Working with TCGAbiolinks package

Antonio Colaprico, Tiago Chedraoui Silva, Luciano Garofano, Catharina Olsen, Davide Cavalli, Claudia Cava, Isabella Castiglioni, Thais, Tathiane, Stefano Pagnotta, Michele Ceccarelli Houtan Noushmehr, Gianluca Bontempi

2015-09-06

Contents

Introduction

Motivation: The Cancer Genome Atlas (TCGA) provides us with an enormous collection of data sets, not only spanning a large number of cancers but also a large number of experimental platforms. Even though the data can be accessed and downloaded from the database, the possibility to analyse these downloaded data directly in one single R package has not yet been available.

TCGAbiolinks consists of three parts or levels. Firstly, we provide different options to query and download from TCGA relevant data from all currently platforms and their subsequent pre processing for commonly used bio informatics (tools) packages in Bioconductor or CRAN. Secondly, the package allows to integrate different data types and it can be used for different types of analyses dealing with all platforms such as diff expression, network inference or survival analysis etc. and thus it allows to visualize the obtained results. Thirdly we added a socio level where a researcher can found a similar interest in a bioinformatic community, and allows both to find a validation of results in literature in pubmed and also to retrieve questions and answers from sites such as support.bioconductor.org, biostars.org, stackoverflow.etc.

This document describes how to search, download and analyze TCGA data using the TCGAbiolinks package.

TCGAquery: Searching TCGA open-access data

You can easily search TCGA samples using the TCGAquery function. Using a summary of filters as used in the TCGA portal, the function works with the following parameters:

- **tumor** Tumor or list of tumors. The list of tumor is shown in the examples.
- **platform** Platform or list of tumors. The list of platforms is shown in the examples.
- **samples** List of TCGA barcodes
- **level** Options: 1,2,3 "mage-tab"
- **center**
- **version** List of Platform/Tumor/Version to be changed

TCGAquery: some filtering examples

TCGAquery: Searching by tumor

You can filter the search by tumor using the tumor parameter

```
query <- TCGAquery(tumor = "gbm")
```

If you don't remember the tumor name, or if you have incorrectly typed it, it will provide you with all the tumor names in TCGA. Also the names can be seen in the help pages TCGAquery

```
query <- TCGAquery(tumor = "")
```

```
##
##
## Table: TCGA tumors
##
## -----  -----  -----  -----  -----  -----  -----  -----
## ACC    CHOL   GBM    LAML   LUSC   PCPG   STAD   UCS
## BLCA   COAD   HNSC   LGG    MESO   PRAD   TGCT   UVM
## BRCA   DLBC   KICH   LIHC   MISC   READ   THCA   ACC
## CESC   KIRC   KIRP   LUAD   OV     SKCM   THYM   BLCA
```

```
## CHOL   FPPP   KIRP   LUAD   PAAD   SKCM   UCEC   BRCA
## -----  -----  -----  -----  -----  -----  -----  -----
## ------------------------------------------------------
## ERROR: Disease not found. Select from the table above.
## ======================================================
```

TCGAquery: Searching by level

You can filter the search by level "1", "2", "3" or "mage-tab"

```
query <- TCGAquery(tumor = "gbm", level = 3)
query <- TCGAquery(tumor = "gbm", level = 2)
query <- TCGAquery(tumor = "gbm", level = 1)
query <- TCGAquery(tumor = "gbm", level = "mage-tab")
```

TCGAquery: Searching by platform

You can filter the search by platform using the platform parameter

```
query <- TCGAquery(tumor = "gbm", platform = "IlluminaHiSeq_RNASeqV2")
```

If you don't remember the platform, or if you have incorrectly typed it, it will provide you with all the platforms names in TCGA. Also the names can be seen in the help pages ??TCGAquery

```
query <- TCGAquery(tumor = "gbm", platform = "")
```

```
##
##
## Table: TCGA Platforms
##
## ------------------  ---------------------------------  ---------------------------------
## 454                 HumanMethylation27                 IlluminaHiSeq_WGBS
## ABI                 HumanMethylation450                Mapping250K_Nsp
## AgilentG4502A_07    IlluminaGAMethylation_OMA002_CPI   Mapping250K_Sty
## AgilentG4502A_07_1  IlluminaGAMethylation_OMA003_CPI   MDA_RPPA_Core
## AgilentG4502A_07_2  IlluminaGA_DNASeq                  microsat_1
## AgilentG4502A_07_3  IlluminaGA_DNASeq_automated        minbio
## bio                 IlluminaGA_DNASeq_Cont             minbiotab
## biotab              IlluminaGA_DNASeq_Cont_automated   Mixed_DNASeq
## CGH-1x1M_G4447A     IlluminaGA_DNASeq_Cont_curated     Mixed_DNASeq_automated
## diagnostic_images   IlluminaGA_DNASeq_curated          Mixed_DNASeq_Cont
## fh_analyses         IlluminaGA_miRNASeq                Mixed_DNASeq_Cont_automated
## fh_reports          IlluminaGA_mRNA_DGE                Mixed_DNASeq_Cont_curated
## fh_stddata          IlluminaGA_RNASeq                  Mixed_DNASeq_curated
## Genome_Wide_SNP_6   IlluminaGA_RNASeqV2                Multicenter_mutation_calling_MC3
## HumanMethylation27  IlluminaGA                         Multicenter_mutation_calling_MC3_Cont
## H-miRNA_8x15K       IlluminaHiSeq_DNASeq               pathology_reports
## H-miRNA_8x15Kv2     IlluminaHiSeq_DNASeq_automated     SOLiD_DNASeq
## H-miRNA_EarlyAccess IlluminaHiSeq_DNASeq_Cont          SOLiD_DNASeq_automated
## H-miRNA_G4470A      IlluminaHiSeq_DNASeq_Cont_automated SOLiD_DNASeq_Cont
## HG-CGH-244A         IlluminaHiSeq_DNASeq_Cont_curated  SOLiD_DNASeq_Cont_automated
## HG-CGH-415K_G4124A  IlluminaHiSeq_DNASeq_curated       SOLiD_DNASeq_Cont_curated
## HG-U133_Plus_2      IlluminaHiSeq_DNASeqC              SOLiD_DNASeq_curated
## HG-U133A_2          IlluminaHiSeq_miRNASeq             supplemental_clinical
## HT_HG-U133A         IlluminaHiSeq_mRNA_DGE             tissue_images
## HuEx-1_0-st-v2      IlluminaHiSeq_RNASeq               WHG-1x44K_G4112A
```

```
## HumanCNDuo              IlluminaHiSeq_RNASeqV2              NHG-4x44K G4112F
## HumanHap550             IlluminaHiSeq_TotalRNASeqV2         NHG-OGH_4x44K
## ------------------  ----------------------------------  ----------------------------------
## ------------------------------------------------------
## ERROR: Platform not found. Select from the table above.
## ------------------------------------------------------
```

TCGAquery: Searching by center

You can filter the search by center using the center parameter

```
query <- TCGAquery(level = "gdac", center = "unc.edu")
```

If you don't remember the center or if you have incorrectly typed it. It will provide you with all the center names in TCGA

```
query <- TCGAquery(level = "gdac", center = "")
```

```
##
##
## Table: TCGA Centers
##
## ------------------  ----------------------  --------------------
## bcgsc.ca            intgen.org              publicgenomics.com
## broad.mit.edu       jhu-usc.edu             sanger.ac.uk
## broadinstitute.org  jhu.edu                 systemsbiology.org
## crghead CSCa        lbl.gov                 ucsc.edu
## genome.wustl.edu    mdanderson.org          unc.edu
## hgsc.bcm.edu        nasc.org                usc.edu
## hms.harvard.edu     nationwidechildrens.org vanderbilt.edu
## huntsmanlpx.org     pnl.gov                 bcgsc.ca
## ------------------  ----------------------  --------------------
## ------------------------------------------------------
## ERROR: Center not found. Select from the table above.
## ------------------------------------------------------
```

TCGAquery: Searching by samples

You can filter the search by samples using the samples parameter. You can give a list of barcodes or only one barcode. These barcode can be partial barcodes.

```
# You can define a list of samples to query and download providing relative TCGA barcodes.
listSamples <- c("TCGA-E9-A1NG-11A-52R-A14M-07","TCGA-BH-A1FC-11A-32R-A13Q-07",
                 "TCGA-A7-A13G-11A-51R-A13Q-07","TCGA-BE-A0BK-11A-13R-A089-07",
                 "TCGA-B9-A1EI-11A-34R-A109-07","TCGA-BH-A0AU-01A-11R-A12P-07",
                 "TCGA-CU-A1EJ-01A-11R-A13Q-07","TCGA-A7-A13D-01A-13R-A12P-07",
                 "TCGA-AO-A03J-01A-31R-A115-07","TCGA-AO-A03L-01A-11R-A149-07")

# Query all available platforms with a list of barcode
query <- TCGAquery(samples = listSamples)

# Query with a partial barcode
query <- TCGAquery(samples = "TCGA-61-1745-01A")
```

TCGAquery_version: Retrieve versions of the data in TCGA

Query version for a specific platform for example IlluminaHiSeq_RNASeqV2

```
library(TCGAbiolinks)

BRCA_RNASeqV2_version <- TCGAquery_Version(tumor = "brca",
                                           platform = "illuminahiseq_rnaseqv2")
```

The result is shown below.

Table 1: Table with version, number of samples and size (Mbyte) of BRCA IlluminaHiSeq_RNASeqV2 Level 3

Version	Date	Samples	Size/Mbyte
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.11.0/	2015-01-28 08:15	1218	1740.6
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.10.0/	2014-10-16 18:00	1215	1736.4
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.9.0/	2014-07-14 18:13	1182	1689.6
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.8.0/	2014-05-30 23:14	1172	1675.2
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.7.0/	2014-02-13 20:07	1160	1657.9
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.6.0/	2014-01-13 09:53	140	1629.1
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.5.0/	2013-08-22 18:05	106	1580.8
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.4.0/	2013-04-28 16:36	1032	1476.5
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.3.0/	2013-04-12 15:29	956	1369.3
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.2.0/	2012-12-17 18:23	956	1366.5
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.1.0/	2012-07-27 17:52	919	1312.9
unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.0.0/	2012-05-18 12:21	858	1226.1

TCGAquery: Searching old versions

The results from TCGAquery are always the last one from the TCGA data portal. As we have a preprocessing table you should always update TCGAbiolinks package. We intent to update the database constantly.

In case you want an old version of the files we have the version parameter that should be a list of triple values(platform, tumor,version). For example the code below will get the LGG and GBM tumor for platform HumanMethylation450 but for the LGG/HumanMethylation450, we want the version 5 of the files instead of the latest. This could take some seconds

```
query <- TCGAquery(tumor = c("LGG","GBM"),platform = c("HumanMethylation450"),
                   level = 3,
                    version = list(c("HumanMethylation450","LGG",1)))
```

TCGAquery_clinic & TCGAquery_clinicFilt: Working with clinical data.

You can retrive clinical data using the `clinic` function. The parameters of this function are:

- cancer ("OV","BRCA","GBM", etc)
- clinical_data_type ("clinical_patient","clinical_drug", etc)

A full list of cancer and clinical data type can be found in the help of the function.

```
# Get clinical data
clinical_brca_data <- TCGAquery_clinic("brca","clinical_patient")
clinical_uva_data_bio <- TCGAquery_clinic("uva","biospecimen_normal_control")
clinical_brca_data_bio <- TCGAquery_clinic("brca","biospecimen_normal_control")
clinical_brca_data <- TCGAquery_clinic("brca","clinical_patient")
```

Also, some functions to work with clinical data are provided. For example the function TCGAquery_clinicFilt will filter your data, returning the list of barcodes that matches all the filter.

The parameters of TCGAquery_clinicFilt are:

- **barcode** List of barcodes
- **clinical_patient_data** clinical patient data obtained with clinic function Ex: clinical_patient_data <- TCGAquery_clinic("LGG","clinical_patient")
- **HER** her2 neu immunohistochemistry receptor status: "Positive" or "Negative"
- **gender** "MALE" or "FEMALE"
- **PR** Progesterone receptor status: "Positive" or "Negative"
- **stage** Pathologic Stage: "stage_IX", "stage_I", "stage_IA", "stage_IB", "stage_IIX", "stage_IIA", "stage_IIB", "stage_IIIX","stage_IIIA", "stage_IIIB", "stage_IIIC", "stage_IV" -
- **ER** Estrogen receptor status: "Positive" or "Negative"

```
bar <- c('TCGA-BH-A0E0-01A-11R-1789-07', 'TCGA-C8-A26Y-04A-11A-1789-07',
         'TCGA-G9-6332-60A-11R-1789-07', 'TCGA-G9-6338-01A-11R-1789-07',
         'TCGA-G9-6336-11A-11R-1789-07', 'TCGA-G9-7336-11A-11R-1789-07',
         'TCGA-G9-7336-04A-11R-1789-07', 'TCGA-G9-7336-14A-11R-1789-07',
         'TCGA-G9-7030-04A-11R-1789-07', 'TCGA-G9-7030-02A-11R-1789-07',
         'TCGA-G9-7336-11A-11R-1789-07', 'TCGA-G9-7038-06A-11R-1789-07',
         'TCGA-G9-7336-10A-11R-1789-07', 'TCGA-BH-A1ES-10A-11R-1789-07',
         'TCGA-BH-A1FD-10A-11R-1789-07', 'TCGA-BH-A0BZ-02A-11R-1789-07',
         'TCGA-E6-A2MY-04A-11R-1789-07', 'TCGA-BH-A1FG-04A-11R-1789-08',
         'TCGA-C8-A12S-04A-11R-2089-08', 'TCGA-A3-A0EN-11A-11R-0788-07',
         'TCGA-AR-A2LJ-10A-11R-0799-08', 'TCGA-A8-A07B-09A-11R-1788-07',
         'TCGA-BH-A1FB-04A-11R-1789-07', 'TCGA-A8-A07T-04A-55R-1789-07',
         'TCGA-AC-A03S-06A-11R-1789-07', 'TCGA-BH-A0B4-11A-12R-1789-07',
         'TCGA-E6-A1XC-60A-13R-1789-07', 'TCGA-AO-A03S-01A-11R-1789-07',
         'TCGA-AC-A03S-01A-11R-1789-07', 'TCGA-G9-6336-11A-11R-1789-07',
         'TCGA-G9-6330-11A-11R-1789-07', 'TCGA-G9-6330-01A-11R-1789-07',
         'TCGA-G9-6340-01A-11R-1789-07','TCGA-G9-6340-11A-11R-1789-07')

S <- TCGAquery_SampleTypes(bar,'TP')
S2 <- TCGAquery_SampleTypes(bar,'NT')

# Retrieve multiple tissue types  NOT FROM THE SAME PATIENTS
SS <- TCGAquery_SampleTypes(bar,c("TP","TR"))

# Retrieve multiple tissue types  FROM THE SAME PATIENTS
SSS <- TCGAquery_MatchedCoupledSampleTypes(bar,c("NT","TP"))

# Get clinical data
clinical_brca_data <- TCGAquery_clinic("brca","clinical_patient")
female_erpos_herpos <- TCGAquery_clinicFilt(bar,clin, HER="Positive", gender="FEMALE", ER="Positive")
```

The result is shown below:

```
## ER Positive Samples:
##
##
## HER Positive Samples:
##
##
## GENDER FEMALE Samples:
##   TCGA-BH-A1FB
```

```
##  TCGA-BH-A1F0
##  TCGA-BH-A0E2
##  TCGA-BG-A0VY
##  TCGA-BH-A1F6
##  TCGA-BH-A1J8
##  TCGA-AR-A0TM
##  TCGA-AR-A1AJ
##  TCGA-AR-A1AH
##  TCGA-BH-A1FD
##  TCGA-AR-A0AT
##  TCGA-AO-A0J5
##  TCGA-BS-A1ZH

## character(0)
```

TCGAquery_subtypes: Working with molecular subtypes data.

```
# Check with subtypes from TCGAquery
require(xlsx)
GBM_path_subtypes <- TCGAquery_subtypes(tumor = "gbm",path ="../dataGBM")
GBM_subtypes <- as.data.frame(read.xlsx2(GBM_path_subtypes,1,stringsAsFactors = NULL))
GBM_subtypes <- GBM_subtypes[,c(1:10)]
GBM_subtypes <- GBM_subtypes[-1,]
colnames(GBM_subtypes) <- gsub("\\.","-",as.character(GBM_subtypes[1,]))
GBM_subtypes <- GBM_subtypes[-1,]
GBM_subtypes <- GBM_subtypes[!duplicated(GBM_subtypes$Sample_Id),]
rownames(GBM_subtypes) <- GBM_subtypes$Sample_Id
dim(GBM_subtypes)
# [1] 308  10

# starting find difference in subtypes

TableSubtypes_filt_GBM <- TableSubtypes_filt[TableSubtypes_filt$tumor.type == "GBM",]
setdiff(TableSubtypes_filt_GBM$id, GBM_subtypes$Sample_Id)
# [1] "TCGA-19-4065"

require(xlsx)
LGG_path_subtypes <- TCGAquery_subtypes(tumor = "lgg",path ="../dataLGG")
LGG_subtypes <- as.data.frame(read.xlsx2(LGG_path_subtypes,1,stringsAsFactors = NULL))
rownames(LGG_subtypes) <- LGG_subtypes$Tumor
dim(LGG_subtypes)
# [1] 208   9

TableSubtypes_filt_LGG <- TableSubtypes_filt[TableSubtypes_filt$tumor.type == "LGG",]
setdiff(TableSubtypes_filt_LGG$id, LGG_subtypes$Tumor)
# [223] "TCGA-CH-A9DS"

LGG_clinic <- TCGAquery_clinic(cancer = "LGG",clinical_data_type = "clinical_patient")
# table(LGG_clinic$idh1_mutation_found)


STAD_path_subtypes <- TCGAquery_subtypes(tumor = "stad",path ="../dataSTAD")
STAD_subtypes <- as.data.frame(read.xlsx2(STAD_path_subtypes,1,stringsAsFactors = FULL))
```

TCGAquery_integrate: Summary of the common numbers of patient samples in different platforms

Some times researches would like to use samples from different platforms from the same patient. In order to help the user to have an overview of the number of samples in common we created the function TCGAquery_integrate that will receive the data frame returned from TCGAquery and produces a matrix n platforms x n platforms with the values of samples in common.

Some search examples are shown below

```
query <- TCGAquery(tumor = "brca", level = 3)
matSamples <- TCGAquery_integrate(query)
```

The result of the 3 platforms of TCGAquery_integrate result is shown below:

Table 2: Table common samples among platforms from TCGAquery

	AgilentG4502A_07_3	HumanMethylation450	IlluminaHiSeq_RNASeqV2
AgilentG4502A_07_3	604	224	530
HumanMethylation450	224	850	790
IlluminaHiSeq_RNASeqV2	530	790	1218

TCGAquery: some examples

Some search examples are shown below:

```
query <- TCGAquery(tumor = c("gbm","lgg"),
                   platform = c("HumanMethylation450","HumanMethylation27"))

query <- TCGAquery(tumor = "gbm", platform = "HumanMethylation450", level = "2")

query <- TCGAquery(samples = "TCGA-61-1743-01A-01D")

query <- TCGAquery(samples = "TCGA-61-1743-01A-01D-0860-04", level = 3)

query <- TCGAquery(samples = "TCGA-61-1743-01A-01D-0860-04",
                   tumor = "OV", platform = "IlluminaHiSeq_RNASeqV2")
```

TCGAdownload: Downloading open-access data

You can easily download data using the TCGAdownload function.

The arguments are:

- data The TCGAquery output
- path location to save the files. Default: "."
- type Filter the files to download by type
- samples List of samples to download
- force Download again if file already exists? Default: FALSE

TCGAdownload: Example of use

```
# get all samples from the query and save them in the TCGA folder
# samples from Illumina HiSeq RNASeqV2 with type rsem.genes.results
# samples to normalize later

TCGAdownload(query, path = "data", type = "rsem.genes.results")

TCGAdownload(query, path = "data", type = "rsem.isoforms.normalized_results")


TCGAdownload(query, path = "dataBRCA", type = "rsem.genes.results",
             samples = c("TCGA-E9-A1RG-11A-32R-A14M-07",
                         "TCGA-BH-A1FC-11A-32R-A13Q-07")
             )
```

Comment: The function will structure the folders to save the data as: *Path given by the user/Experiment folder*

TCGAdownload: Table of types available for downloading

- **RNASeqV2:** junction_quantification,rsem.genes.results, rsem.isoforms.results, rsem.genes.normalized_results, rsem.isoforms.normalized_results, bt.exon_quantification
- **RNASeq:** exon.quantification,spl.exn.quantification, gene.quantification
- **genome_wide_snp_6:** hg18.seg,hg19.seg,nocnv_hg18.seg,nocnv_hg19.seg

TCGAprepare: Preparing the data

You can easily read the downloaded data using the TCGAprepare function. This function will prepare the data into a SummarizedExperiment [Huber, Wolfgang and Carey, Vincent J and Gentleman, Robert and Anders, Simon and Carlson, Marc and Carvalho, Benilton S and Bravo, Hector Corrada and Davis, Sean and Gatto, Laurent and Girke, Thomas and others 2015] object for downstream analysis. For the moment this function is working only with data level 3

The arguments are:

- **query** Data frame as the one returned from TCGAquery
- **dir** Directory with the files
- **type** File to prepare
- **samples** List of samples to prepare.
- **save** Save a rda object with the prepared object? Default: FALSE
- **filename** Name of the rda object that will be saved if save is TRUE
- **toPackage** Name of the package to prepare the data specific to that package.
- **summarizedExperiment** Should the output be a SummarizedExperiment object? Default: TRUE

In order to add useful information to researches we added in the colData of the summarizedExperiment the subtypes classification for the LGG and GBM samples that can be found in the TCGA publication section We intend to add more tumor types in the future.

Also in the metadata of the objet we added the parameters used in TCGAprepare, the query matrix used for preparing, and file information (name,creation time and modification time) in order to help the user know which samples, versions, and parameters they used.

TCGAprepare: Example of use

```
# get all samples from the query and save them in the TCGA folder
# samples from Illumina HiSeq RNASeq2 with type rsem.genes.results
# samples to normalize later
data <- TCGAprepare(query, dir = "data", save = TRUE, filename = "myfile.rda")
```

As an example, for the platform Illumina HiSeq_RNASeqV2 we prepared two samples (TCGA-DY-A1DE-01A-11R-A155-07 and TCGA-DY-A08A-01A-11R-A155-07) for the rsem.genes.normalized_results type. In order to create the object mapped the gene_id to the hg19. The genes_id not found are then removed from the final matrix. The default output is a SummarizedExperiment is shown below.

```
library(TCGAbiolinks)
library(SummarizedExperiment)
```

```
## Loading required package: GenomicRanges
## 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, xtabs
##
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, 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, rep.int, rownames, sapply,
##     setdiff, sort, table, tapply, union, unique, unlist, unsplit
##
## Loading required package: S4Vectors
## Loading required package: stats4
## Loading required package: IRanges
## 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")'.
```

```
head(assay(dataREAD,"normalized_count"))
```

```
##              TCGA-DY-A1DE-01A-11R-A155-07 TCGA-DY-A08A-01A-11R-A155-07
## A1BG|1                            13.6782                      13.0282
## A1CF|29974                        53.4579                     140.8465
## A2M|2                           6630.4792                    1061.9360
```

```
## A2ML1|144568                    0.0000                    18.2001
## A4GALT|53947                  170.1189                    39.5696
## A4GNT|51146                     0.9606                     0.0000
```

In order to create the SummarizedExperiment object we mapped the rows of the experiments into GRanges. In order to map miRNA we used the miRNA from the annotation database TxDb.Hsapiens.UCSC.hg19.knownGene, this will exclude the miRNA from viruses and bacteria. In order to map genes genes alias, we used the biomart hg19 database (hsapiens_gene_ensembl from grch37.ensembl.org).

In case you prefere to have the raw data. You can get a data frame without any modification setting the summarizedExperiment to false.

```
library(TCGAbiolinks)
class(dataBRCA_df)
```

```
## [1] "data.frame"
```

```
dim(dataBRCA_df)
```

```
## [1] 20531     2
```

```
head(dataBRCA_df)
```

```
##              TCGA-BH-A1DE-01A-11R-A155-07 TCGA-BH-A0EA-01A-11R-A155-07
## ?|100130426                       0.0000                       0.0000
## ?|100133144                      11.5808                      32.0877
## ?|100134869                       4.1574                      12.5126
## ?|10357                         222.1458                     102.8808
## ?|10431                        1269.9778                     774.0168
## ?|136542                          0.0000                       0.0000
```

TCGAprepare: Table of types available for the TCGAprepare

- **RNASeqV2:** junction_quantification,rsem.genes.results, rsem.isoforms.results, rsem.genes.normalized_results, rsem.isoforms.normalized_results, bt.exon_quantification
- **RNASeq:** exon.quantification,spljxn.quantification, gene.quantification
- **genome_wide_snp_6:** hg18.seg,hg19.seg,nocnv_hg18.seg,nocnv_hg19.seg

TCGAprepare: Preparing the data with parameter - toPackage

This section will show how to integrate TCGAbiolinks with other packages. Our intention is to provide as many integrations as possible.

The example below shows how to use TCGAbiolinks with ELMER package (expression/methylation analysis). The TCGAprepare for the DNA methylation data will Removing probes with NA values in more than 0.80% samples and remove the annotation data, for the expression data it will take the log2(expression + 1) of the expression matrix in order to To inspect the relation between DNA methylation and expression also it will prepare the rownames as the standard by the package.

```
############## Get tumor samples with TCGAbiolinks
library(TCGAbiolinks)
query <- TCGAquery(tumor = "GBM",level = 3, platform = "HumanMethylation450")
# This function will take a lot of time depens on internet connection
TCGAdownload(query,path = "TCGA/GBM/")
exp <- TCGAprepare(query,dir = "TCGA/GBM",
                   save = TRUE,
                   filename = 'exp.rda',
                   toPackage = 'ELMER')
```

```
query.rna <- TCGAquery(tumor="GBM",level=3, platform="IlluminaHiSeq_RNASeqV2")
TCGAdownload(query.rna,path="TCGA/rna",type = "rsem.genes.normalized_results")
exp <- TCGAprepare(query.rna, dir="TCGA/rna", save = TRUE,
                   filename = "exp.rda",toPackage = "X.PUB")

############ To ENSEA
library(BLNER)

########## gene annotation
geneAnnot <- ten()
geneAnnot$SEMBID <- paste0("ID",geneAnnot$GENEID)
geneInfo <- promoters(geneAnnot,upstream = 0, downstream = 0)
########### probe
probe <- get.feature.probe()

rse.ghm.glial.with.exp <- fetch.mae(rmtch = ghm.glial.m,
                                    exp = exp,
                                    probeInfo = probe,
                                    TCGA = TRUE,
                                    geneInfo = geneInfo)
```

TCGAprepare: Preparing the data with CNV data (Genome_Wide_SNP_6)

You can easily search TCGA samples, download and prepare a matrix of gene expression

```
# Define a list of samples to query and download providing relative TCGA barcodes.
samplesList <- c("TCGA-02-0046-10A-01D-0182-01",
                 "TCGA-02-0003-01A-01D-0182-01",
                 "TCGA-02-0003-10A-01D-0182-01",
                 "TCGA-02-0054-01A-01D-0182-01",
                 "TCGA-02-0007-01A-01D-0182-01")

# Query platform Genome_Wide_SNP_6 with a list of barcode
query <- TCGAquery(tumor = "gbm", level = 3, platform = "Genome_Wide_SNP_6")

# Download a list of barcodes with platform Genome_Wide_SNP_6
TCGAdownload(query, path = "samples")

# Prepare matrix
GBM_CNV <- TCGAprepare(query, dir = "samples", type = ".hg19.seg.txt")
```

TCGAanalyze: Analyze data from TCGA.

You can easily analyze data using following functions:

TCGAanalyze_Preprocessing Preprocessing of Gene Expression data (IlluminaHiSeq_RNASeqV2).

You can easily search TCGA samples, download and prepare a matrix of gene expression

```
# You can define a list of samples to query and download providing relative TCGA barcodes.
```

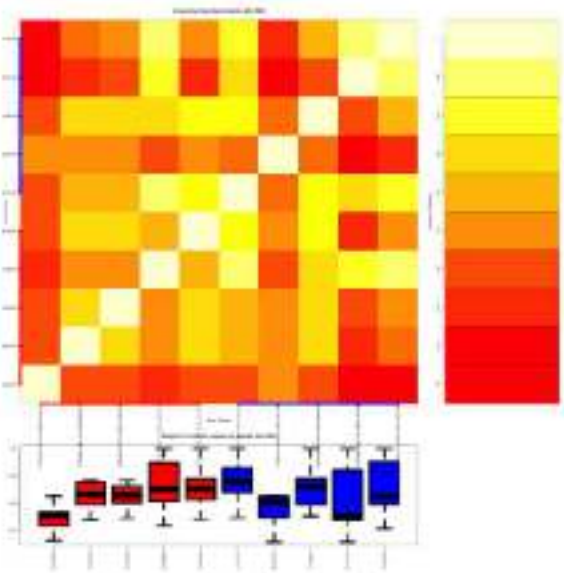
```
listSamples <- c('TCGA-E9-A1RG-11A-52R-A14P-07','TCGA-BH-A1FC-11A-32R-A13Q-07',
                 'TCGA-A7-A13G-11A-51R-A13Q-07','TCGA-BH-A0DK-11A-13R-A089-07',
                 'TCGA-E9-A1RE-11A-34R-A169-07','TCGA-BH-A0AU-01A-11R-A12P-07',
                 'TCGA-C8-A1EJ-01A-11R-A13Q-07','TCGA-A7-A13D-01A-13R-A12P-07',
                 'TCGA-A2-A0CV-01A-31R-A115-07','TCGA-A2-A0YD-01A-11R-A10R-07')

# Query platform Illumina HiSeq RNASeqV2 with a list of barcode
query <- TCGAquery(tumor = "brca", samples = listSamples,
  platform = "Illumina HiSeq_RNASeqV2", level = "3")

# dont run
#TCGAdownload(query, path = "dataBrca", type = "gene.quantification",samples = listSamples)

# Download a list of barcodes with platform IlluminaHiSeq_RNASeqV2
TCGAdownload(query, path = "../dataBrca", type = "raw_gene_results",samples = listSamples)

# Prepare expression matrix with geneID in rows and samples (barcode) in columns
# raw.gene.results as values
BRCARnaseq_assay <- TCGAprepare(query,"../dataBrca",type = "raw_gene_results")

BRCAMatrix <- assay(BRCARnaseq_assay,"raw_counts")

# For gene expression if you need to see a boxplot correlation and AAIC plot
# to define outliers you can run

BRCARnaseq_CorOutliers <- TCGAanalyze_Preprocessing(BRCAMatrix_assay)
```

The result is shown below:

Table 3: Example of a r
7 samples in columns)

	TCGA-BH-A1FC-11A-32R-A13Q-07	TCGA-A2-A0CV-01A-31R-A115-07	TCGA-A7-A13G-11A-51R-A13Q-0
DPAGT1|1798	1255.21	4105.45	1480.0
PI4KB|5298	5093.00	9095.00	4044.0
KRTAP4-8|728024	0.00	0.00	0.0
DNAJC24|120526	656.64	760.78	1073.6
HIRA|7290	1485.00	2014.00	2092.0
ZNF70|7595	145.00	208.00	25.0
SNHG10|283596	130.00	246.00	113.0
MEF2D|4209	2854.00	11778.00	3683.0
PUS7|50769	1329.00	3340.00	1064.0
EIF1B|10289	1722.00	4078.00	2389.0

The result from TCGAanalyze_Preprocessing is shown below:



TCGAanalyze_DEA & TCGAanalyze_LevelTab Differential expression analysis (DEA)

Perform DEA (Differential expression analysis) to identify differentially expressed genes (DEGs) using the `TCGAanalyze_DEA` function.

`TCGAanalyze_DEA` performs DEA using following functions from R edgeR:

1. edgeR::DGEList converts the count matrix into an edgeR object
2. edgeR::estimateCommonDisp each gene gets assigned the same dispersion estimate
3. edgeR::exactTest performs pair wise tests for differential expression between two groups.
4. edgeR::topTags takes the output from exactTest(), adjusts the raw p-values using the False Discovery Rate (FDR) correction, and returns the top differentially expressed genes.

This function receives as parameters:

- **mat1** The matrix of the first group (in the example group 1 is the normal samples).
- **mat2** The matrix of the second group (in the example group 2 is tumor samples)
- **Cond1type** Label for group 1
- **Cond2type** Label for group 2

After, we filter the output of dataDEGs by abs(logFC) >=1, and uses the `TCGAanalyze_LevelTab` function to create a table with DEGs (differentially expressed genes), log Fold Change (FC), false discovery rate (FDR), the gene expression level for samples in Cond1type, and Cond2type, and Delta value (the difference of gene expression between the two conditions multiplied logFC).

```
# Downstream analysis using gene expression data
# TCGA samples from Illumina HiSeq_RNASeqV2 with type rsem.genes.results
# save(dataBRCA, geneInfo , file = "dataDEGsExpression.rda")
library(TCGAbiolinks)

# Diff.expr.analysis (DEA)
```

```
dataDEGs <- TCGAanalyze_DEA(dataFilt[,samplesNT], dataFilt[,samplesTP],
                            "Normal", "Tumor")

# DEGs filter by abs(logFC) >=1
dataDEGsFilt <- dataDEGs[abs(dataDEGs$logFC) >= 1,]

# DEGs table with expression values in normal and tumor samples
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGsFilt,"Tumor","Normal",
                                          dataFilt[,samplesTP],dataFilt[,samplesNT])
```

The result is shown below:

Table 4: Table DEGs after DEA

mRNA	logFC	FDR	Tumor	Normal	Delta
FN1	2.88	1.23615e-19	347787.48	41734.19	1001017.5
COL1A1	1.77	1.690844e-08	356810.32	89293.72	633086.3
C4orf7	3.20	2.820474e-50	87821.36	2132.76	450425.4
COL1A2	1.49	9.490476e-05	272386.44	91241.32	363242.9
GAPDH	1.32	3.290676e-05	179057.44	63663.00	238255.5
CLEC3A	6.79	7.971002e-74	27257.16	259.60	185158.6
IGFBP5	1.94	1.090717e-04	126186.88	53323.12	158674.6
CPB1	4.27	3.044021e-37	37001.76	2637.72	157068.8
CARTPT	6.72	1.023191e-72	21700.96	215.16	145672.8
DCD	7.26	1.017983e-80	19911.20	81.80	144606.3

TCGAanalyze_EAcomplete & TCGAvisualize_EAbarplot: Enrichment Analysis

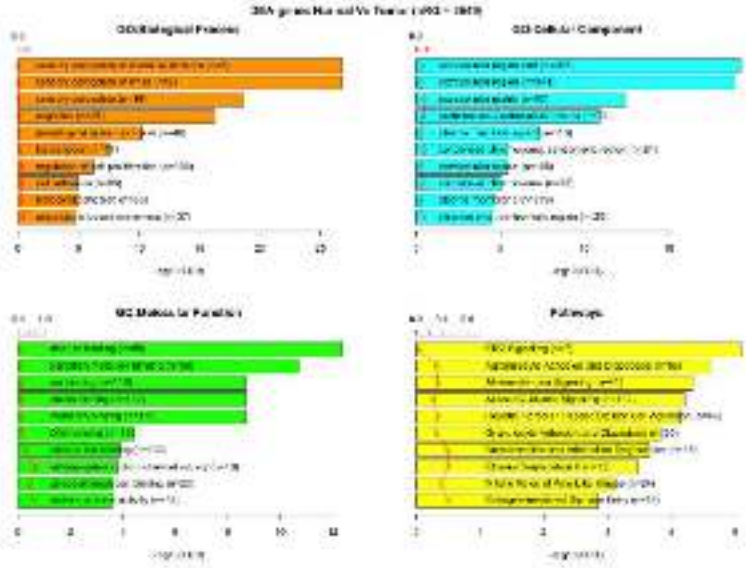
Researchers, in order to better understand the underlying biological processes, often want to retrieve a functional profile of a set of genes that might have an important role. This can be done by performing an enrichment analysis.

We will perform an enrichment analysis on gene sets using the `TCGAanalyze_EAcomplete` function. Given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find identify classes of genes or proteins that are over-represented using annotations for that gene set.

To view the results you can use the `TCGAvisualize_EAbarplot` function as shown below

```
library(TCGAbiolinks)
# Enrichment Analysis EA
# Gene Ontology (GO) and Pathway enrichment by Edge list
Genelist <- rownames(dataDEGsFiltLevel)

system.time(ansEA <- TCGAanalyze_EAcomplete(TFname="DEA genes Normal Vs Tumor",Genelist))

# Enrichment Analysis EA (TCGAVisualize)
# Gene Ontology (GO) and Pathway enrichment barplot

TCGAvisualize_EAbarplot(tf = rownames(ansEA$ResBP),
            GOBPTab = ansEA$ResBP,
            GOCCTab = ansEA$ResCC,
            GOMFTab = ansEA$ResMF,
            PathTab = ansEA$ResPat,
            nRGTab = Genelist,
            nBar = 10)
```

The result is shown below:



TCGAanalyze_survival Survival Analysis: Cox Regression and diet package

When analyzing survival times, different problems come up than the ones discussed so far. One question is how do we deal with subjects dropping out of a study. For example, assume that we test a new cancer drug. While some subjects die, others may believe that the new drug is not effective, and decide to drop out of the study before the study is finished. A similar problem would be faced when we investigate how long a machine lasts before it breaks down.

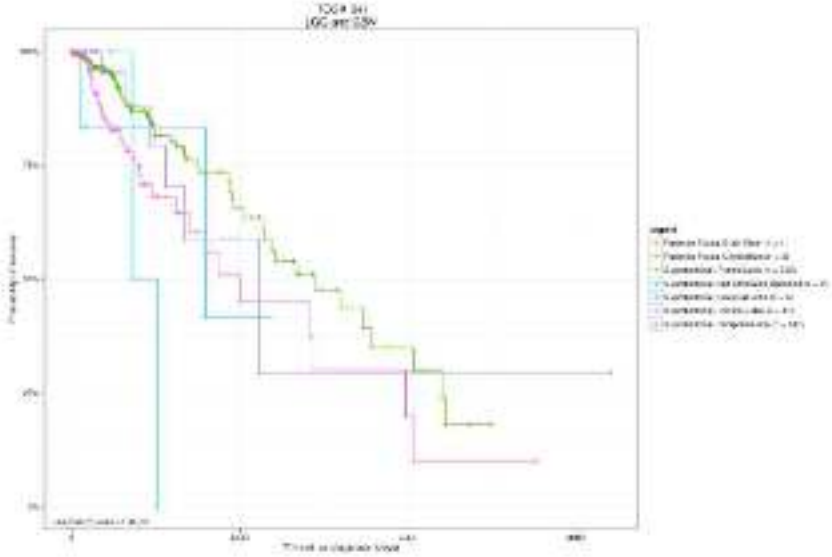
Using the clinical data, it is possible to create a survival plot with the function TCGAanalyze_survival as follows.

```
clin.gbm <- TCGAquery_clinic('gbm', 'clinical_patient')
clin.lgg <- TCGAquery_clinic('lgg', 'clinical_patient')

TCGAanalyze_survival(plyr::rbind.fill(clin.lgg,clin.gbm),
                     "radiation_therapy",
                     main = "TCGA Set\nLGG and GBM",height = 10, width=10)
```

The arguments of TCGAanalyze_survival are:

- **clinical_patient** TCGA Clinical patient with the information days_to_death
- **clusterCol** Column with groups to plot. This is a mandatory field, the caption will be based in this column
- **legend** Legend title of the figure
- **cutoff** xlim This parameter will be a limit in the x-axis. That means, that patients with days_to_death > cutoff will be set to Alive
- **main** main title of the plot
- **ylab** y-axis text of the plot
- **xlab** x-axis text of the plot
- **filename** The name of the pdf file
- **color** Define the colors of the lines

The result is shown below:



```
library(TCGAbiolinks)
# Survival Analysis SA
clinical_patient_Cancer <- TCGAquery_clinic("brca","clinical_patient")
dataBRCAcomplete <- log2(BRCA_rnaseqv2)

tokenStop<- 1

tabSurvKMcomplete <- NULL

for( i in 1: round(nrow(dataBRCAcomplete)/100)){
message( paste( i, "of ", round(nrow(dataBRCAcomplete)/100)))
tokenStart <- tokenStop
tokenStop <-100*i
tabSurvKM<-TCGAanalyze_SurvivalKM(clinical_patient_Cancer,dataBRCAcomplete,
                                  Genelist = rownames(dataBRCAcomplete)[tokenStart:tokenStop],
                                        Survresult = F,ThreshTop=0.67,ThreshDown=0.33)

tabSurvKMcomplete <- rbind(tabSurvKMcomplete,tabSurvKM)
}

tabSurvKMcomplete <- tabSurvKMcomplete[tabSurvKMcomplete$pvalue < 0.01,]
tabSurvKMcomplete <- tabSurvKMcomplete[!duplicated(tabSurvKMcomplete$mRNA),]
rownames(tabSurvKMcomplete) <-tabSurvKMcomplete$mRNA
tabSurvKMcomplete <- tabSurvKMcomplete[,-1]
tabSurvKMcomplete <- tabSurvKMcomplete[order(tabSurvKMcomplete$pvalue, decreasing F),]

tabSurvKMcompleteDEGs <- tabSurvKMcomplete[rownames(tabSurvKMcomplete) %in% dataDEGsFiltLevel$mRNA,]
```

The result is shown below.

Table 5: Table KM-survival genes after SA

	pvalue	Cancer Deaths	Cancer Deaths with Top	Cancer Deaths with Down	Mean Tumor Top	Mean Tu
DCTPP1	6.2941/0e-08	66	46	20	13.31	
APOO	9.390193e-06	65	49	16	11.49	
LOC387646	1.036097e-05	69	48	21	7.97	
PGK1	1.196577e-05	71	49	22	15.06	
CCNE2	2.100348e-05	65	48	17	11.07	
CCDC76	2.920614e-05	74	46	28	9.47	
FGD5	3.039936e-05	69	23	46	12.30	
FAM168B	3.575856e-05	68	25	43	6.82	
MMP28	3.702351e-05	70	17	53	8.55	
ADHFE1	3.907108e-05	67	22	45	9.34	

TCGAvisualize: Visualize results from analysis functions with TCGA's data.

You can easily visualize results from some following functions:

TCGAvisualize_PCA: Principal Component Analysis plot for differentially expressed genes

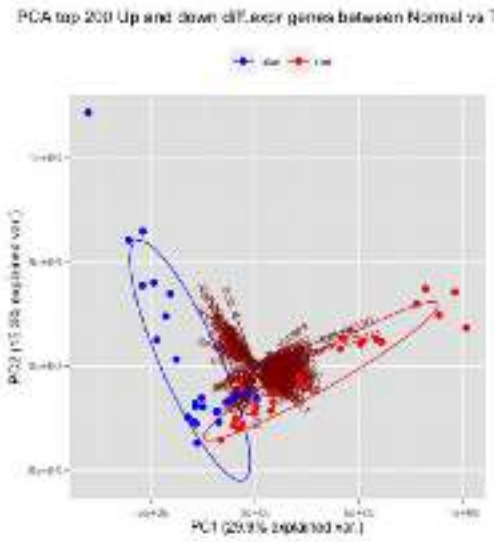
In order to understand better our genes we can perform a PCA to reduce the number of dimensions of our gene set. The function TCGAvisualize_PCA will plot the PCA for different groups

The parameters of this function are:

- **dataFilt** The expression matrix after normalization and quantile filter
- **dataDEGsFiltLevel** The TCGAanalyze_LevelTab output
- **ntopgenes** number of DEGs genes to plot in PCA

```
library(TCGAbiolinks)

# normalization of genes
dataNorm <- TCGAbiolinks::TCGAanalyze_Normalization(dataBRCA, geneInfo)

# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(dataNorm, 0.25)

# Principal Component Analysis plot for ntop selected DEGs
TCGAvisualize_PCA(dataFilt,dataDEGsFiltLevel, ntopgenes = 200)

# boxplot of normalized data
#sample0ones <- rownames(dataDEGsFilt[dataDEGsFilt$logFC >=1,])[1:20]
#boxplot(log(dataBRCA[sample0genes,]), las = 2)
#boxplot(log(dataFilt[sample0genes,]), las = 2)
```

The result is shown below:



TCGAvisualize_SurvivalCoxNET Survival Analysis: Cox Regression and dnet package

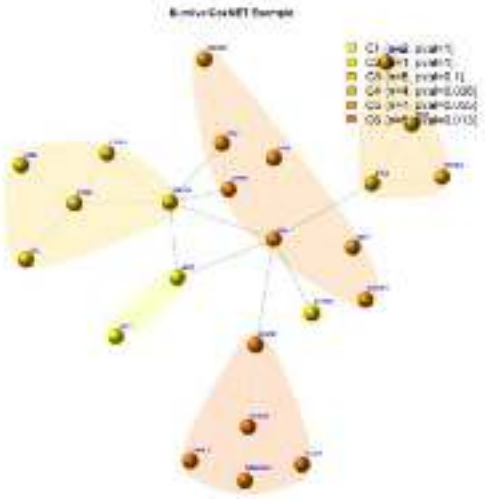
TCGAvisualize_SurvivalCoxNET can help an user to identify a group of survival genes that are significant from univariate Kaplan Meier Analysis and also for Cox Regression. It shows in the end a network build with community of genes with similar range of pvalues from Cox regression (same color) and that interaction among those genes is already validated in literatures using the STRING database (version 9.1)

```
library(TCGAbiolinks)
# Survival Analysis SA

clinical_patient_Cancer <- TCGAquery_clinic("brca","clinical_patient")
dataBRCAcomplete <- log2(BRCA_rnaseqv2)

tokenStop<- 1

tabSurvKMcomplete <- NULL

for( i in 1: round(nrow(dataBRCAcomplete)/100)){
message( paste( i, "of ", round(nrow(dataBRCAcomplete)/100)))
tokenStart <- tokenStop
tokenStop <-100*i
tabSurvKM<-TCGAanalyse_SurvivalKM(clinical_patient_Cancer,
                                  dataBRCAcomplete,
                                  Genelist = rownames(dataBRCAcomplete)[tokenStart:tokenStop],
                                  Survresult = F,ThreshTop=0.67,ThreshDown=0.33)

tabSurvKMcomplete <- rbind(tabSurvKMcomplete,tabSurvKM)
}
```

```
tabSurvKMcomplete <- tabSurvKMcomplete[tabSurvKMcompleteSpvalue < 0.01,]
tabSurvKMcomplete <- tabSurvKMcomplete[!duplicated(tabSurvKMcompleteSmRNA),]
rownames(tabSurvKMcomplete) <-tabSurvKMcompleteSmRNA
tabSurvKMcomplete <- tabSurvKMcomplete[,-1]
tabSurvKMcomplete <- tabSurvKMcomplete[order(tabSurvKMcompleteSpvalue, decreasing=F),]

tabSurvKMcompleteDEGs <- tabSurvKMcomplete[rownames(tabSurvKMcomplete) %in% dataDEGsFiltLevel$mRNA,]

tflist <- EAGenes[EAGenes$Family == "transcription regulator","Genes"]
tabSurvKMcomplete_onlyTF <- tabSurvKMcomplete[rownames(tabSurvKMcomplete) %in% tflist,]

TabCoxNet <- TCGAvisualize_SurvivalCoxNET(clinical_patient_Cancer,dataBRCAcomplete,
                          Genelist = rownames(tabSurvKMcomplete_onlyTF),
                          scoreConfidence = 700,titlePlot = "TCGAvisualize_SurvivalCoxNET Example")
```

In particular the survival analysis with kaplan meier and cox regression allow user to reduce the feature / number of genes significant for survival. And using 'dnet' pipeline with 'TCGAvisualize_SurvivalCoxNET' function the user can further filter those genes according some already validated interaction according STRING database. This is important because the user can have an idea about the biology inside the survival discrimination and further investigate in a sub group of genes that are working in as synergistic effect influencing the risk of survival. In the following picture the user can see some communities of genes with same color and survival pvalue.

The result is shown below:



TCGA Downstream Analysis some workflows and pipelines

Downstream Analysis n.1

After preparing the gene expression from TCGA data using the TCGAprepare function, you can do a normalization of genes using the function TCGAanalyze_Normalization, do a quantile filter of genes with the TCGAanalyze_Filtering function.

TCGAanalyze_Normalization allows user to normalize mRNA transcripts and miRNA, using R EDASeq package. Normalization for RNA-Seq Numerical and graphical summaries of RNA-Seq read data. Within-lane normalization procedures to adjust for GC-content effect (or other gene-level effects) on read counts: loess robust local regression, global-scaling, and full-quantile normalization (Risso, Davide and Schwartz, Katja and Sherlock, Gavin and Dudoit, Sandrine 2011). Between-lane normalization procedures to adjust for distributional differences between lanes (e.g., sequencing depth): global scaling and full quantile normalization (Bullard, James H and Purdom, Elizabeth and Hansen, Kasper D and Dudoit, Sandrine 2010).

For istance returns all mRNA or miRNA with mean across all samples higher than the threshold defined quantile mean across all samples.

Also, in order to classify your samples (barcode) you can use the TCGAquery_SampleTypes function, the typeSample 'NT' will return the "Solid Tissue Normal"' samples, while the typeSample 'TP' will return "Primary Solid Tumor"' samples.

```
# Downstream analysis using gene expression data
# TCGA samples from Illumina HiSeq_RNASeqV2 with type exon.gene.rsem.txt

library(TCGAbiolinks)

# dataBRCA in TCGAbiolinks package is a table from TCGA BRCA [10 samples] and comes from
# BRCAMatrix <- TCGAprepare(query,"dataBrca") from above example
# dataBRCA <- BRCAMatrix

# normalization of genes
dataNorm <- TCGAbiolinks::TCGAanalyze_Normalization(dataBRCA, geneInfo)

# quantile filter of genes
dataFilt <- TCGAanalyze_Filtering(dataNorm, 0.25)

# selection of normal samples "NT"
samplesNT <- TCGAquery_SampleTypes(colnames(dataFilt), typesample = c("NT"))

# selection of tumor samples "TP"
samplesTP <- TCGAquery_SampleTypes(colnames(dataFilt), typesample = c("TP"))
```

Downstream Analysis n.2 IlluminaHiSeq_RNASeq data

You can easily search TCGA samples, download and prepare a matrix of gene expression.

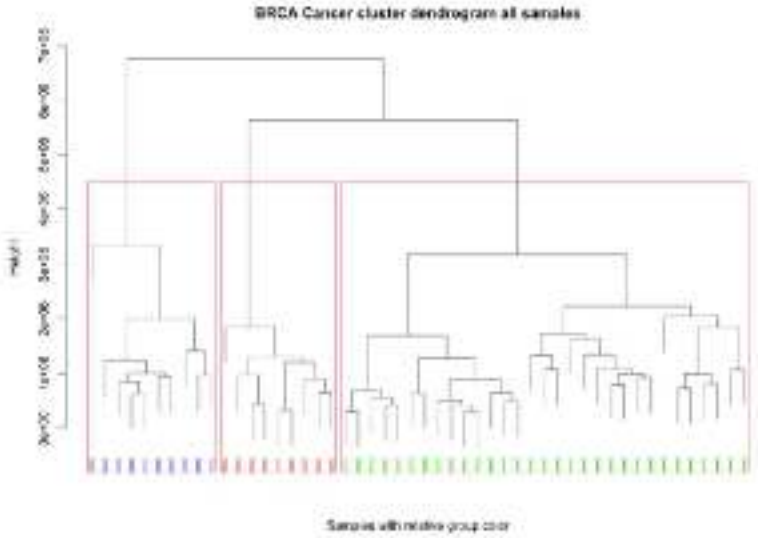
```
# Query platform Illumina HiSeq_RNASeq within a list of barcode
query <- TCGAquery(level = "3rd", platform = "IlluminaHiSeq_RNASeq", level = "3")

# You can define a list of samples to query and download providing relative TCGA barcodes.
listSamples <- TCGAquery_samplesfilter(query)

# Download only first 6 samples for test.

TCGAdownload(query, path = "dataBrca", type = "gene.quantification",
             samples = listSamplesIlluminaHiSeq_RNASeq[1:6])
```

```
# Prepare expression matrix with gene id in rows and samples (barcode) in columns
# rows.genes.results as values
BRCAMatrix <- TCGAprepare(query,'dataBrca',type = 'geneExpression')
```

Downstream Analysis n.3 LGG and GBM Integration (Heatmap and Cluster)

```
library(TCGAbiolinks)
library(genefilter)
library(clue)

BRCArnaseqV2 <- dataBRCA
BRCArnaseqV2mostVar <- varFilter(BRCArnaseqV2, var.func = IQR,  var.cutoff = 0.75,
                                 filterByQuantile = TRUE)

sData <- t(BRCArnaseqV2mostVar)
ddist <- dist(sData, method = 'euclidean')
dHc <- hclust(ddist, method = "ward.D")

plot(dHc, labels = FALSE, main ="UNCA Cancer cluster Sendrogram all samples",
     xlab = 'Samples with relative group color',sub '')

rect.hclust(dHc, k=3, border="red")
tabCluster <- as.matrix(cutree(dHc, k = 3))
colnames(tabCluster)<-"Cluster"
tabCluster<-cbind(Sample = rownames(tabCluster),Color = rownames(tabCluster), tabCluster)
tabCluster<-as.data.frame(tabCluster)
tabCluster<-tabCluster[order(tabCluster$Cluster,decreasing = FALSE),]
tabCluster<-as.data.frame(tabCluster)
tabCluster$Color<-as.character(tabCluster$Color)

ccol <- palette()[1 + 1:3]

for( cc in 1:3){
  tabCluster[tabCluster[, "Cluster"] == cc, "Color"] <- ccol[cc]
}

tabCluster <- tabCluster[dHc$labels, ]

rug(which(tabCluster[dHc$order, 'Color'] == "blue"), col = "blue", lwd = 3)
rug(which(tabCluster[dHc$order, 'Color'] == "green3"), col = "green3", lwd = 3)
rug(which(tabCluster[dHc$order, 'Color'] == "red"), col = "red", lwd = 3)
```

The result is shown below:

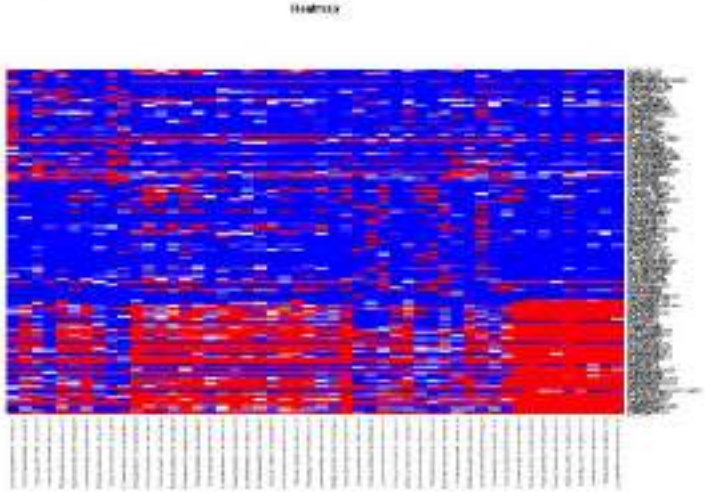



```
library(TCGAbiolinks)


### Differential analysis
GroupBlueData <- BRCAmrnaseqV2[, as.character(tabCluster[tabCluster$Color ==
    'blue', 'Sample'])]
GroupGreen3Data <- BRCAmrnaseqV2[, as.character(tabCluster[tabCluster$Color ==
    'green3', 'Sample'])]
GroupRedData <- BRCAmrnaseqV2[, as.character(tabCluster[tabCluster$Color ==
    'red', 'Sample'])]

DEGsBlue <- TCGAanalyze_DEA(cbind(GroupGreen3Data, GroupRedData), GroupBlueData,
    'GroupOther', 'GroupBlue')
DEGsGreen3 <- TCGAanalyze_DEA(cbind(GroupBlueData, GroupRedData), GroupGreen3Data,
    'GroupOther', 'GroupGreen3')
DEGsRed <- TCGAanalyze_DEA(cbind(GroupBlueData, GroupGreen3Data), GroupRedData,
    'GroupOther', 'GroupRed')


dataDEGs <- TCGAanalyze_DEA(dataFilt[, samplesNT], dataFilt[, samplesTP], "Normal",
    'Tumor')

# DEGs filter by abs(logFC) >=1
dataDEGsFilt <- dataDEGs[abs(dataDEGs$logFC) >= 1, ]

dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGsFilt, 'Tumor', 'Normal', dataFilt[,
    samplesTP], dataFilt[, samplesNT])

DEGsBlue_aval <- TCGAanalyze_LevelTab(DEGsBlue, 'GroupBlue', 'GroupOther', GroupBlueData,
```

```
    rbind(GroupGreen3Data, GroupRedData), typeOrder = TRUE)
DEGsGreen3Level <- TCGAanalyze_LevelTab(DEGsGreen3, "GroupGreen3", "GroupOther",
    GroupGreen3Data, rbind(GroupBlueData, GroupRedData), typeOrder = TRUE)
DEGsRedLevel <- TCGAanalyze_LevelTab(DEGsRed, "GroupRed", "GroupOther", GroupRedData,
    rbind(GroupBlueData, GroupGreen3Data), typeOrder = TRUE)

blueDEGs <- DEGsBlueLevel[DEGsBlueLevel$FDR < 0.01 & DEGsBlueLevel$logFC >=
    1, ]
blueDEGs <- blueDEGs[order(blueDEGs$FDR), ]
green3DEGs <- DEGsGreen3Level[DEGsGreen3Level$FDR < 0.01 & DEGsGreen3Level$logFC >=
    1, ]
green3DEGs <- green3DEGs[order(green3DEGs$FDR), ]
redDEGs <- DEGsRedLevel[DEGsRedLevel$FDR < 0.01 & DEGsRedLevel$logFC >=
    1, ]
redDEGs <- redDEGs[order(redDEGs$FDR), ]

blueDEGsSpec <- blueDEGs[setdiff(rownames(blueDEGs), union(rownames(green3DEGs),
    rownames(redDEGs))), ]
green3DEGsSpec <- green3DEGs[setdiff(rownames(green3DEGs), union(rownames(blueDEGs),
    rownames(redDEGs))), ]
redDEGsSpec <- redDEGs[setdiff(rownames(redDEGs), union(rownames(blueDEGs),
    rownames(green3DEGs))), ]

blueDEGsSpec <- blueDEGsSpec[1:50, ]
green3DEGsSpec <- green3DEGsSpec[1:50, ]
redDEGsSpec <- redDEGsSpec[1:50, ]

tabCluster <- tabCluster[order(tabCluster$Color), ]

MFiltQuantileOrdered <- BRCAexpressQN2[c(rownames(blueDEGsSpec), rownames(green3DEGsSpec),
    rownames(redDEGsSpec)), rownames(tabCluster)]

MFactivity <- t(MFiltQuantileOrdered)

HMactivity <- MFactivity
threshold_quantile <- 0.75
HMactivity[HMactivity >= quantile(MFactivity, threshold_quantile)] <- quantile(HMactivity,
    threshold_quantile)

summary(as.vector(HMactivity))
quantile(HMactivity, 0.15)
quantile(HMactivity, 0.85)
HMactivity[HMactivity <= quantile(HMactivity, 0.15)] <- quantile(HMactivity,
    0.15)
HMactivity[HMactivity >= quantile(HMactivity, 0.85)] <- quantile(HMactivity,
    0.85)

column_annotation <- matrix(" ", nrow = nrow(HMactivity), ncol = 1)
column_annotation[, 1] <- tabCluster$Color

row_annotation <- matrix(" ", nrow = 1, ncol = ncol(HMactivity))
row_annotation[1, ] <- c(rep('blue', nrow(blueDEGsSpec)), rep('green3',
    nrow(green3DEGsSpec)), rep('red', nrow(redDEGsSpec)))
```

```
library("GMD")

png("BRCA_heatmap.png", width = 1200, height = 800)
heatmap.2(t(DMAactivity), ColSideColors = column_annotation, RowSideColors = row_annotation,
          key = FALSE, Colv = NA, Rowv = NA,
          scale = 'none',
          #col = greenred(75),
          dendrogram = "none",
          #lwidth = 50, lmat=c( = 50,
          margins = c(1, 8), side.height.fraction = 0.25, keysize = 1.4, cexRow = 1.6)
dev.off()
```

The result is shown below:



Downstream Analysis n.4 DNA methylation analysis

Some downstream analysis from DNA methylation data can be done with TCGAbiolinks. An example is shown below. Firstly, we search, download and prepare data from the HumanMethylation450 platform for the GBM tumor and also get the clinical information from the patients. In this case, we will have a SummarizedExperiment object, where the rows are the probes and the columns the samples. For more information about this object you can take a look in the documentation with the command ?SummarizedExperiment

```
library(TCGAbiolinks)

# Getting the data
query <- TCGAquery(tumor = "gbm", platform = "HumanMethylation450", level = 3)
TCGAdownload(query,path=".")
data <- TCGAprepare(query,dir = ".",save = T)
clinical <- TCGAquery_clinic("gbm","clinical_patient")
```

```
#Preprocessing
# Remove probes with NA level
data <- subset(data,subset =(rowSums(is.na(assay(data)))!=0))
```

As an example, we divided the data into groups in order to analyze the data.

```
# random split of patients into groups
clinical$group <- c(rep("group1",nrow(clinical)/4),
                    rep("group2",nrow(clinical)/4),
                    rep("group3",nrow(clinical)/4),
                    rep("group4",nrow(clinical)-3*(floor(nrow(clinical)/4))))

 colData(data)$group <- c(rep("group1",ncol(data)/2), rep("group2",ncol(data)/2))
```

TCGAvisualize_meanMethylation: Sample Mean DNA Methylation Analysis

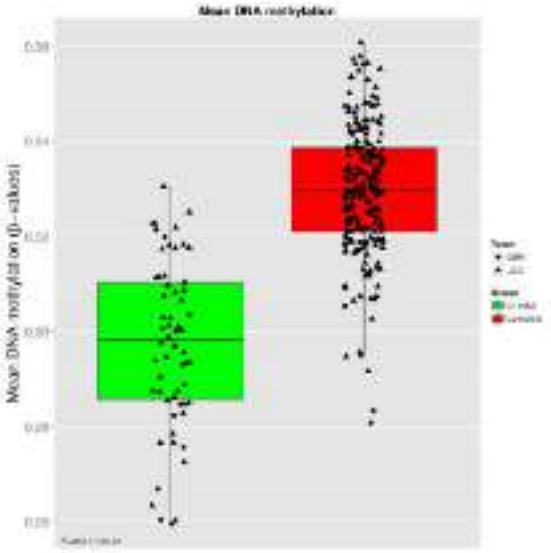
Using the data and calculating the mean DNA methylation per group, it is possible to create a mean DNA methylation boxplot with the function TCGAvisualize_meanMethylation as follows:

```
TCGAvisualize_meanMethylation(data,"group")
```

The arguments of TCGAvisualize_meanMethylation are

- **data** SummarizedExperiment object obtained from TCGAprepare
- **groupCol** Columns in colData(data) that defines the groups. If no columns defined a column called "Patients" will be used
- **subgroupCol** Columns in colData(data) that defines the subgroups
- **shapes** Shape vector of the subgroups. It must have the size of the levels of the subgroups. Example: shapes = c(21,23) if 2a two levels
- **filename** The name of the pdf that will be saved
- **subgroup.legend** Name of the subgroup legend. **DEFAULT: subgroupCol**
- **group.legend** Name of the group legend. **DEFAULT: groupCol**
- **color** vector of colors to be used in graph
- **title** main title in the plot
- **ylab** y axis text in the plot
- **print.pvalue** Print p value for two groups in the plot
- **xlab** x axis text in the plot
- **labels** Labels of the groups

The result is shown below:



TCGAanalyze_DMR: Differentially methylated regions Analysis

We will search for differentially methylated CpG sites using the TCGAanalyze_DMR function. In order to find these regions we use the beta values (methylation values ranging from 0.0 to 1.0) to compare two groups.

Firstly, it calculates the difference between the mean DNA methylation of each group for each probe.

Secondly, it calculates the p value using the wilcoxon test adjusting by the Benjamini Hochberg method. The default parameters was set to require a minimum absolute beta values difference of 0.2 and a p value adjusted of < 0.01.

After these analysis, we save a volcano plot (x axis:dif mean methylation, y axis: significance) that will help the user identify the differentially methylated CpG sites and return the object with the calculus in the rowRanges.

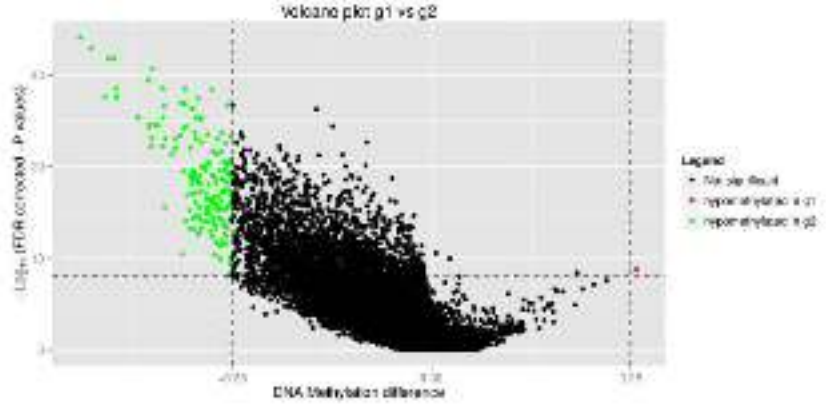
The arguments of volcanoPlot are

- **data** SummarizedExperiment obtained from the TCGAPrepare
- **groupCol** Columns with the groups inside the SummarizedExperiment object. (This will be obtained by the function colData(data))
- **group1** In case our object has more than 2 groups, you should set the name of the group
- **group2** In case our object has more than 2 groups, you should set the name of the group
- **filename** pdf filename. Default: volcano.pdf
- **legend** Legend title
- **color** vector of colors to be used in graph
- **title** main title. If not specified it will be "Volcano plot (group1 vs group2)"
- **ylab** y axis text
- **xlab** x axis text
- **xlim** x limits to cut image
- **ylim** y limits to cut image
- **label** vector of labels to be used in the figure. Example: c("1" = "Not Significant", "2" = "Hypermethylated in group1", "3" = "Hypomethylated in group1"))

- **p.cut** p value threshold. *Default: 0.01*
- **diffmean.cut** diffmean threshold. *Default: 0.2*
- **adj.method** Adjusted method for the p-value calculation
- **paired** Wilcoxon paired parameter. Default: *FALSE*
- **overwrite** Overwrite the pvalues and diffmean values if already in the object for both groups? *Default: FALSE*

```
data <- TCGAanalyze_DMR(data, groupCol = "cluster.sch",subgroupCol = "disease",
                              group.legend  = "Groups", subgroup.legend = "Tumor",
                              print.pvalue = TRUE)
```

The output will be a plot such as the figure below. The green dots are the probes that are hypomethylated in group 1 compared to group 2, while the red dots are the hypermethylated probes in group 1 compared to group 2.



Also, the TCGAanalyze_DMR function will save the plot as pdf and return the same SummarizedExperiment that was given as input with the values of p.value, p value adjusted, diffmean and the group it belongs in the graph (non significant, hypomethylated, hypermethylated) in the rowRanges. The columns will be (where group1 and group2 are the names of the groups):

- diffmean.group1.group2
- p.value.group1.group2
- p.value.adj.group1.group2
- status.group1.group2

This values can be view/accessed using the rowRanges accessor (rowRanges(data)).

Observation: Calling the same function again, with the same arguments will only plot the results, as it was already calculated. With you want to have them recalculated, please set **overwrite** to TRUE or remove the calculated columns.

TCGAvisualize_starburst: Analyzing expression and methylation together

The starburst plot is proposed to combine information from two volcano plots, and is applied for a study of DNA methylation and gene expression. In order to reproduce this plot, we will use the TCGAvisualize_starburst function.

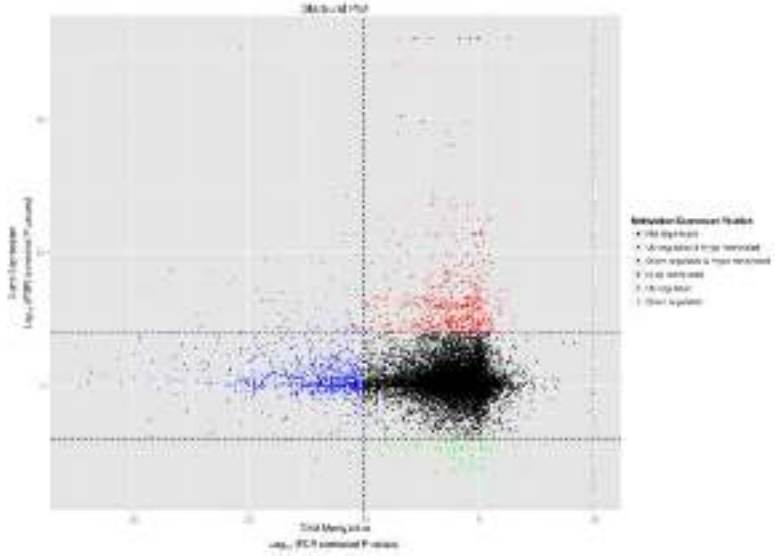
The function creates Starburst plot for comparison of DNA methylation and gene expression. The log10 (FDR corrected P value) is plotted for beta value for DNA methylation (x axis) and gene expression (y axis) for each gene. The black dashed line shows the FDR-adjusted P value of 0.01.

The parameters of this function are:

- **met** SummarizedExperiment with methylation data obtained from the TCGAprepare and processed by TCGAanalyze_DMR function. Expected colData columns: diffmean and p.value.adj
- **exp** Matrix with expression data obtained from the TCGAanalyze_DEA function. Expected colData columns: logFC, FDR
- **filename** pdf filename
- **legend** legend title
- **color** vector of colors to be used in graph
- **label** vector of labels to be used in graph
- **title** main title
- **ylab** y axis text
- **xlab** x axis text
- **xlim** x limits to cut image
- **ylim** y limits to cut image
- **p.cut** p value cut off
- **group1** The name of the group 1. Obs: Column p.value.adj.group1.group2 should exist
- **group2** The name of the group 2. Obs: Column p.value.adj.group1.group2 should exist

```
result <- TCGAvisualize_starburst(met,exp,"g1","g2",p.cut = 0.05)
```

As result the function will a plot the figure below and return a matrix with The Gene_symbol and it status in relation to expression(up regulated/down regulated) and methylation (Hyper/Hypo methylated)



TCGAinvestigate: Searching questions, answers and literature

TCGAinvestigate: Find most studied TFs in pubmed

Find most studied TFs in pubmed related to a specific cancer, disease, or tissue

```
# First perform DEGs with TCGAanalyze
# See previous section
library(TCGAbiolinks)

# Select only transcription factors (TFs) from DEGs
TFs <- TCGAgenes[TCGAgenes$type == "transcription regulator",]
TFs_inDEGs <- intersect(TFs$Gene, dataDEGsFiltLevel$mRNA )
dataDEGsFiltLevelTFs <- dataDEGsFiltLevel[TFs_inDEGs,]

# Order table DEGs TFs according to Delta decrease
dataDEGsFiltLevelTFs <- dataDEGsFiltLevelTFs[order(dataDEGsFiltLevelTFs$Delta,decreasing = TRUE),]

# Find Pubmed of TF studied related to cancer
tabDEGsTFPubmed <- TCGAinvestigate("breast", dataDEGsFiltLevelTFs, topgenes = 10)
```

The result is shown below:

Table 6: Table with most studied TF in pubmed related to a specific cancer

mRNA	logFC	FDR	Tumor	Normal	Delta	Pubmed	PMID
MUC1	2.46	0	58436.56	6469.40	94525.36	627	26016502; 25680764; 25682681; 25973571;
EGR1	-2.46	0	14080.32	60543.24	34627.41	513	26011749; 25656606; 25824686; 25788830;
MDM2	1.41	0	16132.29	4959.92	22024.14	441	26042502; 26031071; 25814189; 25805170;
GATA3	1.38	0	29394.60	8804.72	66410.03	180	26028330; 26008346; 25994056; 25805123;
FOXA1	1.45	0	16179.96	5378.58	23465.63	167	26008346; 25995231; 25994056; 25762479;
EGR1	-2.44	0	16073.08	74647.28	30275.29	77	25703526; 24680816; 24742492; 24675612;
TCEB1	1.43	0	17706.96	6260.08	25470.20	13	25798844; 23589165; 23162035; 21937081;
MAGEL2	1.18	0	20850.16	8246.32	24633.09	5	24225485; 23884293; 22035435; 21618023;
PTK6	1.72	0	19200.12	44252.52	26104.62	5	25945613; 25214712; 21913217; 20427076;
ILF2	1.27	0	22250.32	7854.44	26206.23	0	0

TCGAsocial: Searching questions,answers and literature

The TCGAsocial function has two type of searches, one that searches for most downloaded packages in CRAN or BioConductor and one that searches the most related question in biostar.

TCGAsocial with BioConductor

Find most downloaded packages in CRAN or BioConductor

```
library(TCGAbiolinks)

# Define a list of package to find number of downloads
listPackage <-c('limma','edgeR','survcomp')

tabPackage <- TCGAsocial(siteToFind ="bioconductor.org",listPackage)

# define a keyword to find in support.bioconductor.org returning a table with suggested packages
tabPackageKey <- TCGAsocial(siteToFind ="support.bioconductor.org" ,KeyInfo = "edgeR")
```

The result is shown below:

Table 7: Table with number of downloads about a list of packages

Package	NumberDownload
limma	70749
edgeR	33531
survcomp	3961

Table 8: Find most related question in support.bioconductor.org with keyword = tcga

question	BiostarsSite	PackageSuggested
A: Calculating Ibd Using R Package	/35461/	IBIN
A: How To Identify Rotamer States From A Pdb ?	/96579/	SIM
A: Pathway Analysis In R	/14316/	sigPathway
A: Ngs Question – Consensus	/17535/	sigPathway
A: How to read .bim file in Rsamtools R package?	/97978/	Rsamtools
A: Best Practices/Software To Calculate Ka/Ks Ratio	/3817/#	ks
A: Trouble With Local Psiblast	/79246/	ks
A: R Package For Annotations Of Genomic Regions	/43313/	ks
A: Question About Multip Methylation Array	/69357/	LEA;MEDIPS
A: Find Out The Genes That Correspond To My Coordinates	/47836/	ChIPpeakAnno
Mirna Sequence Using Biomart R Package	/96700/	biomaRt
A: Annotating Expression Profile Data	/60864/	AnnotationDbi;LEA
A: How to generate a Venn diagram	/102393	0
A: CNV calling for illumina 550k array	/108029	0
A: Error: could not find function "heatmap.2"	/106843	0
A: Extracting Probeset IDs from .CEL files	/135942	0
A: Bam to nucleotide frequencies	/100798	0
A: Gene Regulatory Network using micro array data	/121070	0
A: R programming question: insert alternately	/139129	0
A: Ignoring Ns on Each Side of the Chromosome	/146815	0
** MISSING ***	NA	0
A: GSeq rat genome	/139732	0

TCGAsocial with Biostar

Find most related question in biostar

```
library(TCGAbiolinks)

# Find most related question in biostar with TCGA
tabPackage1 <- TCGAsocial(SiteToFind ="biostars.org",keyinfo = "TCGA")

# Find most related question in biostar with package
tabPackage2 <- TCGAsocial(SiteToFind ="biostars.org",KeyInfo = "package")
```

The result is shown below:

Table 9: Find most related question in biostar with TCGA

question	BiostarsSite	PackageSuggested
A: Question About Tcga Snp Array Data	/89341/	LEA;PROcess;ROC

question	BiostarsSite	PackageSuggested
A: Csv Data	/95763/	DNAcopy HELP
A: Csv Data	/95763/	DNAcopy HELP
A: Where To Find Test Datasets For Data Classification Problems	/60604/	convert;GEOquery;LEA;rMAT;rpart;SIM
A: How to get public cancer RNA seq data?	/13770	0
A: Microarray And Epigenomic Data For Same Cancer Cell Line?	/95724/	0

Table 10: Find most related question in biostar web package

question	BiostarsSite	PackageSuggested
A: Calculating Ibd Using R Package	/85461/	TIN
A: How To Identify Sstamir Status From A Fdb ?	/96579/	SIM
A: Pathway Analysis In R	/14316/	sigPathway
A: Ngs Question -- Consensus	/17535/	sigPathway
A: How to read .bam file in Rsamtools R package?	/9799/8/	Rsamtools
A: Best Practices/Software To Calculate Ks/Ks Ratio	/5817/#	lea
A: Trouble With Local Fpkmst	/79246/	lea
A: R Package For Annotation Of Genomic Regions	/43313/	lea
A: Question About Medip Methylation Array	/89357/	LEA;MEDIPS
A: Find Out The Genes That Correspond To My Coordinates	/47926/	ChIPpeak/vmo
Mirna Sequence Using Biomart R Package	/96708/	biomaRt
A: Annotating Expression Profile Data	/68964/	AnnotationDbi;LEA
A: How to generate a Venn diagram	/102393	0
A: CNV calling for Illumina bulk array	/109029	0
A: Error: could not find function "heatmap.2"	/105643	0
A: Extracting Probeset IDs from .CEL files	/135942	0
A: Bam to nucleotide frequencies	/109798	0
A: Gene Regulatory Network using micro array data	/121073	0
A: R programming question: insert alternately	/139129	0
A: Ignoring Ns on Each Side of the Chromosome	/145613	0
*** MISSING ***	NA	0
A: rSeqc not genome	/135732	0

Session Information

```
sessionInfo()
## R version 3.2.1 (2015-06-18)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.4 (Yosemite)
##
## locale:
## [1] pt_BR.UTF-8/pt_BR.UTF-8/pt_BR.UTF-8/C/pt_BR.UTF-8/pt_BR.UTF-8
##
## attached base packages:
##  [1] grid      stats4    parallel  stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
```

```
## other attached packages:
##  [1] png_0.1-7                   SummarizedExperiment_0.8.2
##  [3] Biobase_2.29.1              GenomicRanges_1.21.17
##  [5] GenomeInfoDb_1.5.9          IRanges_2.3.17
##  [7] S4Vectors_0.7.12            BiocGenerics_0.15.5
##  [9] TCGAbiolinks_0.99.1         BiocStyle_1.7.6
##
## loaded via a namespace (and not attached):
##  [1] xtable_1.8-0
##  [2] edgeR_3.11.2
##  [3] splines_3.2.1
##  [4] R.utils_2.1.0
##  [5] bigmemory_4.5.19
##  [6] biganalytics_1.1.1
##  [7] latticeExtra_0.6-25
##  [8] XVector_0.9.1
##  [9] coin_1.0-24
## [10] RCurl_1.95-4.7
## [11] yaml_2.1.13
## [12] RSQLite_1.0.0
## [13] lattice_0.20-33
## [14] limma_3.25.14
## [15] downloader_0.4
## [16] chron_2.3-47
## [17] digest_0.6.9
## [18] RColorBrewer_1.1-2
## [19] GenomicFeatures_1.21.13
## [20] rmeta_2.2.0
## [21] colorspace_1.2-6
## [22] Matrix_1.2-2
## [23] xtable_1.8-0
## [24] R.oo_1.19.0
## [25] plyr_1.8.3
## [26] XML_3.98-1.3
## [27] devtools_1.9.0
## [28] ShortRead_1.27.5
## [29] stringr_1.0.0
## [30] genefilter_1.51.0
## [31] zlibbioc_1.15.0
## [32] xtable_1.7-4
## [33] survMisc_1.0-3
## [34] scales_0.2.5
## [35] supraHex_1.7.2
## [36] BiocParallel_1.3.47
## [37] gtools_3.5.0
## [38] annotate_1.47.4
## [39] ggplot2_1.0.1
## [40] GenomicFeatures_1.21.13
## [41] bumphunter_1.27.0
## [42] proto_0.3-10
## [43] survival_2.38-3
## [44] magrittr_1.5
## [45] memoise_0.2.1
## [46] evaluate_0.7
```

```
## [47] GGally_0.5.0
## [48] R.methodsS3_1.7.0
## [49] plyr_1.8-4
## [50] KEGG.db_3.2.3
## [51] xml2_0.1.1
## [52] hwriter_1.3.2
## [53] graph_1.47.2
## [54] tools_3.2.1
## [55] data.table_1.9.6
## [56] formatR_1.2
## [57] matrixStats_0.14.2
## [58] stringi_1.0.0
## [59] xlsx_0.5.7
## [60] gunzoll_0.4.2
## [61] AnnotationDbi_1.31.17
## [62] lambda.r_1.1.7
## [63] rvmtstens_1.0.2
## [64] Biostrings_2.37.2
## [65] DESeq_1.21.0
## [66] futile.logger_1.4.1
## [67] XML_3.98-1.2
## [68] rjson_0.2.15
## [69] lgrgph_1.0.1
## [70] bitops_1.0-5
## [71] rmarkdown_0.7.1
## [72] dnsh_1.0.7
## [73] gtable_0.1.2
## [74] DBI_0.3.1
## [75] reshape_0.8.5
## [76] roxygen2_4.1.1
## [77] curl_0.8.1
## [78] RG_2.1.0
## [79] reshape2_1.4.1
## [80] EDASeq_2.3.2
## [81] GenomicAlignments_1.5.12
## [82] knitr_1.10.5
## [83] rtracklayer_1.29.13
## [84] futile.options_1.0.0
## [85] Rgraphviz_2.15.0
## [86] ape_3.3
## [87] rJava_0.9-7
## [88] TxDb.Hsapiens.UCSC.hg19.knownGene_3.1.3
## [89] xtable_1.2-21
## [90] stringi_0.5-5
## [91] Rcpp_0.12.0
## [92] genoplotter_1.47.0
```

References

Bullard, James H and Purdom, Elizabeth and Hansen, Kasper D and Dudoit, Sandrine. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in MRNA-Seq Experiments."

Huber, Wolfgang and Carey, Vincent J and Gentleman, Robert and Anders, Simon and Carlson, Marc and Carvalho,

Benilton S and Bravo, Hector Corrada and Davis, Sean and Gatto, Laurent and Girke, Thomas and others. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor."

Risso, Davide and Schwartz, Katja and Sherlock, Gavin and Dudoit, Sandrine. 2011. "GC-Content Normalization for RNA-Seq Data."