/package=EntropyExplorer

6. Landan G, Cohen NM, et al. "Epigenetic Polymorphism and the Stochastic Formation of Differentially Methylated Regions in Normal and Cancerous Tissues." *Nature Genetics* **44** (2012): 12071214. 10.1038/ng.2442

7. Landau, Dan et al. "Locally Disordered Methylation forms the Basis of Intra tumor Methylome Variation in Chronic Lymphocytic Leukemia." *Cancer Cell* **26(6)** (2014): 813825. 10.1016/j.ccell.2014.10.012

8. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8): e1003118.
   doi:10.1371/journal.pcbi.1003118

9. Li, Sheng et al. "Distinct Evolution and Dynamics of Epigenetic and Genetic Heterogeneity in Acute Myeloid Leukemia." *Nature Medicine* **22.7** (2016): 792799. 10.1038/nm.4125

10. Li, Sheng et al. "Dynamic Evolution of Clonal Epialleles Revealed by Methclone." *Genome Biology* **15** (2014).

11. Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.Rproject.org/package=ggfortify

12. Matt Dowle and Arun Srinivasan (2017). data.table: Extension of ‘data.frame‘. R package version 1.10.4. https://CRAN.Rproject.org/package=data.table

13. Pan, Heng et al. “Epigenomic Evolution in Diffuse Large Bcell Lymphomas.” *Nature Communications* **6** (2015). 10.1038/ncomms7921

14. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.Rproject.org/.

15. Raivo Kolde (2015). pheatmap: Pretty Heatmaps. R package version 1.0.8. https://CRAN.Rproject.org/package=pheatmap

16. Revolution Analytics and Steve Weston (2015). doMC: Foreach Parallel Adaptor for ’parallel’. R package version 1.3.4. https://CRAN.Rproject.org/package=doMC

17. Revolution Analytics and Steve Weston (2015). foreach: Provides Foreach Looping Construct for R. R package version 1.4.3. https://CRAN.Rproject.org/package=foreach

18. Sheffield et al. “DNA Methylation Heterogeneity Defines a Disease Spectrum in Ewing Sarcoma.” *Nature Medicine* **23** (2017): 386395. 10.1038/nm.4273

19. Sherwin, William. “Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography.” *Entropy* **12** (2010): 17651798. 10.3390/e12071765

20. Yihui Xie (2017). knitr: A GeneralPurpose Package for Dynamic Report Generation in R. R package version 1.16.

21. Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. The R Journal, 2016.