survClip package (Version)

Paolo Martini

September 25, 2017

1 survClip: finding prognostic modules exploiting pathway topology

When working with survival analysis the most important things are a good (big enought) batch of patients and an accurate annotations of events and other covariate. Many cancer datasets has these features expecially those collected from TCGA project. In R bioconductor, we can find TCGA data about OV cancer in a package called *curatedOvarianData*. In this brief example we are going to give an overview of survClip package.

We should start from loading the library and the dataset. We used the microarray dataset beacuse RNASeq data row counts are not available. For an example using RNASeq data please refer to our online example at *romauldi.bio.unipd.it*.

```
> # setwd("Documents/work/survClip/vignettes")
> library(curatedOvarianData)
> data(TCGA_eset)
> TCGA_eset
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 13104 features, 578 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: TCGA.20.0987 TCGA.23.1031 ...
    TCGA.13.1819 (578 total)
  varLabels: alt_sample_name unique_patient_ID ...
    uncurated_author_metadata (31 total)
  varMetadata: labelDescription
featureData
  featureNames: A1CF A2M ... ZZZ3 (13104 total)
  fvarLabels: probeset gene
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 21720365
Annotation: hthgu133a
```

This dataset consist of ĩ3000 genes measured over 578 patients. All the patients has associated clinical data that include relapse events, vital status and survival rate. In the following chunck of code we are going to format the clinical data of the phenoData to get them ready to use in survClip.

```
> names(phenoData(TCGA_eset)@data)
```

```
 [1] "alt_sample_name"
 [2] "unique_patient_ID"
 [3] "sample_type"
 [4] "histological_type"
 [5] "primarysite"
```

```
 [6] "arrayedsite"
 [7] "summarygrade"
 [8] "summarystage"
 [9] "tumorstage"
[10] "substage"
[11] "grade"
[12] "age_at_initial_pathologic_diagnosis"
[13] "pltx"
[14] "tax"
[15] "neo"
[16] "days_to_tumor_recurrence"
[17] "recurrence_status"
[18] "days_to_death"
[19] "vital_status"
[20] "os_binary"
[21] "relapse_binary"
[22] "site_of_tumor_first_recurrence"
[23] "primary_therapy_outcome_success"
[24] "debulking"
[25] "percent_normal_cells"
[26] "percent_stromal_cells"
[27] "percent_tumor_cells"
[28] "batch"
[29] "flag"
[30] "flag_notes"
[31] "uncurated_author_metadata"
```

```
> annot <- phenoData(TCGA_eset)@data
> pid <- annot$unique_patient_ID
> days.recurrence <- annot$days_to_tumor_recurrence
> status.recurrence <- annot$recurrence_status
> days <- annot$days_to_death
> status <- annot$vital_status
> table(status.recurrence)
```

```
status.recurrence
norecurrence   recurrence
         279          299
```

```
> table(status)
```

```
status
deceased   living
     290      270
```

```
> status[status=="living"] <- 0
> status[status=="deceased"] <- 1
> status.recurrence[status.recurrence=="norecurrence"] <- 0
> status.recurrence[status.recurrence=="recurrence"] <- 1
> survAnnot.os <- data.frame(status=as.numeric(status), days=as.numeric(days),
+                            row.names=pid, stringsAsFactors=F)
> survAnnot.pfs <- data.frame(status=as.numeric(status.recurrence), days=as.numeric(days.recurrence),
+                             row.names=pid, stringsAsFactors=F)
```

As results, we build two data.frames that represent the minimal information to run survClip analysis: vital status (status) and the days to death or last follow up (days) fro each patients. In this example we are going to analyze the overall survival. We remove NAs and we sort the samples and the expression matrix according to the survival annotation.

```
> survAnnot <- na.omit(survAnnot.os)
> exp <- exprs(TCGA_eset)
> samples <- colnames(exp)
> samples <- gsub('.', replacement = '-', fixed = T, x = samples)
> colnames(exp)<- samples
> samples <- intersect(samples, row.names(survAnnot))
> survAnnot <- survAnnot[samples,]
> exp <- exp[, samples, drop=F]
```

The expression data are almost ready. We go rapdly througth a step of normalization with limma.

```
> library(limma)
> expN <- normalizeQuantiles(exp)
```

Now we can analyze this subset of patients with survClip. First we need to load pathways. The source of pathway we choose is KEGG form graphite Bioconductor package.

```
> library(graphite)
> kegg<- pathways("hsapiens", "kegg")
```

Then we need to convert the identifier in geneSymbol since our matrix has been summarized by gene symbols.

```
> cancerPathways <- names(kegg)[grep("cancer", names(kegg))]
> kegg <- convertIdentifiers(kegg[cancerPathways], "symbol")
```

At the moment we have all we need to perform the analysis with survClip: an expression matrix, survival annotations and a graph. Let's do it! To speed up analysis we are going to extract a selection of cancer related pathway. In the followings, you will find how to run whole pathway survival analysis. To improve readability I reformat results in a table.

```
> library(survClip)
> cancerRelated<-lapply(cancerPathways, function(p) {
+   graph <- pathwayGraph(kegg[[p]])
+   pathwaySurvivalTest(expN, survAnnot, graph,
+                       pcsSurvCoxMethod = "topological", maxPCs=5)
+ })
> pMat <-t(sapply(cancerRelated, function(cr) {
+   crp <- cr@pvalues
+   sapply(crp, function(x) {x[[1]]})
+ }))
> row.names(pMat)<- cancerPathways
> pMat[1:10,4:5]
```

```
                                  topoPvalue.pvalue
Bladder cancer                          0.4503073811
Breast cancer                           0.0008077811
Central carbon metabolism in cancer     0.0708754375
Choline metabolism in cancer            0.0049718530
Colorectal cancer                       0.1951006304
Endometrial cancer                      0.4216828506
MicroRNAs in cancer                     0.0510126971
Non-small cell lung cancer              0.5039552224
Pancreatic cancer                       0.3626715406
Pathways in cancer                      0.0010651762
                                  topoShrinkPvalue.pvalue
Bladder cancer                                 0.4496668684
Breast cancer                                  0.0005264046
Central carbon metabolism in cancer            0.0619453296
Choline metabolism in cancer                   0.0050860369
Colorectal cancer                              0.2038925491
```

```
Endometrial cancer                          0.3942436163
MicroRNAs in cancer                         0.0510126971
Non-small cell lung cancer                  0.4352725627
Pancreatic cancer                           0.4814462061
Pathways in cancer                          0.0064150121
```

Among the other, "Breast cancer" pathway is particularly significan. Let's try to decompose the pathway and see the survival modules. Please note that it is not mandatory to perform whole pathway test in advance.

```
> pathName = "Breast cancer"
> graph <- pathwayGraph(kegg[[pathName]])
> ct <- cliqueSurvivalTest(expN, graph, survAnnot, pcsSurvCoxMethod = "sparse", maxPCs=5)
> getTopLoadGenes(ct)
```

```
   feature clId           geneLoad whichPC
1    FZD10   13  -0.98666989311384     PC1
2     FZD1   13 -0.622352100171049     PC5
3     FZD3   13   0.60301210677208     PC5
4    FZD10   14  0.993520586336023     PC1
5     FZD1   14    0.7152215867918     PC5
6     FZD7   14  0.644920046564349     PC5
7     FZD5   15  0.870705558033245     PC3
8    WNT7A   15 -0.775667314223233     PC5
9      FOS   23 -0.989189427759661     PC1
10     APC   42                  1     PC1
```