TCGA analysis

Paulyna Magana

Contents

0.1 TCGA Analysis

```
library("TCGAbiolinks")
library("limma")
library("edgeR")
library("caret")
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library("SummarizedExperiment")
```

```
## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: ’BiocGenerics’

## The following object is masked from ’package:limma’:
##
##     plotMA

## The following objects are masked from ’package:stats’:
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from ’package:base’:
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: ’S4Vectors’

## The following objects are masked from ’package:base’:
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: ’IRanges’

## The following object is masked from ’package:grDevices’:
##
##     windows

## Loading required package: GenomeInfoDb
## 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")’.
```

```
##
## Attaching package: ’Biobase’
```

```
## The following object is masked from ’package:MatrixGenerics’:
##
##     rowMedians
```

```
## The following objects are masked from ’package:matrixStats’:
##
##     anyMissing, rowMedians
```

```r
library("gplots")
```

```
##
## Attaching package: ’gplots’
```

```
## The following object is masked from ’package:IRanges’:
##
##     space
```

```
## The following object is masked from ’package:S4Vectors’:
##
##     space
```

```
## The following object is masked from ’package:stats’:
##
##     lowess
```

```r
library("survival")
```

```
##
## Attaching package: ’survival’
```

```
## The following object is masked from ’package:caret’:
##
##     cluster
```

```r
library("survminer")
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library("RColorBrewer")
library("genefilter")
```

```
##
## Attaching package: 'genefilter'

## The following objects are masked from 'package:MatrixGenerics':
##
##     rowSds, rowVars

## The following objects are masked from 'package:matrixStats':
##
##     rowSds, rowVars
```

Query the TCGA database through R with the function GDCquery.

Check all the available projects at TCGA with the command bellow.

```
GDCprojects = getGDCprojects()

head(GDCprojects[c("project_id", "name")] )
```

```
##   project_id
## 1 TARGET-NBL
## 2 GENIE-GRCC
## 3 GENIE-DFCI
## 4  GENIE-NKI
## 5 GENIE-VICC
## 6  GENIE-UHN
##                                                                 name
## 1                                                      Neuroblastoma
## 2         AACR Project GENIE - Contributed by Institut Gustave Roussy
## 3    AACR Project GENIE - Contributed by Dana-Farber Cancer Institute
## 4    AACR Project GENIE - Contributed by Netherlands Cancer Institute
## 5 AACR Project GENIE - Contributed by Vanderbilt-Ingram Cancer Center
## 6 AACR Project GENIE - Contributed by Princess Margaret Cancer Centre
```

```
dplyr::filter(GDCprojects, grepl('Lung', name))
```

```
##          id primary_site dbgap_accession_number project_id disease_type
## 1   CPTAC-3 Pancreas....              phs001287    CPTAC-3 Gliomas,....
## 2 TCGA-LUAD Bronchus....                   <NA>  TCGA-LUAD Cystic, ....
## 3 TCGA-LUSC Bronchus....                   <NA>  TCGA-LUSC Squamous....
##                                                           name releasable state
## 1 CPTAC-Brain, Head and Neck, Kidney, Lung, Pancreas, Uterus       TRUE  open
## 2                                        Lung Adenocarcinoma       TRUE  open
## 3                               Lung Squamous Cell Carcinoma       TRUE  open
##   released tumor
## 1     TRUE     3
## 2     TRUE  LUAD
## 3     TRUE  LUSC
```

1 For TCGA-LUSC, get details on all the data deposited

```
TCGAbiolinks:::getProjectSummary("TCGA-LUSC")
```

```
## $file_count
## [1] 23893
##
## $data_categories
##   file_count case_count               data_category
## 1       3146        504       Copy Number Variation
## 2       3350        504            Sequencing Reads
## 3       7791        497 Simple Nucleotide Variation
## 4       1719        503             DNA Methylation
## 5        577        504                    Clinical
## 6       2148        504      Transcriptome Profiling
## 7       2630        504                 Biospecimen
## 8        328        328          Proteome Profiling
## 9       2204        501        Structural Variation
##
## $case_count
## [1] 504
##
## $file_size
## [1] 3.568685e+13
```

Of note, not all patients were measured for all data types. Also, some data types have more files than samples. This is the case when more experiments were performed per patient, i.e. transcriptome profiling was performed both in mRNA and miRNA, or that data have been analysed by distinct computational strategies.

Let us start by querying all RNA-seq data from LUSC project.

When using GDCquery we always need to specify the id of the project, i.e. “TCGA_LUSC”, and the data category we are interested in, i.e. “Transcriptome Profiling”. Here, we will focus on a particular type of data summarization for mRNA-seq data (workflow.type), which is based on raw counts estimated with HTSeq.

Note that performing this query will take a few of minutes

```
query_TCGA = GDCquery(
  project = "TCGA-LUSC",
  data.category = "Transcriptome Profiling", # parameter enforced by GDCquery
  experimental.strategy = "RNA-Seq",
  workflow.type = "STAR - Counts")
```

```
## --------------------------------------
## o GDCquery: Searching in GDC database
## --------------------------------------
## Genome of reference: hg38
## --------------------------------------------
## oo Accessing GDC. This might take a while...
## --------------------------------------------
## ooo Project: TCGA-LUSC
## --------------------
## oo Filtering results
## --------------------
## ooo By experimental.strategy
## ooo By workflow.type
## ----------------
## oo Checking data
## ----------------
## ooo Checking if there are duplicated cases
## Warning: There are more than one file for the same case. Please verify query results. You can use th
## ooo Checking if there are results for the query
## -------------------
## o Preparing output
## -------------------
```

To visualize the query results in a more readable way, we can use the command getResults.

```
lusc_res = getResults(query_TCGA)

colnames(lusc_res)
```

```
##  [1] "id"                        "data_format"              
##  [3] "cases"                     "access"                   
##  [5] "file_name"                 "submitter_id"             
##  [7] "data_category"             "type"                     
##  [9] "file_size"                 "created_datetime"         
## [11] "md5sum"                    "updated_datetime"         
## [13] "file_id"                   "data_type"                
## [15] "state"                     "experimental_strategy"    
## [17] "version"                   "data_release"             
## [19] "project"                   "analysis_id"              
## [21] "analysis_state"            "analysis_submitter_id"    
## [23] "analysis_workflow_link"    "analysis_workflow_type"   
## [25] "analysis_workflow_version" "sample_type"              
## [27] "is_ffpe"                   "cases.submitter_id"       
## [29] "sample.submitter_id"
```

```
head(lusc_res)
```

```
##                                      id data_format                        cases
## 1 a4b51d89-f5bf-44c1-9822-9bd033709681         TSV TCGA-18-4083-01A-01R-1100-07
## 2 a79deb09-f575-42d9-976b-10ea495e95f7         TSV TCGA-77-8150-01A-11R-2247-07
## 3 6602a055-2305-48a2-9c54-8a29778c5644         TSV TCGA-34-8454-01A-11R-2326-07
## 4 95abf543-3987-4436-ba2a-a15e9c244d77         TSV TCGA-66-2727-01A-01R-0980-07
## 5 f9c280f7-4975-4970-806d-2cf4b94ccf74         TSV TCGA-43-6770-01A-11R-1820-07
## 6 6661360a-3532-4876-a994-580d1bba454e         TSV TCGA-22-5491-01A-01R-1635-07
##       access
## 1       open
## 2 controlled
## 3       open
## 4       open
## 5       open
## 6       open
##                                                                      file_name
## 1 60438408-6d49-446f-a182-57732b78c6d8.rna_seq.augmented_star_gene_counts.tsv
## 2   1bd807e3-4eea-4a60-9df3-ca0d26fe5080.rna_seq.star_splice_junctions.tsv.gz
## 3 e83faec0-340e-4d7a-b56f-dfe285e6b794.rna_seq.augmented_star_gene_counts.tsv
## 4 0ed46774-ff35-4335-9123-6e39bb29a13a.rna_seq.augmented_star_gene_counts.tsv
## 5 7733c278-8e2b-4d6f-9e28-82256ea05035.rna_seq.augmented_star_gene_counts.tsv
## 6 399ad5f4-8c56-4de9-82ec-bcddc45700b3.rna_seq.augmented_star_gene_counts.tsv
##                            submitter_id           data_category            type
## 1 b79f85e4-4815-48fa-95d5-77498762efc6 Transcriptome Profiling gene_expression
## 2 0e271ef7-5ea1-44a5-a95c-b5abd1d5b54f Transcriptome Profiling gene_expression
## 3 5758fa25-cdf3-4f6b-90b9-d476026fd7b7 Transcriptome Profiling gene_expression
## 4 ef705cd6-174c-4385-af3c-048d69ac1ea3 Transcriptome Profiling gene_expression
## 5 580f6b41-9837-46b8-befb-d016e7fb1d36 Transcriptome Profiling gene_expression
## 6 bd8283b6-9ae7-4e88-81c6-7d70e0c01170 Transcriptome Profiling gene_expression
##   file_size                 created_datetime                           md5sum
## 1   4254197 2021-12-13T19:56:32.595477-06:00 3be38dfda35b9d3239db9d3e7c82c1ae
## 2   2027168 2021-12-13T19:45:56.852365-06:00 aa0d4e5a411df7a676b1f4dcf43b2e7e
## 3   4242218 2021-12-13T20:11:01.707109-06:00 183aad4f4cd026fa502c492ea11c3b16
## 4   4242658 2021-12-13T19:59:20.016308-06:00 da2dd64c24f6b0a92404c118d8f34a87
## 5   4266117 2021-12-13T20:03:24.254088-06:00 19285a7294a41a1a4ba469e0122aafbb
## 6   4251146 2021-12-13T19:58:27.817949-06:00 57fe562c9b4368a7e514aa5c7d3af7ba
##                   updated_datetime                              file_id
## 1 2022-01-19T14:45:53.298207-06:00 a4b51d89-f5bf-44c1-9822-9bd033709681
## 2 2022-01-19T13:48:50.358334-06:00 a79deb09-f575-42d9-976b-10ea495e95f7
## 3 2022-01-19T14:45:30.592276-06:00 6602a055-2305-48a2-9c54-8a29778c5644
## 4 2022-01-19T14:45:48.166544-06:00 95abf543-3987-4436-ba2a-a15e9c244d77
## 5 2022-01-19T14:46:26.205497-06:00 f9c280f7-4975-4970-806d-2cf4b94ccf74
## 6 2022-01-19T14:45:30.745815-06:00 6661360a-3532-4876-a994-580d1bba454e
##                        data_type    state experimental_strategy version
## 1 Gene Expression Quantification released               RNA-Seq       1
## 2 Splice Junction Quantification released               RNA-Seq       1
## 3 Gene Expression Quantification released               RNA-Seq       1
## 4 Gene Expression Quantification released               RNA-Seq       1
## 5 Gene Expression Quantification released               RNA-Seq       1
## 6 Gene Expression Quantification released               RNA-Seq       1
##   data_release   project                          analysis_id analysis_state
## 1  32.0 - 35.0 TCGA-LUSC 827759b6-0077-48bb-9187-b6554db3e9fc       released
## 2  32.0 - 35.0 TCGA-LUSC 48f2ed18-e137-4ac2-b6a6-ecd819041ee3       released
## 3  32.0 - 35.0 TCGA-LUSC 64b7e505-7027-4073-9f45-42d581b52289       released
## 4  32.0 - 35.0 TCGA-LUSC 3c8c5d6b-6ef0-487a-b2cd-b04c19cd5d07       released
## 5  32.0 - 35.0 TCGA-LUSC d8d51545-4100-4f33-a406-5b6b54cecb2f       released
## 6  32.0 - 35.0 TCGA-LUSC e5018327-dd2f-4084-a211-8c071aedb132       released
##                             analysis_submitter_id
## 1 60438408-6d49-446f-a182-57732b78c6d8_star__counts
## 2 1bd807e3-4eea-4a60-9df3-ca0d26fe5080_star__counts
## 3 e83faec0-340e-4d7a-b56f-dfe285e6b794_star__counts
## 4 0ed46774-ff35-4335-9123-6e39bb29a13a_star__counts
## 5 7733c278-8e2b-4d6f-9e28-82256ea05035_star__counts
## 6 399ad5f4-8c56-4de9-82ec-bcddc45700b3_star__counts
##
## 1 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
## 2 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
## 3 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
## 4 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
## 5 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
## 6 https://github.com/NCI-GDC/gdc-rnaseq-cwl/blob/5d8c131bbff59fb0c969217fc1d44e6d1503cd1f/rnaseq-sta
##   analysis_workflow_type                analysis_workflow_version   sample_type
## 1          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
## 2          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
## 3          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
## 4          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
## 5          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
## 6          STAR - Counts 5d8c131bbff59fb0c969217fc1d44e6d1503cd1f Primary Tumor
##   is_ffpe cases.submitter_id sample.submitter_id
## 1      NA       TCGA-18-4083    TCGA-18-4083-01A
## 2      NA       TCGA-77-8150    TCGA-77-8150-01A
## 3      NA       TCGA-34-8454    TCGA-34-8454-01A
## 4      NA       TCGA-66-2727    TCGA-66-2727-01A
## 5      NA       TCGA-43-6770    TCGA-43-6770-01A
## 6      NA       TCGA-22-5491    TCGA-22-5491-01A
```