Integrative analysis: Methylation Data (450K) and RNA-seq

Ruth Barral Arca

30 de mayo de 2019

Introduction

Colorectal cancer is cancer that originates in the colon or rectum. These cancers can also be called colon cancer or rectal cancer (rectal) depending on where they originate. Colon cancer and rectal cancer are often grouped because they have many common characteristics.

The treatment of rectal cancer is largely based on the stage (extent) of the cancer, although other factors may also be important.People with rectal cancers that have not spread to distant sites are usually treated with surgery. Radiation therapy and chemotherapy may also be given before or after surgery.

Stage I rectal cancers have grown in the deeper layers of the rectum wall, but they have not spread outside of the rectum itself. Whereas Stage IV rectal cancers have spread to distant organs and tissues, such as the liver or lungs. The treatment options for stage IV disease depend to some degree on how widespread the cancer is.

In this project we wish to study the genes responsible of rectal cancer progression comparing the methylation and transcriptomic patterns of patients in stage I versus patients in stage IV following an integrative approach.

Objetive

We will follow a holistic approach integrating transcriptomic and Epigenomic data to study samples from The Cancer Genome Atlas Rectum Adenocarcinoma (TCGA-READ), in order to asses which genes associated to tumor progression have both their methylation and transcriptional status alterated

##Prepare the datasets

Download data from TGCA

Set the enviroment

```
setwd("/home/ruth/Dropbox/TFM_RUTH/third_approach")
library(RTCGAToolbox)
library(Biobase)
library(limma)
library(MetKMR)
library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
```

We will use the R package RTCGAToolbox to download the files from the FIREHOSE repository

```
readData = getFirehoseData (dataset="READ", runDate="20150402",forceDownload = TRUE,
                            Clinic=TRUE, Methylation=TRUE, RNASeq2GeneNorm=TRUE)
save(readData,file="/home/ruth/Dropbox/TFM_RUTH/third_approach/readData.rda")
```

Get the clinical data

```
clin = getData(readData, "clinical")
names(clin)
```

```
##  [1] "Composite Element REF"
##  [2] "years_to_birth"
##  [3] "vital_status"
##  [4] "days_to_death"
##  [5] "days_to_last_followup"
##  [6] "primary_site_of_disease"
##  [7] "neoplasm_diseasestage"
##  [8] "pathology_T_stage"
##  [9] "pathology_N_stage"
## [10] "pathology_M_stage"
## [11] "dcc_upload_date"
## [12] "gender"
## [13] "date_of_initial_pathologic_diagnosis"
## [14] "days_to_last_known_alive"
## [15] "radiation_therapy"
## [16] "histological_type"
## [17] "radiations_radiation_regimenindication"
## [18] "completeness_of_resection"
## [19] "number_of_lymph_nodes"
## [20] "race"
## [21] "ethnicity"
## [22] "batch_number"
```

```
clin$t_stage = factor(substr(clin$pathology_T_stage,1,2))
table(clin$t_stage)
```

```
##
##  t1  t2  t3  t4
##   9  32 115  14
```

Get the expression data (Normalized RNA-seq counts)

```
rnaseq = getData(readData, "RNASeq2GeneNorm")
rnaseq[1:4,1:4]
```

```
##       TCGA-AF-2687-01A-02R-1736-07 TCGA-AF-2689-11A-01R-A32Z-07
## A1BG                      20.1873                       43.4263
## A1CF                      51.0856                      313.3531
## A2BP1                      0.4257                       18.9911
## A2LD1                     90.2639                       92.2611
##       TCGA-AF-2690-01A-02R-1736-07 TCGA-AF-2691-11A-01R-A32Z-07
## A1BG                      56.4619                       35.9451
## A1CF                      24.9913                      218.1571
## A2BP1                      0.9256                       22.6758
## A2LD1                    164.3365                      113.6528
```

```
#Patient identifiers transformation
rid = tolower(substr(colnames(rnaseq),1,12))
rid = gsub("-", ".", rid)
```

```r
mean(rid %in% rownames(clin))
```

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

```r
colnames(rnaseq) = rid
which(duplicated(colnames(rnaseq)))
```

```
## [1]  11  21  23  25  27  35 104
```

```r
rnaseq = rnaseq[,-which(duplicated(colnames(rnaseq)))]


readES = ExpressionSet(log2(rnaseq+1))
pData(readES) = clin[sampleNames(readES),]
readES
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 20501 features, 98 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: tcga.af.2687 tcga.af.2689 ... tcga.g5.6641 (98
##     total)
##   varLabels: Composite Element REF years_to_birth ... t_stage (23
##     total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

```r
clin$pathology_T_stage
```

```
##   [1] "t3"  "t3"  "t3"  "t1"  "t3"  "t2"  "t3"  "t3"  "t3"  "t3"  "t4a"
##  [12] "t2"  "t3"  "t2"  "t4a" "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t1"
##  [23] "t2"  "t3"  "t3"  "t3"  "t3"  "t2"  "t3"  "t3"  "t3"  "t3"  "t3"
##  [34] "t2"  "t3"  "t4"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t2"
##  [45] "t3"  "t3"  "t3"  "t2"  "t1"  "t2"  "t3"  "t2"  "t2"  "t3"  "t3"
##  [56] "t2"  "t2"  "t1"  "t3"  "t3"  "t2"  "t3"  "t3"  "t3"  "t2"  "t3"
##  [67] "t3"  "t3"  "t3"  "t4"  "t3"  "t2"  "t3"  "t3"  "t3"  "t2"  "t2"
##  [78] "t3"  "t3"  "t3"  "t3"  "t2"  "t1"  "t3"  "t3"  "t3"  "t2"  "t3"
##  [89] "t3"  "t3"  "t4"  "t1"  "t4"  "t2"  "t3"  "t2"  "t3"  "t3"  "t3"
## [100] "t3"  "t3"  "t3"  "t3"  "t2"  "t3"  "t3"  "t3"  "t3"  "t3"  "t4"
## [111] "t1"  "t2"  "t3"  "t3"  "t3"  "t3"  "t2"  "t1"  "t3"  "t4a" "t2"
## [122] "t4a" "t2"  "t2"  "t2"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"
## [133] "t3"  "t3"  "t3"  "t2"  "t4a" "t3"  "t3"  "t3"  "t3"  "t3"  "t2"
## [144] "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t3"
## [155] "t4a" "t4b" "t3"  "t3"  "t3"  "t3"  "t3"  "t3"  "t4a" "t2"  "t3"
## [166] "t4a" "t3"  "t3"  "t3"  NA    "t1"
```

```r
#There are patients with NA tumor stage therefore we need to eliminate those patients.
readES$t_stage
```

```
##  [1] t3   t3   t3   t1   t3   t2   t3   t3   t4   t2   t3   t2   t4   t3
## [15] t3   t3   t3   t3   t3   t3   t2   t1   t3   t3   t3   t3   t3   t3
## [29] t3   t3   t2   t3   t3   t3   t3   t3   t4   t1   t2   t3   t3   t3
## [43] t3   t2   t1   t3   t4   t2   t4   t2   t2   t2   t3   t3   t3   t3
## [57] t3   t3   t3   t3   t3   t3   t2   t4   t3   t3   t3   t3   t3   t2
```

```
## [71] t3   t3   t3   t3   t3   t3   t3   t3   t3   t3   t3   t4   t4   t3
## [85] t3   t3   t3   t3   t3   t4   t2   t3   t4   t3   t3   t3   <NA> t1
## Levels: t1 t2 t3 t4
```

```
readES = readES[,!is.na(readES$t_stage)]

 #check that all the samples have an associated tumor stage
table(is.na(readES$t_stage))
```

```
##
## FALSE
##    97
```

Diferential expression analysis:

```
design<-model.matrix(~0+t_stage,data=pData(readES))
head(design)
```

```
##              t_staget1 t_staget2 t_staget3 t_staget4
## tcga.af.2687         0         0         1         0
## tcga.af.2689         0         0         1         0
## tcga.af.2690         0         0         1         0
## tcga.af.2691         1         0         0         0
## tcga.af.2692         0         0         1         0
## tcga.af.2693         0         1         0         0
```

```
fit<-lmFit(readES,design)
contrast.matrix<-makeContrasts(t_staget1-t_staget4,levels=design)
fit2<-contrasts.fit(fit,contrast.matrix)
fite<-eBayes(fit2)
```

```
## Warning: Zero sample variances detected, have been offset away from zero
```

```
top.table<-topTable(fite,coef=1,number=Inf,adjust="BH")
results<-decideTests(fite)
table(results)
```

```
## results
##    -1     0     1
##     2 20488    11
```

```
results.p0.05<-top.table[top.table$adj.P.Val<0.05,]
dim(results.p0.05)
```

```
## [1] 13  6
```

```
results.p0.05[1:5,]
```

```
##                    logFC    AveExpr         t      P.Value   adj.P.Val
## LOC100128977  0.2196981 0.01132465  5.819536 8.150998e-08 0.001671036
## SEMA5B       -2.4269386 4.52996758 -5.213876 1.096380e-06 0.011238439
## GDEP          0.8484637 0.09571759  4.966822 3.041876e-06 0.020787166
## C18orf26      0.3168032 0.02516016  4.813776 5.650601e-06 0.022131855
## GFRA4         0.3481737 0.04383371  4.765126 6.865180e-06 0.022131855
##                      B
## LOC100128977 7.089247
## SEMA5B       4.893415
```

```
## GDEP         4.031006
## C18orf26     3.507857
## GFRA4        3.343442
```

Interactomic analysis

Preparation of methylation data

We selected methylation data from de platform Infinium Human Methylation 450K BeadChip. This data needs to be trated to: 1) transform identifiers to match the ones used in the transcriptomic dataset & clinical data 2) remove duplicates

```
me450k = getData(readData, "Methylation", 2)
fanno = me450k[,1:3]
me450k = data.matrix(me450k[,-c(1:3)])
med = tolower(substr(colnames(me450k),1,12))
med = gsub("-", ".", med)
mean(med %in% rownames(clin))
```

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

```
sum(duplicated(med))
```

```
## [1] 8
```

```
todrop = which(duplicated(med))
me450k = me450k[,-todrop]
med = med[-todrop]
colnames(me450k) = med
ok = intersect(rownames(clin), colnames(me450k))
me450kES = ExpressionSet(me450k[,ok])
pData(me450kES) = clin[ok,]
fData(me450kES) = fanno
me450kES = me450kES[,-which(is.na(me450kES$t_stage))]
```

Keep only common samples to transcriptomic and epigenomic data

```
ok = intersect(sampleNames(me450kES), sampleNames(readES))
meMatch = me450kES[,ok]
esMatch = readES[,ok]

esMatch$t_stage==meMatch$t_stage
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
meMatch = meMatch[,meMatch$t_stage=="t1" |meMatch$t_stage=="t4"]
esMatch = esMatch[,esMatch$t_stage=="t1" |esMatch$t_stage=="t4"]
```

```
colnames(meMatch)==colnames(esMatch)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Get the Infinium Human Methylation 450K BeadChip annotation

```
ann450k<-getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
head(ann450k)
```

```
## DataFrame with 6 rows and 33 columns
##                     chr       pos      strand        Name    AddressA
##             <character> <integer> <character> <character> <character>
## cg00050873        chrY   9363356           -  cg00050873    32735311
## cg00212031        chrY  21239348           -  cg00212031    29674443
## cg00213748        chrY   8148233           -  cg00213748    30703409
## cg00214611        chrY  15815688           -  cg00214611    69792329
## cg00455876        chrY   9385539           -  cg00455876    27653438
## cg01707559        chrY   6778695           +  cg01707559    45652402
##                AddressB                                          ProbeSeqA
##             <character>                                        <character>
## cg00050873    31717405 ACAAAAAAACAACACACAACTATAATAATTTTTAAAATAAATAAACCCCA
## cg00212031    38703326 CCCAATTAACCACAAAAACTAAACAAATTATACAATCAAAAAAACATACA
## cg00213748    36767301 TTTTAACACCTAACACCATTTTAACAATAAAAATTCTACAAAAAAAAACA
## cg00214611    46723459 CTAACTTCCAAACCACACTTTATATACTAAACTACAATATAACACAAACA
## cg00455876    69732350 AACTCTAAACTACCCAACACAAACTCCAAAAACTTCTCAAAAAAAACTCA
## cg01707559    64689504 ACAAATTAAAAACACTAAAACAAACACAACAACTACAACAACAAAAAACA
##                                                    ProbeSeqB        Type
##                                                  <character> <character>
## cg00050873 ACGAAAAAACAACGCACAACTATAATAATTTTTAAAATAAATAAACCCCG           I
## cg00212031 CCCAATTAACCGCAAAAACTAAACAAATTATACGATCGAAAAAACGTACG           I
## cg00213748 TTTTAACGCCTAACACCGTTTTAACGATAAAAATTCTACAAAAAAAAACG           I
## cg00214611 CTAACTTCCGAACCGCGCTTTATATACTAAACTACAATATAACGCGAACG           I
## cg00455876 AACTCTAAACTACCCGACACAAACTCCAAAAACTTCTCGAAAAAAACTCG           I
## cg01707559 GCGAATTAAAAACACTAAAACGAACGCGACGACTACAACGACAAAAAACG           I
##                NextBase       Color    Probe_rs Probe_maf      CpG_rs
##             <character> <character> <character> <numeric> <character>
## cg00050873           A         Red          NA        NA          NA
## cg00212031           T         Red          NA        NA          NA
## cg00213748           A         Red          NA        NA          NA
## cg00214611           A         Red          NA        NA          NA
## cg00455876           A         Red          NA        NA          NA
## cg01707559           A         Red          NA        NA          NA
##               CpG_maf      SBE_rs   SBE_maf            Islands_Name
##             <numeric> <character> <numeric>             <character>
## cg00050873         NA          NA        NA   chrY:9363680-9363943
## cg00212031         NA          NA        NA chrY:21238448-21240005
## cg00213748         NA          NA        NA   chrY:8147877-8148210
## cg00214611         NA          NA        NA chrY:15815488-15815779
## cg00455876         NA          NA        NA   chrY:9385471-9385777
## cg01707559         NA          NA        NA   chrY:6778574-6780028
##             Relation_to_Island
##                    <character>
## cg00050873            N_Shore
```

```
## cg00212031             Island
## cg00213748            S_Shore
## cg00214611             Island
## cg00455876             Island
## cg01707559             Island
##
##
## cg00050873 TATCTCTGTCTGGCGAGGAGGCAACGCACAACTGTGGTGGTTTTTGGAGTGGGTGGACCC[CG]GCCAAGACGGCCTGGGCTGACCAGAC
## cg00212031 CCATTGGCCCGCCCCAGTTGGCCGCAGGGACTGAGCAAGTTATGCGGTCGGGAAGACGTG[CG]TTAAAGGGCTGAAGGGGAGGGACGGA
## cg00213748 TCTGTGGGACCATTTTAACGCCTGGCACCGTTTTAACGATGGAGGTTCTGCAGGAGGGGG[CG]ACCTGGGGTAGGAGGCGTGCTAGTGC
## cg00214611 GCGCCGGCAGGACTAGCTTCCGGGCCGCGCTTTGTGTGCTGGGCTGCAGTGTGGCGCGGG[CG]AGGAAGCTGGTAGGGCGGTTGTCGCA
## cg00455876 CGCGTGTGCCTGGACTCTGAGCTACCCGGCACAAGCTCCAAGGGCTTCTCGGAGGAGGCT[CG]GGGACGGAAGGCGTGGGGTGAGTGGC
## cg01707559 AGCGGCCGCTCCCAGTGGTGGTCACCGCCAGTGCCAATCCCTTGCGCCGCCGTGCAGTCC[CG]CCCTCTGTCGCTGCAGCCGCCGCGCC
##                                                  SourceSeq Random_Loci
##                                                <character> <character>
## cg00050873 CGGGGTCCACCCACTCCAAAAACCACCACAGTTGTGCGTTGCCTCCTCGC
## cg00212031 CGCACGTCTTCCCGACCGCATAACTTGCTCAGTCCCTGCGGCCAACTGGG
## cg00213748 CGCCCCCTCCTGCAGAACCTCCATCGTTAAAACGGTGCCAGGCGTTAAAA
## cg00214611 CGCCCGCGCCACACTGCAGCCCAGCACACAAAGCGCGGCCCGGAAGCTAG
## cg00455876 GACTCTGAGCTACCCGGCACAAGCTCCAAGGGCTTCTCGGAGGAGGCTCG
## cg01707559 CGCCCTCTGTCGCTGCAGCCGCCGCGCCCGCTCCAGTGCCCCCAATTCGC
##            Methyl27_Loci UCSC_RefGene_Name        UCSC_RefGene_Accession
##              <character>       <character>                   <character>
## cg00050873                 TSPY4;FAM197Y2        NM_001164471;NR_001553
## cg00212031                        TTTY14                     NR_001543
## cg00213748
## cg00214611                  TMSB4Y;TMSB4Y           NM_004202;NM_004202
## cg00455876
## cg01707559              TBL1Y;TBL1Y;TBL1Y NM_134259;NM_033284;NM_134258
##              UCSC_RefGene_Group     Phantom         DMR    Enhancer
##                     <character> <character> <character> <character>
## cg00050873         Body;TSS1500
## cg00212031              TSS200
## cg00213748
## cg00214611        1stExon;5'UTR
## cg00455876
## cg01707559 TSS200;TSS200;TSS200
##                     HMM_Island Regulatory_Feature_Name
##                    <character>             <character>
## cg00050873   Y:9973136-9976273
## cg00212031 Y:19697854-19699393
## cg00213748   Y:8207555-8208234
## cg00214611 Y:14324883-14325218     Y:15815422-15815706
## cg00455876   Y:9993394-9995882
## cg01707559   Y:6838022-6839951
##                           Regulatory_Feature_Group         DHS
##                                       <character> <character>
## cg00050873
## cg00212031
## cg00213748
## cg00214611 Promoter_Associated_Cell_type_specific
## cg00455876
## cg01707559
```

```
ann450k <-ann450k [grep("TSS1500|TSS200|5'UTR|1stExon|Body",ann450k$UCSC_RefGene_Group),]
betas<-rownames(meMatch)
table( betas %in% rownames(ann450k))
```

```
##
##  FALSE   TRUE
## 135096 350481
```

```
meMatch<-meMatch[betas %in% rownames(ann450k),]

annotation2 <- data.frame(row = 1:length(ann450k$UCSC_RefGene_Name),
                          pos = ann450k$pos,
                          site=rownames(ann450k),
                          chr=ann450k$chr,
                          gene = ann450k$UCSC_RefGene_Name,
                          stringsAsFactors = F)

#kept only the genes differentially expressed in the expression matrix
DE_genes<-rownames(results.p0.05)
expr<-exprs(esMatch[DE_genes,])

#remove NAs from methylation data
me.data<-as.data.frame(na.omit(exprs(meMatch)))
me.data[1:3,1:3]
```

```
##            tcga.af.4110 tcga.af.6672 tcga.ag.3742
## cg00000029    0.1566724    0.1414851    0.0977719
## cg00000292    0.7305760    0.5594742    0.8403807
## cg00000321    0.3006971    0.1299963    0.7509142
```

MetKMR analysis

We will perform an MetKMR analysis grouping the positions according to the gene they belong (thefore the window size and gap parameter will be ignored).

```
ID<-NULL
pvalue<-NULL

for (i in 1:length(DE_genes)){
  print(paste("Completed %",round(100*i/length(rownames(expr)),2)))
  gene<-DE_genes[i]
  annotation <- annotation2[!is.na(annotation2$gene),]
  annotation <-annotation [grep(gene,annotation$gene),]
  annotation<-annotation[annotation$site %in% rownames(me.data),]
  any (annotation$site %in% rownames(me.data))
  if (any (annotation$site %in% rownames(me.data))){


  analysis2 <- new("MetRKAT",
                   data =data.matrix(na.omit(exprs(meMatch)) ),
                   annotation =annotation,
                   distmethod = c("euclidean"),
                   wsize = 10, gap = 0,
                   max.na = 0.3,
```

```
                  wmethod = 'genes')

  analysis2@intervals <- createIntervals(analysis2)
  #remember to set the output time to C "CONTINUOUS" as our output
  #variable will be the RNA-seq counts
  analysis2@results <- applyRKAT(analysis2, y = expr[gene,], out_type = 'C')

  if (any(analysis2@results$pval < 0.05)) {
    filtered_results <- analysis2@results[analysis2@results$pval <= 0.05, ]
    intervalnames<-analysis2@annotation[filtered_results$first_row, 'gene']
    ID<-c(ID,intervalnames)
    pvalue<-c(pvalue,filtered_results$pval)
    result<-cbind(ID,pvalue)
    print(result)
  }else {
    print (paste(c(DE_genes[i], "is not significant")))
  } }else {next}
}
```

```
## [1] "Completed % 7.69"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
##      ID
## [1,] "MAPT;MAPT;LOC100128977;MAPT;MAPT;MAPT;LOC100130148;MAPT"
##      pvalue
## [1,] "0.00460353203158026"
## [1] "Completed % 15.38"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "SEMA5B"             "is not significant"
## [1] "Completed % 23.08"
## [1] "Completed % 30.77"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "C18orf26"           "is not significant"
## [1] "Completed % 38.46"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "GFRA4"              "is not significant"
## [1] "Completed % 46.15"
```

```
## Discarding/imputing NA values... Done!
```

```
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "TXNDC17"            "is not significant"
## [1] "Completed % 53.85"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "TAS2R41"            "is not significant"
## [1] "Completed % 61.54"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
##      ID
## [1,] "MAPT;MAPT;LOC100128977;MAPT;MAPT;MAPT;LOC100130148;MAPT"
## [2,] "DCST1;DCST2;DCST2;DCST1"
##      pvalue
## [1,] "0.00460353203158026"
## [2,] "0.0378887676174533"
## [1] "Completed % 69.23"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
##      ID
## [1,] "MAPT;MAPT;LOC100128977;MAPT;MAPT;MAPT;LOC100130148;MAPT"
## [2,] "DCST1;DCST2;DCST2;DCST1"
## [3,] "COX10"
##      pvalue
## [1,] "0.00460353203158026"
## [2,] "0.0378887676174533"
## [3,] "0.0305765463945251"
## [1] "Completed % 76.92"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "FOLR4"              "is not significant"
## [1] "Completed % 84.62"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
##      ID
## [1,] "MAPT;MAPT;LOC100128977;MAPT;MAPT;MAPT;LOC100130148;MAPT"
## [2,] "DCST1;DCST2;DCST2;DCST1"
```

```
## [3,] "COX10"
## [4,] "SCO1"
##      pvalue
## [1,] "0.00460353203158026"
## [2,] "0.0378887676174533"
## [3,] "0.0305765463945251"
## [4,] "0.0299962190655586"
## [1] "Completed % 92.31"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "HADH"               "is not significant"
## [1] "Completed % 100"
```

```
## Discarding/imputing NA values... Done!
## Preparing annotation dataset... Done!
## Creating window intervals...
## Retrieving p-value for each window...
```

```
## [1] "CELA2B"             "is not significant"
```

```
result<-as.data.frame(result)


#remove possible  grep pattern matching errors
x<-0
for (i in 1:length(result[,1])) {
  print(strsplit(as.character(result[,1]),";")[i])
  EVAL=any(unlist(strsplit(as.character(result[,1]),";")[i]) %in% DE_genes)
  if (EVAL==TRUE){ x<-c(x,i)}
}
```

```
## [[1]]
## [1] "MAPT"         "MAPT"         "LOC100128977" "MAPT"
## [5] "MAPT"         "MAPT"         "LOC100130148" "MAPT"
##
## [[1]]
## [1] "DCST1" "DCST2" "DCST2" "DCST1"
##
## [[1]]
## [1] "COX10"
##
## [[1]]
## [1] "SCO1"
```

```
#Final results genes associated to rectum adenocarcioma progression
#whose methylation and expression status are altered
result_filtered<-result[x,]
result_filtered
```

```
##                                                          ID
## 1 MAPT;MAPT;LOC100128977;MAPT;MAPT;MAPT;LOC100130148;MAPT
## 3                                                    COX10
## 4                                                     SCO1
##                pvalue
```

```
## 1 0.00460353203158026
## 3  0.0305765463945251
## 4  0.0299962190655586
```

References

- "Working with TCGA data: clinical, expression, mutation and methylation Introduction" https://genomicsclass.github.io/book/pages/tcga.html
- "MiRKAT: Microbiome Regression-Based Kernel Association Test" https://cran.r-project.org/web/packages/MiRKAT/index.html
- American Cancer Society https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/tratamiento/por-etapas-recto.html