Exercício 3: Visualizando dados da coorte TCGA-LIHC

Professor: Mauro Castro

Estudante: Clístenes Grizafis Bento

Para responder a questão abaixo, considere o conjunto de dados disponível no seguinte endereço da internet:

https://github.com/csgroen/RTN_example_TCGA_LIHC

Este conjunto de dados foi pré-processado a partir do estudo TCGA et al. (2017), que descreve uma coorte de câncer de fígado. Neste conjunto de dados está incluindo uma matriz de valores numéricos e metadados correspondentes, disponibilizado em formato .RData no arquivo:

"tcgaLIHCdata_preprocessed.RData"

Faça o download do arquivo .RData, e observe o tipo de objeto carregado no **RStudio**. Trata-se de um objeto da classe SummarizedExperiment, que representa um "container", contendo uma matriz numérica juntamente com os metadados. Extraia a matriz de dados numéricos do objeto SummarizedExperiment e observe sua estrutura no **RStudio**: as colunas representam casos (n=371) e linhas representam variáveis moleculares (n=29885). Em seguida, selecione as 100 variáveis moleculares que mais contribuem para a distinção de casos (alternativamete, você pode selecionar features usando Coeficiente de Variação ou Abundância) e gere uma matriz filtrada contendo apenas as variáreis selecionadas. Por fim, execute uma análise de clusterização não-supervisionada, e visualize o resultado usando o pacote ComplexHeatmap.

FORMA DE ENVIO: Arquivo PDF

REFERÊNCIAS

The Cancer Genome Atlas Research Network (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell, **169**, 1327–1341.e23.

Example of data preprocessing for RTN and RTNsurvival using the TCGA-LIHC cohort.

< https://github.com/csgroen/RTN_example_TCGA_LIHC >

Morgan M, Obenchain V, Hester J, Pagès H (2022). SummarizedExperiment: SummarizedExperiment container. R package version 1.26.1.

https://bioconductor.org/packages/SummarizedExperiment

In [4]:
```
BiocManager::install("ComplexHeatmap")
```

```
'getOption("repos")' replaces Bioconductor standard repositories, see
'?repositories' for details

replacement repositories:
    CRAN: https://cran.r-project.org

Bioconductor version 3.10 (BiocManager 1.30.15), R 3.6.1 (2019-07-05)
Installing package(s) 'ComplexHeatmap'
Warning message in .inet_warning(msg):
"dependency 'rjson' is not available"also installing the dependencies 'shape', 'circl
ize', 'GetoptLong', 'clue', 'GlobalOptions', 'png'
```

```
  There are binary versions available but the source versions are later:
        binary source needs_compilation
shape     1.4.5  1.4.6             FALSE
circlize 0.4.12 0.4.15             FALSE
clue     0.3-59 0.3-64              TRUE
png       0.1-7  0.1-8              TRUE

  Binaries will be installed
package 'GetoptLong' successfully unpacked and MD5 sums checked
package 'clue' successfully unpacked and MD5 sums checked
package 'GlobalOptions' successfully unpacked and MD5 sums checked
package 'png' successfully unpacked and MD5 sums checked
package 'ComplexHeatmap' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\clist\AppData\Local\Temp\RtmpGW05i4\downloaded_packages
```

```
installing the source packages 'shape', 'circlize'

Old packages: 'askpass', 'backports', 'BH', 'BiocManager', 'boot', 'broom',
  'callr', 'caret', 'class', 'cli', 'clipr', 'clue', 'cluster', 'codetools',
  'colorspace', 'crayon', 'curl', 'data.table', 'DBI', 'dbplyr', 'dichromat',
  'digest', 'dplyr', 'ellipsis', 'evaluate', 'fansi', 'forcats', 'foreach',
  'formatR', 'fs', 'generics', 'ggplot2', 'glmnet', 'glue', 'gower', 'gtable',
  'haven', 'hexbin', 'highr', 'hms', 'htmltools', 'htmlwidgets', 'httpuv',
  'httr', 'ipred', 'IRdisplay', 'IRkernel', 'iterators', 'jsonlite',
  'KernSmooth', 'knitr', 'labeling', 'later', 'lava', 'lubridate', 'magrittr',
  'maps', 'markdown', 'Matrix', 'matrixStats', 'mgcv', 'mime', 'ModelMetrics',
  'modelr', 'nlme', 'nnet', 'numDeriv', 'openssl', 'pbdZMQ', 'pillar',
  'pkgconfig', 'plyr', 'png', 'prettyunits', 'processx', 'prodlim', 'progress',
  'promises', 'ps', 'purrr', 'quantmod', 'R6', 'RColorBrewer', 'Rcpp', 'RCurl',
  'readr', 'readxl', 'recipes', 'repr', 'reprex', 'reshape2', 'rlang',
  'rmarkdown', 'rpart', 'rstudioapi', 'rvest', 'scales', 'selectr', 'shiny',
  'sourcetools', 'spatial', 'SQUAREM', 'stringi', 'stringr', 'survival', 'sys',
  'tibble', 'tidyr', 'tidyselect', 'tidyverse', 'timeDate', 'tinytex', 'TTR',
  'utf8', 'uuid', 'viridisLite', 'whisker', 'withr', 'xfun', 'xml2', 'xts',
  'yaml', 'zoo'
```

Importando pacotes necessários para análise

In [6]:

```
library(ComplexHeatmap)
library(SummarizedExperiment)
library(circlize)
library(RColorBrewer)
```

```
========================================
ComplexHeatmap version 2.2.0
Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
Github page: https://github.com/jokergoo/ComplexHeatmap
Documentation: http://jokergoo.github.io/ComplexHeatmap-reference

If you use it in published research, please cite:
Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
  genomic data. Bioinformatics 2016.
========================================

Warning message:
"package 'circlize' was built under R version 3.6.3"========================================
=======
circlize version 0.4.12
CRAN page: https://cran.r-project.org/package=circlize
Github page: https://github.com/jokergoo/circlize
Documentation: https://jokergoo.github.io/circlize_book/book/

If you use it in published research, please cite:
Gu, Z. circlize implements and enhances circular visualization
  in R. Bioinformatics 2014.

This message can be suppressed by:
  suppressPackageStartupMessages(library(circlize))
========================================
```

Carregando dados

In [8]:
```
load(file = "./data/tcgaLIHCdata_preprocessed.RData")
class(tcgaLIHCdata)
```

'RangedSummarizedExperiment'

Verificando dimensões

In [9]:
```
dim(tcgaLIHCdata)
```

1. 29885
2. 371

Está de acordo com o enunciado

Extração de dados da matrix e metadados

In [10]:
```
gexp <- assay(tcgaLIHCdata)
rowAnnotation <- rowData(tcgaLIHCdata)
colAnnotation <- colData(tcgaLIHCdata)
```

Verificando dados dos objetos

In [11]:
```
class(gexp)
```

'matrix'

In [12]:
```
class(rowAnnotation)
```

'DFrame'

In [13]:
```
class(colAnnotation)
```

'DataFrame'

In [14]:
```
gexp[1:3,1:4]
```

	TCGA-DD-A3A3-01A-11R-A22L-07	TCGA-DD-A1EF-01A-11R-A131-07	TCGA-ED-A627-01A-12R-A311-07	TCGA-DD-AACB-01A-11R-A41C-07
ENSG00000000003	22.37358577	27.57565722	18.8409124	20.16986881
ENSG00000000005	0.04209836	0.01943315	0.0246897	0.02132653
ENSG00000000419	13.36474791	32.92042747	20.6326769	29.75572654

In [15]:
```
rowAnnotation
```

```
DataFrame with 29885 rows and 3 columns
                       ENSEMBL        SYMBOL         OG_ENSEMBL
                   <character>   <character>        <character>
ENSG00000000003 ENSG00000000003        TSPAN6 ENSG00000000003.13
ENSG00000000005 ENSG00000000005          TNMD  ENSG00000000005.5
ENSG00000000419 ENSG00000000419          DPM1 ENSG00000000419.11
ENSG00000000457 ENSG00000000457         SCYL3 ENSG00000000457.12
ENSG00000000460 ENSG00000000460      C1orf112 ENSG00000000460.15
...                         ...           ...                ...
ENSG00000281883 ENSG00000281883    AL512506.3  ENSG00000281883.1
ENSG00000281887 ENSG00000281887 GIMAP1-GIMAP5  ENSG00000281887.1
ENSG00000281903 ENSG00000281903     LINC02246  ENSG00000281903.1
ENSG00000281910 ENSG00000281910      SNORA50A  ENSG00000281910.1
ENSG00000281912 ENSG00000281912     LINC01144  ENSG00000281912.1
```

A estratégia adotada para redução de dados foi a de coeficiente da variação

Atualizando nome das linhas usando rowAnnotation

In [16]:
```
all(rownames(rowAnnotation)==rownames(gexp))
# [1] TRUE
rownames(gexp) <- rowAnnotation$SYMBOL
```

TRUE

Removendo genes com baixa contagem

In [17]:
```
idx <- rowSums(gexp!=0)/ncol(gexp)
gexp <- gexp[idx>0.3,]
dim(gexp)
```

1. 21547
2. 371

Filtrando matrix usando correlação com a variável "Tumor_Stage"

In [18]:
```
idx <- cor(t(gexp), colAnnotation$Tumor_Stage, method = "spearman",
           use="complete.obs")
idx <- sort.list(abs(idx), decreasing = T)[1:100]
gexp_filt <- gexp[idx,]
dim(gexp_filt)
```

1. 100
2. 371

Removendo NAs usando colAnnotation

In [19]:
```
colAnnotation_filt <- colAnnotation[,c("Tumor_Stage"), drop=F]
colAnnotation_filt <- colAnnotation_filt[complete.cases(colAnnotation_filt),, drop=F]
dim(colAnnotation_filt)
```

1. 347
2. 1

In [20]:
```
gexp_filt <- gexp_filt[ ,rownames(colAnnotation_filt)]
dim(gexp_filt)
```

1. 100
2. 347

Mudando escala dos dados

In [21]:
```
x <- gexp_filt
x <- t(apply(x, 1, rank));x <- x/max(x)
x <- t(scale(t(x), center = TRUE, scale = F))
dim(x)
```

1. 100
2. 347

Aplicando clustenização

Ajustando nomes das colunas

In [22]:
```
colAnnotation_filt$Tumor_Stage <- as.factor(colAnnotation_filt$Tumor_Stage)
levels(colAnnotation_filt$Tumor_Stage)
```

1. '1'
2. '2'
3. '3'
4. '4'

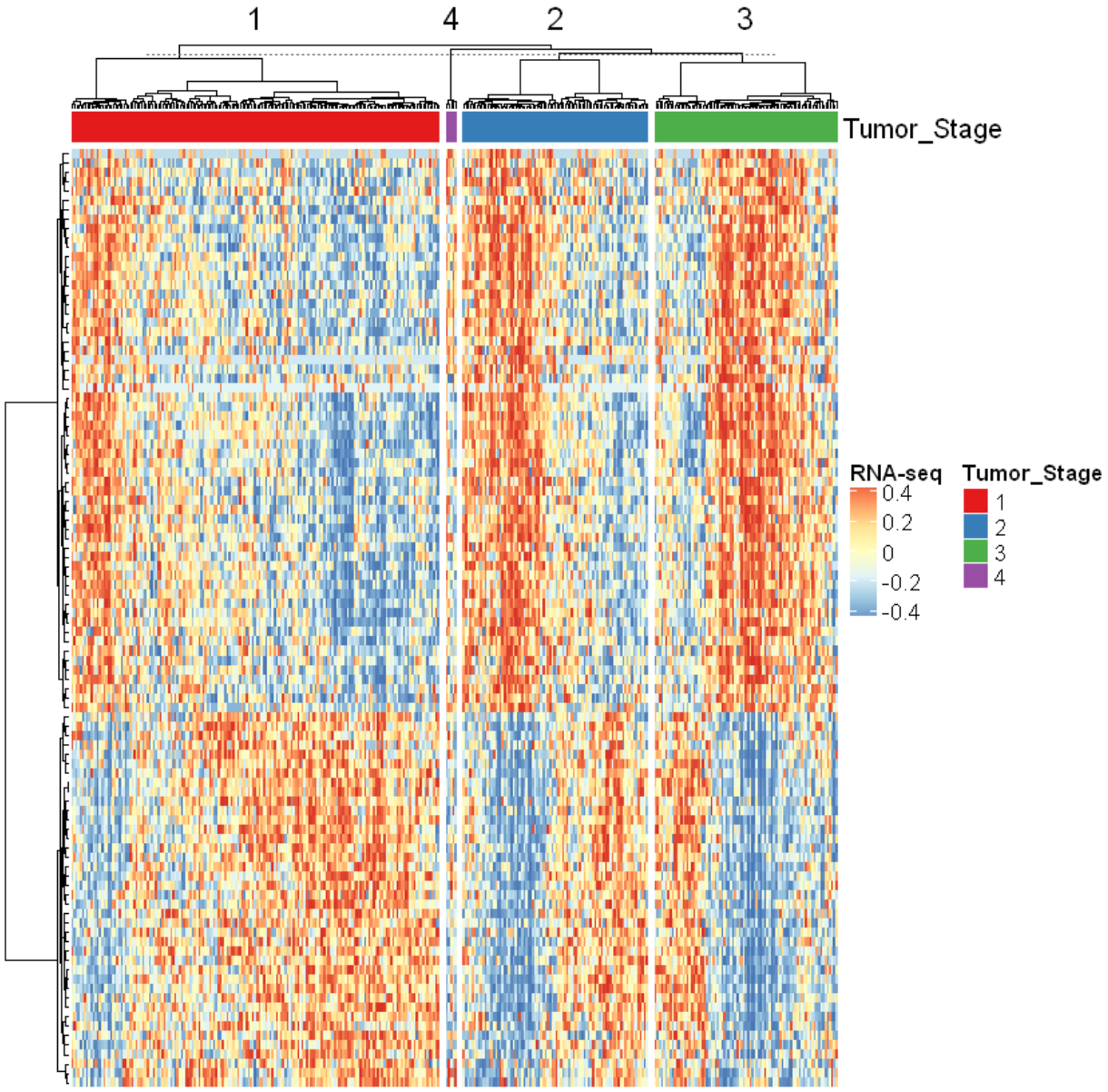
In [23]:
```
pal1 <- brewer.pal(4,"Set1")
names(pal1) <- levels(colAnnotation_filt$Tumor_Stage)
top_annotation <- columnAnnotation(df=colAnnotation_filt,
                                   col=list('Tumor_Stage'=pal1))
```

Ajustando esquema de cores

In [24]:
```
pal2 <- rev(brewer.pal(7,"RdYlBu"))
bks <- quantile(as.numeric(x), probs = seq(0,1, length.out = length(pal2)))
colors <- colorRamp2(breaks = bks, colors = pal2)
```

Executando clustenização e plotando heatmap usando ComplexHeatMap

In [25]:
```
Heatmap(x, col = colors, name = "RNA-seq",
        column_split = colAnnotation_filt$Tumor_Stage,
        show_row_names = F, show_column_names = F,
        top_annotation=top_annotation,
        clustering_method_rows = "ward.D2",
        clustering_distance_rows="spearman",
        clustering_method_columns = "ward.D2",
        clustering_distance_columns = "spearman")
```


1
4
2
3
Tumor_Stage
RNA-seq
0.4
0.2
0
-0.2
-0.4
Tumor_Stage
1
2
3
4