Análisis de datos. Parte 1

Datos clínicos y Datos de Expresión génica

Carmen Lebrero Cia

19/10/2020

Contents

Archivos de datos clínicos

Si utilizamos la función `str()` sobre `ClinicKIRC2` (datos clínicos de TCGA-KIRC) observamos que se trata de un archivo Dataframe con 537 observaciones y 66 variables.

```
str(ClinicKIRC2)
```

```
## 'data.frame':    537 obs. of  66 variables:
##  $ bcr_patient_barcode                       : chr  "TCGA-3Z-A93Z" "TCGA-6D-AA2E" "TCGA-A3-3306" "TC
##  $ additional_studies                        : Factor w/ 2 levels "","TCGAFPPP": 1 1 1 1 1 1 1 1 1 1
##  $ tumor_tissue_site                         : Factor w/ 1 level "Kidney": 1 1 1 1 1 1 1 1 1 1 ...
##  $ histological_type                         : Factor w/ 1 level "Kidney Clear Cell Renal Carcinoma"
##  $ other_dx                                  : Factor w/ 4 levels "No","Yes","Yes, History of Prior M
##  $ gender                                    : Factor w/ 2 levels "FEMALE","MALE": 2 1 2 2 1 2 2 2 2 2
##  $ vital_status                              : Factor w/ 3 levels "","Alive","Dead": 2 2 2 2 2 3 3 2
##  $ days_to_birth                             : int  -25205 -25043 -24569 -24315 -28287 -21183 -21556
##  $ days_to_last_known_alive                  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ days_to_death                             : int  NA NA NA NA NA 1191 735 NA NA NA ...
##  $ days_to_last_followup                     : int  4 135 1120 1436 16 NA NA 1493 1491 1130 ...
##  $ race_list                                 : Factor w/ 4 levels "","ASIAN","BLACK OR AFRICAN AMERIC
```

```
##  $ tissue_source_site                          : Factor w/ 20 levels "3Z","6D","A3",..: 1 2 3 3 3 3 3 3
##  $ patient_id                                  : chr  "A93Z" "AA2E" "3306" "3307" ...
##  $ bcr_patient_uuid                            : chr  "2B1DEA0A-6D55-4FDD-9C1C-0D9FBE03BD78" "D3B47E53-
##  $ history_of_neoadjuvant_treatment            : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
##  $ informed_consent_verified                   : Factor w/ 1 level "YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ icd_o_3_site                                : Factor w/ 1 level "C64.9": 1 1 1 1 1 1 1 1 1 1 ...
##  $ icd_o_3_histology                           : Factor w/ 2 levels "8310/3","8312/3": 1 1 1 1 1 1 1 1
##  $ icd_10                                      : Factor w/ 1 level "C64.9": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tissue_prospective_collection_indicator     : Factor w/ 3 levels "","NO","YES": 3 3 2 2 2 2 2 2 2 2
##  $ tissue_retrospective_collection_indicator   : Factor w/ 3 levels "","NO","YES": 2 2 3 3 3 3 3 3 3 3
##  $ ethnicity                                   : Factor w/ 3 levels "","HISPANIC OR LATINO",..: 3 3 3 1
##  $ person_neoplasm_cancer_status               : Factor w/ 3 levels "","TUMOR FREE",..: 2 2 1 3 2 2 2 1
##  $ performance_status_scale_timing             : Factor w/ 3 levels "","Other","Preoperative": 2 1 1 1
##  $ days_to_initial_pathologic_diagnosis        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age_at_initial_pathologic_diagnosis         : int  69 68 67 66 77 57 59 57 67 70 ...
##  $ year_of_initial_pathologic_diagnosis        : int  2013 2013 2005 2005 2006 2005 2005 2005 2005 2006
##  $ day_of_form_completion                      : int  11 17 23 13 12 20 14 14 14 16 ...
##  $ month_of_form_completion                    : int  11 3 8 4 4 4 4 4 4 4 ...
##  $ year_of_form_completion                     : int  2014 2014 2010 2010 2010 2010 2010 2010 2010 2010
##  $ laterality                                  : Factor w/ 3 levels "Bilateral","Left",..: 3 3 2 3 3 3
##  $ lactate_dehydrogenase_result                : Factor w/ 3 levels "","Elevated",..: 3 1 1 1 1 1 1 1
##  $ serum_calcium_result                        : Factor w/ 4 levels "","Elevated",..: 4 1 1 1 4 1 1 1 1
##  $ hemoglobin_result                           : Factor w/ 4 levels "","Elevated",..: 4 1 1 1 4 3 3 1 1
##  $ platelet_qualitative_result                 : Factor w/ 4 levels "","Elevated",..: 4 1 1 1 4 1 1 1 1
##  $ white_cell_count_result                     : Factor w/ 4 levels "","Elevated",..: 4 1 1 1 4 1 1 1 1
##  $ erythrocyte_sedimentation_rate_result       : Factor w/ 3 levels "","Elevated",..: 1 1 1 1 1 1 1 1 1
##  $ lymph_node_examined_count                   : int  NA NA NA 4 NA NA NA NA NA NA ...
##  $ number_of_lymphnodes_positive               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ karnofsky_performance_score                 : int  100 NA NA NA NA NA NA NA NA NA ...
##  $ eastern_cancer_oncology_group               : int  0 1 NA NA NA NA NA NA NA NA ...
##  $ primary_lymph_node_presentation_assessment: Factor w/ 3 levels "","NO","YES": 2 2 2 3 2 2 1 2 2 2
##  $ neoplasm_histologic_grade                   : Factor w/ 6 levels "","G1","G2","G3",..: 3 3 4 4 3 3 4
##  $ tobacco_smoking_history                     : int  5 1 NA NA NA NA NA NA NA NA ...
##  $ year_of_tobacco_smoking_onset               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ stopped_smoking_year                        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ number_pack_years_smoked                    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ targeted_molecular_therapy                  : Factor w/ 3 levels "","NO","YES": 2 2 1 1 1 1 1 1 1 1
##  $ radiation_therapy                           : Factor w/ 2 levels "","NO": 2 2 1 1 1 1 1 1 1 1 ...
##  $ primary_therapy_outcome_success             : Factor w/ 4 levels "","Complete Remission/Response",.
##  $ has_new_tumor_events_information            : chr  "NO" "NO" "NO" "NO" ...
##  $ has_drugs_information                       : chr  "NO" "NO" "NO" "NO" ...
##  $ has_radiations_information                  : chr  "NO" "NO" "NO" "NO" ...
##  $ has_follow_ups_information                  : chr  "YES" "YES" "NO" "NO" ...
##  $ project                                     : Factor w/ 1 level "TCGA-KIRC": 1 1 1 1 1 1 1 1 1 1 ..
##  $ stage_event_system_version                  : Factor w/ 4 levels "","5th","6th",..: 4 4 1 1 1 1 1 1
##  $ stage_event_clinical_stage                  : logi  NA NA NA NA NA NA ...
##  $ stage_event_pathologic_stage                : Factor w/ 5 levels "","Stage I","Stage II",..: 2 2 2 4
##  $ stage_event_tnm_categories                  : Factor w/ 59 levels "M0T1aN0M0","M0T1aNX",..: 1 15 23
##  $ stage_event_psa                             : logi  NA NA NA NA NA NA ...
##  $ stage_event_gleason_grading                 : logi  NA NA NA NA NA NA ...
##  $ stage_event_ann_arbor                       : logi  NA NA NA NA NA NA ...
##  $ stage_event_serum_markers                   : logi  NA NA NA NA NA NA ...
##  $ stage_event_igcccg_stage                    : logi  NA NA NA NA NA NA ...
##  $ stage_event_masaoka_stage                   : logi  NA NA NA NA NA NA ...
```

De entre todas las variables las más interesantes parecen ser `bcr_patient_barcode`, `vital_status` o `stage_event_pathologic_stage`. Vamos a encontrar los índices de cada variable en el dataframe y podemos hacer un dataframe más pequeño.

```
grep("bcr_patient_barcode", colnames(ClinicKIRC2))
```

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

```
grep("vital_status", colnames(ClinicKIRC2))
```

```
## [1] 7
```

```
grep("stage_event_pathologic_stage", colnames(ClinicKIRC2))
```

```
## [1] 59
```

```
ClinicS <- ClinicKIRC2[,c(1,7,59)]
head(ClinicS)
```

```
##   bcr_patient_barcode vital_status stage_event_pathologic_stage
## 1        TCGA-3Z-A93Z        Alive                      Stage I
## 2        TCGA-6D-AA2E        Alive                      Stage I
## 3        TCGA-A3-3306        Alive                      Stage I
## 4        TCGA-A3-3307        Alive                     Stage III
## 5        TCGA-A3-3308        Alive                     Stage III
## 6        TCGA-A3-3311         Dead                      Stage I
```

Vamos a ver si existen missing values con `is.na()`.

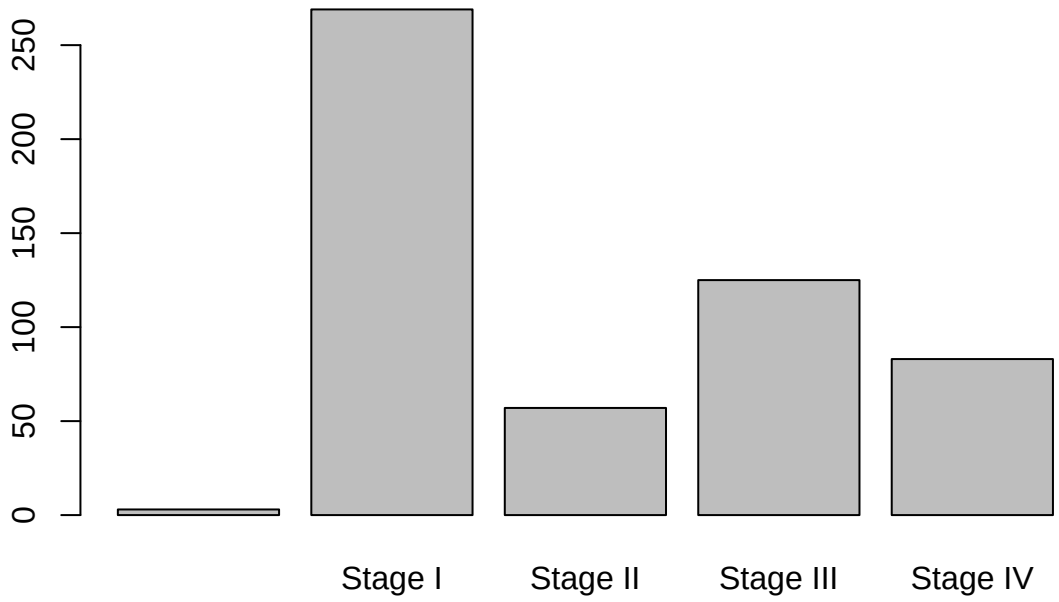
```
sum(is.na(ClinicS))
```

```
## [1] 0
```

Observamos que no tenemos ningún NA en este Dataset, por lo que tenemos datos de nuestra variable de interés para los 537 pacientes del estudio TCGA-KIRC.

Podemos sacar más información a partir de estas variables. Por ejemplo, podemos observar con una gráfica según la variable de estadío patológico que la mayoría de nuestras pertenecen al estadío I, seguido de las muestras en el estadío III, IV, II y 0.

```
plot(ClinicKIRC2$stage_event_pathologic_stage)
```



Podemos obtener el número exacto de pacientes con la función `summmary()`.

```
summary(ClinicKIRC2$stage_event_pathologic_stage)
```

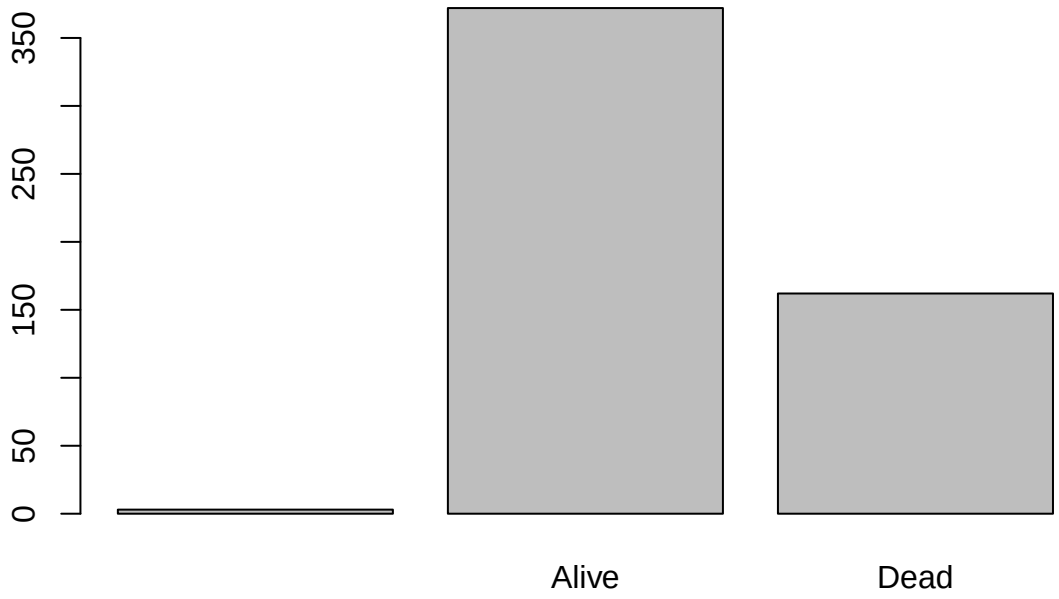
```
##             Stage I  Stage II Stage III  Stage IV
##         3       269        57       125        83
```

Esta clasificación por estadíos se refiere a lo siguiente a un método de agrupación de los pacientes según una serie de características de los tumores como la localización y el tamaño del tumor (T), si se observan ganglios linfáticos cerca (N) o si hay metástasis (M).

- **Estadío 0**. Describe cáncer "in situ" que están localizados en el lugar de origen y no se han esparcido a tejidos cercanos. Son tumores fácilmente curables que se pueden quitar con una cirugía.
- **Estadío 1**. Se trata de un tumor pequeño que no se ha extendido de forma muy profunda a tejidos colindantes ni a los ganglios linfáticos. A veces se le denomina cáncer temprano.
- **Estadíos II y III**. Estos dos estadíos indican tumores más grandes que se han extendido de forma más profunda a tejidos y que pueden haber llegado a los ganglios linfáticos pero no a otros órganos.
- **Estadío IV**. Este estadío significa que el cáncer se ha extendido a otros órganos. También conocido como cáncer metastásico avanzado.

También podemos obtener información acerca de los fallecimientos de nuestra muestra. Y observamos que 372 siguen vivos y 162 han fallecido. Volvemos a ver 3 muestras que no están catalogadas en ninguno de estos dos grupos, debemos ver si esas muestras, aunque hayamos hecho anteriormente la búsqueda de missing values, no contienen información.

```
plot(ClinicKIRC2$vital_status)
```

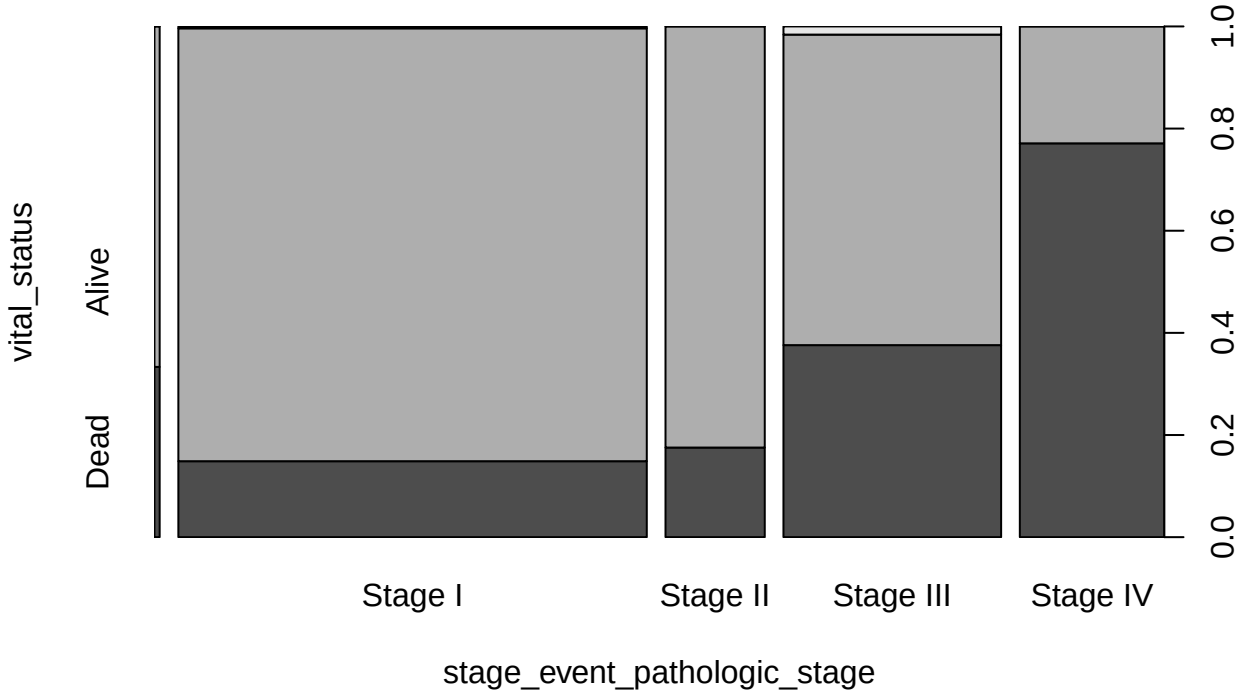



```
summary(ClinicKIRC2$vital_status)
```

```
##       Alive  Dead
##     3   372   162
```

Otro análisis interesante sería saber cuántos fallecidos hay según el estadío del tumor.

```
plot(vital_status ~ stage_event_pathologic_stage, data = ClinicKIRC2)
```



Fase I (n=269)

Fallecidos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Dead" & ClinicKIRC2$stage_event_patl
```

```
## [1] 40
```

Vivos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Alive" & ClinicKIRC2$stage_event_pa
```

```
## [1] 228
```

Fase II (n=57)

Fallecidos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Dead" & ClinicKIRC2$stage_event_patl
```

```
## [1] 10
```

Vivos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Alive" & ClinicKIRC2$stage_event_pa
```

```
## [1] 47
```

Fase III (n=125)

Fallecidos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Dead" & ClinicKIRC2$stage_event_pat
```

```
## [1] 47
```

Vivos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Alive" & ClinicKIRC2$stage_event_pa
```

```
## [1] 76
```

Fase IV (n=83)

Fallecidos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Dead" & ClinicKIRC2$stage_event_pat
```

```
## [1] 64
```

Vivos

```
length(subset(ClinicKIRC2$patient_id , ClinicKIRC2$vital_status == "Alive" & ClinicKIRC2$stage_event_pa
```

```
## [1] 19
```

Vamos a buscar esos pacientes que no tienen vital_status asignado:

```
subset(ClinicKIRC2$bcr_patient_barcode , ClinicKIRC2$vital_status != "Dead" & ClinicKIRC2$vital_status
```

```
## [1] "TCGA-BP-4326" "TCGA-BP-4329" "TCGA-BP-4334"
```

Y ver si estos tres son los mismos que no tienen asignada la fase del tumor:

```
subset(ClinicKIRC2$bcr_patient_barcode , ClinicKIRC2$stage_event_pathologic_stage != "Stage I" & Clinic
```

```
## [1] "TCGA-B4-5838" "TCGA-BP-4798" "TCGA-MM-A563"
```

No son los mismos pacientes los que no tienen etiqueta para vital_status y para la Fase del cancer. Creamos dos Datasets. `ClinicSVital` sin los tres pacientes que no tienen etiqueta para Vital_status y `ClinicSStage` sin los pacientes que no tienen etiqueta para Stage. Además se crearán dos cadenas de caracteres con los barcodes de cada Dataframe.

```
ClinicSVital <- subset(ClinicS, ClinicS$bcr_patient_barcode != "TCGA-BP-4326" & ClinicS$bcr_patient_bar
```

```
ClinicSStage <- subset(ClinicS, ClinicS$bcr_patient_barcode != "TCGA-B4-5838" & ClinicS$bcr_patient_bar
```

Si elegimos como variable a modelar `vital_status`se tratará de un problema de clasificaicón binaria. Mientras que si elegimos el estadío del tumor, sería un problema de clasificación multiclase.

Datos de Expresión Génica

Vamos a explorar nuestros datos de Expresión Génica `ExpGenKIRC3`. Observamos que se trata de un archivo `RangedSummarizedExperiment`, para tratar este tipo de archivos podemos seguir la siguiente guía: https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html.

Introducción

La clase `SummarizedExperiment` se usa para llenar matrices rectangulares de resultados experimentales producidos normalmente en experimentos de secuenciación o microarrays. Cada objeto almacena observaciones de una o más muestras, junto con metadatos adicionales que describen obsevaciones (características) y muestras (fenotipos).

Un aspecto clave de la clase `SummarizedExperiment` es la coordinación de los metadatos y los ensayos cuando se realizan subagrupaciones o subconjuntos. Por ejemplo, si quieres excluir una muestra se puede hacer en los metadatos y en los ensayos en una única operación, lo que asegura que los metadatos y los datos observados permanezcan sincronizados.

Anatomía de SummarizedExperiment

El paquete `SummarizedExperiment` contiene dos clases: `SummarizedExperiment` y `RangedSummarizedExperiment`.

`SummarizedExperiment` es un contenedor similar a una matriz donde las filas representan características de interés (por ejemplo: genes, transcritos, exones, etc.) y las columnas representan las muestras. Los objetos contienen uno o más ensayos, cada uno representado por un objeto matriz de número u otro modo. Las filas del objeto `SummarizedExperiment` representan características de interés. La información acerca de estas características está almacenada en un objeto Dataframe, accesible usando la función `rowData()`. Cada fila del Dataframe aporta información de la característica en la fila correspondiente del objeto `SummarizedExperiment`. Las columnas del Datafram representan diferentes atributos de las características de interés como IDs de genes o transcritos.

`RangedSummarizedExperiment` es el "hijo" de la clase `SummarizedExperiment`, lo que significaa que todos los métodos de `SummarizedExperiment`también funcionan sobre `RangedSummarizedExperiment`.

La diferencia fundamental entre las dos clases es que las filas de `RangedSummarizedExperiment`representan rangos genómicos de interés en vez de un Dataframe de carcterísticas. Los rangos de `RangedSummarizedExperiment`se describen en el objeto `GRanges`o `GRangesList`, accesible utilizando la función `rowRanges()`.

Assays

```
library(SummarizedExperiment)
se <- ExpGenKIRC3
se
```

```
## class: RangedSummarizedExperiment
## dim: 19947 606
## metadata(1): data_release
## assays(2): raw_count scaled_estimate
## rownames(19947): A1BG|1 A2M|2 ... TMED7-TICAM2|100302736
##   LOC100303728|100303728
## rowData names(4): gene_id entrezgene ensembl_gene_id
##   transcript_id.transcript_id_TCGA-B0-5694-01A-11R-1541-07
## colnames(606): TCGA-B0-5694-01A-11R-1541-07
##   TCGA-CJ-4637-01A-02R-1325-07 ... TCGA-CJ-4871-01A-01R-1305-07
##   TCGA-AK-3460-01A-02R-1277-07
## colData names(70): barcode patient ... paper_mRNA_cluster
##   paper_microRNA_cluster
```

Para recuperar los datos a partir del experimento a partir de un objeto `SummarizedExperiment` se puede utilizar `assays()`. Un objeto puede tener múltiples dataset de ensayos a los que se puede acceder usando el operador `$`. En nuestro caso tenemos dos datasets: `raw_count` y `scaled_estimate`. Partiremos desde el archivo `raw_count` para realizar nuestro análisis.

```
assays(se)$raw_count[1:3,1:4]
```

```
##        TCGA-B0-5694-01A-11R-1541-07 TCGA-CJ-4637-01A-02R-1325-07
## A1BG|1                        36.00                        87.66
## A2M|2                      71105.61                     47586.92
## NAT1|9                       111.00                       295.00
##        TCGA-CZ-4860-01A-01R-1305-07 TCGA-B0-4706-01A-01R-1503-07
## A1BG|1                        60.00                       180.09
## A2M|2                      39077.92                     69333.51
## NAT1|9                       208.00                       140.00
```

```
rowRanges(se)
```

```
## GRanges object with 19947 ranges and 4 metadata columns:
##                            seqnames              ranges strand |      gene_id
##                               <Rle>           <IRanges>  <Rle> |  <character>
##                   A1BG|1    chr19   58856544-58864865      - |          A1BG
##                    A2M|2    chr12     9220260-9268825      - |           A2M
##                   NAT1|9     chr8   18027986-18081198      + |          NAT1
##                  NAT2|10     chr8   18248755-18258728      + |          NAT2
##              SERPINA3|12    chr14   95058395-95090983      + | RP11-986E7.7
##                      ...      ...                 ...    ... .           ...
##   LOC100302401|100302401     chr1 178060643-178063119      - |   RASAL2-AS1
##   LOC100302640|100302640     chr3 106555658-106959488      - |    LINC00882
##     NCRNA00182|100302692     chrX   73183790-73513409      - |           FTX
##   TMED7-TICAM2|100302736     chr5 114914339-114961876      - |       TICAM2
##   LOC100303728|100303728     chrX 118599997-118603061      - |  SLC25A5-AS1
##                            entrezgene ensembl_gene_id
##                             <integer>     <character>
##                   A1BG|1          1 ENSG00000121410
##                    A2M|2          2 ENSG00000175899
##                   NAT1|9          9 ENSG00000171428
##                  NAT2|10         10 ENSG00000156006
```

```
##              SERPINA3|12         12 ENSG00000273259
##                      ...        ...             ...
##   LOC100302401|100302401  100302401 ENSG00000224687
##   LOC100302640|100302640  100302640 ENSG00000242759
##     NCRNA00182|100302692  100302692 ENSG00000230590
##   TMED7-TICAM2|100302736  100302736 ENSG00000243414
##   LOC100303728|100303728  100303728 ENSG00000224281
##                          transcript_id.transcript_id_TCGA-B0-5694-01A-11R-1541-07
##                                                                      <character>
##                   A1BG|1                                       uc002qsd.3,uc002qsf.1
##                    A2M|2                                      uc001qvj.1,uc001qvk...
##                   NAT1|9                                      uc003wyq.2,uc003wyr...
##                  NAT2|10                                                uc003wyw.1
##              SERPINA3|12                                      uc001ydo.3,uc001ydp...
##                      ...                                                        ...
##   LOC100302401|100302401                                       uc001gln.1,uc001glo.1
##   LOC100302640|100302640                                       uc003dwf.3,uc011bhk.1
##     NCRNA00182|100302692                                       uc004ebr.1,uc010nlq.1
##   TMED7-TICAM2|100302736                                       uc003krd.2,uc003kre.2
##   LOC100303728|100303728                                       uc004ere.1,uc004erg.1
##   -------
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

Columnas (datos de las muestras)

Se puede acceder a los metadatos que describen las muestras usando '`colData()`, y es un Dataframe que puede almacenar cualquier número de columna.

```
colData(se)
```

```
## DataFrame with 606 rows and 70 columns
##                                           barcode      patient
##                                       <character>  <character>
## TCGA-B0-5694-01A-11R-1541-07 TCGA-B0-5694-01A-11R.. TCGA-B0-5694
## TCGA-CJ-4637-01A-02R-1325-07 TCGA-CJ-4637-01A-02R.. TCGA-CJ-4637
## TCGA-CZ-4860-01A-01R-1305-07 TCGA-CZ-4860-01A-01R.. TCGA-CZ-4860
## TCGA-B0-4706-01A-01R-1503-07 TCGA-B0-4706-01A-01R.. TCGA-B0-4706
## TCGA-B4-5844-01A-11R-1672-07 TCGA-B4-5844-01A-11R.. TCGA-B4-5844
## ...                                             ...          ...
## TCGA-B4-5836-01A-11R-1672-07 TCGA-B4-5836-01A-11R.. TCGA-B4-5836
## TCGA-A3-3308-01A-02R-1325-07 TCGA-A3-3308-01A-02R.. TCGA-A3-3308
## TCGA-AK-3440-01A-02R-1277-07 TCGA-AK-3440-01A-02R.. TCGA-AK-3440
## TCGA-CJ-4871-01A-01R-1305-07 TCGA-CJ-4871-01A-01R.. TCGA-CJ-4871
## TCGA-AK-3460-01A-02R-1277-07 TCGA-AK-3460-01A-02R.. TCGA-AK-3460
##                                         sample shortLetterCode
##                                     <character>     <character>
## TCGA-B0-5694-01A-11R-1541-07 TCGA-B0-5694-01A              TP
## TCGA-CJ-4637-01A-02R-1325-07 TCGA-CJ-4637-01A              TP
## TCGA-CZ-4860-01A-01R-1305-07 TCGA-CZ-4860-01A              TP
## TCGA-B0-4706-01A-01R-1503-07 TCGA-B0-4706-01A              TP
## TCGA-B4-5844-01A-11R-1672-07 TCGA-B4-5844-01A              TP
## ...                                       ...             ...
## TCGA-B4-5836-01A-11R-1672-07 TCGA-B4-5836-01A              TP
```

```
## TCGA-A3-3308-01A-02R-1325-07 TCGA-A3-3308-01A              TP
## TCGA-AK-3440-01A-02R-1277-07 TCGA-AK-3440-01A              TP
## TCGA-CJ-4871-01A-01R-1305-07 TCGA-CJ-4871-01A              TP
## TCGA-AK-3460-01A-02R-1277-07 TCGA-AK-3460-01A              TP
##                                      definition sample_submitter_id   is_ffpe
##                                     <character>         <character> <logical>
## TCGA-B0-5694-01A-11R-1541-07 Primary solid Tumor    TCGA-B0-5694-01A     FALSE
## TCGA-CJ-4637-01A-02R-1325-07 Primary solid Tumor    TCGA-CJ-4637-01A     FALSE
## TCGA-CZ-4860-01A-01R-1305-07 Primary solid Tumor    TCGA-CZ-4860-01A     FALSE
## TCGA-B0-4706-01A-01R-1503-07 Primary solid Tumor    TCGA-B0-4706-01A     FALSE
## TCGA-B4-5844-01A-11R-1672-07 Primary solid Tumor    TCGA-B4-5844-01A     FALSE
## ...                                           ...             ...       ...
## TCGA-B4-5836-01A-11R-1672-07 Primary solid Tumor    TCGA-B4-5836-01A     FALSE
## TCGA-A3-3308-01A-02R-1325-07 Primary solid Tumor    TCGA-A3-3308-01A     FALSE
## TCGA-AK-3440-01A-02R-1277-07 Primary solid Tumor    TCGA-AK-3440-01A     FALSE
## TCGA-CJ-4871-01A-01R-1305-07 Primary solid Tumor    TCGA-CJ-4871-01A     FALSE
## TCGA-AK-3460-01A-02R-1277-07 Primary solid Tumor    TCGA-AK-3460-01A     FALSE
##                                  tissue_type initial_weight   sample_type
##                                  <character>      <numeric>   <character>
## TCGA-B0-5694-01A-11R-1541-07 Not Reported             NA Primary Tumor
## TCGA-CJ-4637-01A-02R-1325-07 Not Reported             NA Primary Tumor
## TCGA-CZ-4860-01A-01R-1305-07 Not Reported             NA Primary Tumor
## TCGA-B0-4706-01A-01R-1503-07 Not Reported             NA Primary Tumor
## TCGA-B4-5844-01A-11R-1672-07 Not Reported             NA Primary Tumor
## ...                                   ...            ...           ...
## TCGA-B4-5836-01A-11R-1672-07 Not Reported             NA Primary Tumor
## TCGA-A3-3308-01A-02R-1325-07 Not Reported             NA Primary Tumor
## TCGA-AK-3440-01A-02R-1277-07 Not Reported             NA Primary Tumor
## TCGA-CJ-4871-01A-01R-1305-07 Not Reported             NA Primary Tumor
## TCGA-AK-3460-01A-02R-1277-07 Not Reported             NA Primary Tumor
##                                 shortest_dimension oct_embedded submitter_id
##                                          <numeric>  <character>  <character>
## TCGA-B0-5694-01A-11R-1541-07                   0.3           NA TCGA-B0-5694
## TCGA-CJ-4637-01A-02R-1325-07                   0.6           NA TCGA-CJ-4637
## TCGA-CZ-4860-01A-01R-1305-07                   0.2           NA TCGA-CZ-4860
## TCGA-B0-4706-01A-01R-1503-07                   0.3           NA TCGA-B0-4706
## TCGA-B4-5844-01A-11R-1672-07                   0.3           NA TCGA-B4-5844
## ...                                            ...          ...          ...
## TCGA-B4-5836-01A-11R-1672-07                   0.4           NA TCGA-B4-5836
## TCGA-A3-3308-01A-02R-1325-07                   0.4           NA TCGA-A3-3308
## TCGA-AK-3440-01A-02R-1277-07                   0.3           NA TCGA-AK-3440
## TCGA-CJ-4871-01A-01R-1305-07                   0.9           NA TCGA-CJ-4871
## TCGA-AK-3460-01A-02R-1277-07                   0.4           NA TCGA-AK-3460
##                                 longest_dimension             sample_id
##                                         <numeric>           <character>
## TCGA-B0-5694-01A-11R-1541-07                 2.5 dce44e3e-0bed-43e1-b..
## TCGA-CJ-4637-01A-02R-1325-07                 2.0 c395332c-5886-44dd-a..
## TCGA-CZ-4860-01A-01R-1305-07                 1.3 34223e9d-5e0a-46ce-a..
## TCGA-B0-4706-01A-01R-1503-07                 3.0 b48f69f1-9023-4bab-9..
## TCGA-B4-5844-01A-11R-1672-07                 1.1 82ae96b4-29c9-45cc-9..
## ...                                          ...                   ...
## TCGA-B4-5836-01A-11R-1672-07                 1.7 e6979738-59b0-4965-8..
## TCGA-A3-3308-01A-02R-1325-07                 1.2 0b935adb-0e19-4fcc-9..
## TCGA-AK-3440-01A-02R-1277-07                 1.2 a1f24fc5-19a7-474c-9..
```

```
## TCGA-CJ-4871-01A-01R-1305-07              3.3 ef710e85-bc6d-42bc-a..
## TCGA-AK-3460-01A-02R-1277-07              1.0 fb676be0-8122-4775-a..
##                              intermediate_dimension days_to_collection
##                                           <numeric>          <integer>
## TCGA-B0-5694-01A-11R-1541-07                   1.5                 NA
## TCGA-CJ-4637-01A-02R-1325-07                   0.8                 NA
## TCGA-CZ-4860-01A-01R-1305-07                   1.0                 NA
## TCGA-B0-4706-01A-01R-1503-07                   1.8                 NA
## TCGA-B4-5844-01A-11R-1672-07                   0.6                 NA
## ...                                            ...                ...
## TCGA-B4-5836-01A-11R-1672-07                   0.7                 NA
## TCGA-A3-3308-01A-02R-1325-07                   1.0                 NA
## TCGA-AK-3440-01A-02R-1277-07                   0.9                 NA
## TCGA-CJ-4871-01A-01R-1305-07                   0.9                 NA
## TCGA-AK-3460-01A-02R-1277-07                   0.6                 NA
##                               pathology_report_uuid       state sample_type_id
##                                         <character> <character>    <character>
## TCGA-B0-5694-01A-11R-1541-07 c28c9873-1da9-4647-8..    released             01
## TCGA-CJ-4637-01A-02R-1325-07 69449fd5-f715-4387-b..    released             01
## TCGA-CZ-4860-01A-01R-1305-07 a19c8a6e-471a-4bb8-8..    released             01
## TCGA-B0-4706-01A-01R-1503-07 9e87fe25-1a78-446c-a..    released             01
## TCGA-B4-5844-01A-11R-1672-07 4d21d547-3865-4e10-a..    released             01
## ...                                              ...         ...            ...
## TCGA-B4-5836-01A-11R-1672-07 f39308e8-58c1-4aa8-b..    released             01
## TCGA-A3-3308-01A-02R-1325-07 de8eb09e-bec5-4967-8..    released             01
## TCGA-AK-3440-01A-02R-1277-07 57353f14-2db0-4dae-8..    released             01
## TCGA-CJ-4871-01A-01R-1305-07 e011c8a3-ae9b-41dc-8..    released             01
## TCGA-AK-3460-01A-02R-1277-07 95fc84f5-4fb3-4ce7-a..    released             01
##                              site_of_resection_or_biopsy prior_treatment
##                                              <character>     <character>
## TCGA-B0-5694-01A-11R-1541-07                Kidney, NOS              No
## TCGA-CJ-4637-01A-02R-1325-07                Kidney, NOS              No
## TCGA-CZ-4860-01A-01R-1305-07                Kidney, NOS              No
## TCGA-B0-4706-01A-01R-1503-07                Kidney, NOS              No
## TCGA-B4-5844-01A-11R-1672-07                Kidney, NOS              No
## ...                                                 ...             ...
## TCGA-B4-5836-01A-11R-1672-07                Kidney, NOS              No
## TCGA-A3-3308-01A-02R-1325-07                Kidney, NOS              No
## TCGA-AK-3440-01A-02R-1277-07                Kidney, NOS              No
## TCGA-CJ-4871-01A-01R-1305-07                Kidney, NOS              No
## TCGA-AK-3460-01A-02R-1277-07                Kidney, NOS              No
##                              age_at_diagnosis tissue_or_organ_of_origin
##                                     <integer>               <character>
## TCGA-B0-5694-01A-11R-1541-07            26060               Kidney, NOS
## TCGA-CJ-4637-01A-02R-1325-07            19101               Kidney, NOS
## TCGA-CZ-4860-01A-01R-1305-07            22115               Kidney, NOS
## TCGA-B0-4706-01A-01R-1503-07            22331               Kidney, NOS
## TCGA-B4-5844-01A-11R-1672-07            22409               Kidney, NOS
## ...                                       ...                     ...
## TCGA-B4-5836-01A-11R-1672-07            22341               Kidney, NOS
## TCGA-A3-3308-01A-02R-1325-07            28287               Kidney, NOS
## TCGA-AK-3440-01A-02R-1277-07            21337               Kidney, NOS
## TCGA-CJ-4871-01A-01R-1305-07            23195               Kidney, NOS
## TCGA-AK-3460-01A-02R-1277-07            21281               Kidney, NOS
```

```
##                              tumor_stage ajcc_pathologic_stage
##                              <character>           <character>
## TCGA-B0-5694-01A-11R-1541-07   stage iii             Stage III
## TCGA-CJ-4637-01A-02R-1325-07    stage iv              Stage IV
## TCGA-CZ-4860-01A-01R-1305-07    stage iv              Stage IV
## TCGA-B0-4706-01A-01R-1503-07   stage iii             Stage III
## TCGA-B4-5844-01A-11R-1672-07    stage ii              Stage II
## ...                                 ...                   ...
## TCGA-B4-5836-01A-11R-1672-07    stage i               Stage I
## TCGA-A3-3308-01A-02R-1325-07   stage iii             Stage III
## TCGA-AK-3440-01A-02R-1277-07    stage i               Stage I
## TCGA-CJ-4871-01A-01R-1305-07    stage iv              Stage IV
## TCGA-AK-3460-01A-02R-1277-07    stage i               Stage I
##                                      diagnosis_id icd_10_code
##                                       <character> <character>
## TCGA-B0-5694-01A-11R-1541-07 9ddd0dac-eba5-5649-a..       C64.9
## TCGA-CJ-4637-01A-02R-1325-07 4c3e5037-7899-50b6-b..       C64.9
## TCGA-CZ-4860-01A-01R-1305-07 ce3952ae-b70e-57f4-9..       C64.9
## TCGA-B0-4706-01A-01R-1503-07 7bf96daf-b83a-509f-8..       C64.9
## TCGA-B4-5844-01A-11R-1672-07 747fc6cc-1aa5-5972-a..       C64.9
## ...                                               ...         ...
## TCGA-B4-5836-01A-11R-1672-07 e12798de-68d1-5549-9..       C64.9
## TCGA-A3-3308-01A-02R-1325-07 be58a956-e0f9-5dbb-9..       C64.9
## TCGA-AK-3440-01A-02R-1277-07 8688cd3f-804b-5d65-b..       C64.9
## TCGA-CJ-4871-01A-01R-1305-07 17c6dd92-3f34-573e-9..       C64.9
## TCGA-AK-3460-01A-02R-1277-07 56761b0d-ed62-50ed-9..       C64.9
##                              year_of_diagnosis ajcc_pathologic_m
##                                      <integer>       <character>
## TCGA-B0-5694-01A-11R-1541-07              2008                M0
## TCGA-CJ-4637-01A-02R-1325-07              2004                M1
## TCGA-CZ-4860-01A-01R-1305-07              2005                M1
## TCGA-B0-4706-01A-01R-1503-07              2007                M0
## TCGA-B4-5844-01A-11R-1672-07              2010                M0
## ...                                        ...               ...
## TCGA-B4-5836-01A-11R-1672-07              2010                M0
## TCGA-A3-3308-01A-02R-1325-07              2006                M0
## TCGA-AK-3440-01A-02R-1277-07              2005                M0
## TCGA-CJ-4871-01A-01R-1305-07              2004                M1
## TCGA-AK-3460-01A-02R-1277-07              2007                M0
##                              ajcc_pathologic_t  morphology  tumor_grade
##                                    <character> <character>  <character>
## TCGA-B0-5694-01A-11R-1541-07               T3a      8310/3 not reported
## TCGA-CJ-4637-01A-02R-1325-07               T2b      8310/3 not reported
## TCGA-CZ-4860-01A-01R-1305-07                T4      8310/3 not reported
## TCGA-B0-4706-01A-01R-1503-07               T3a      8310/3 not reported
## TCGA-B4-5844-01A-11R-1672-07                T2      8310/3 not reported
## ...                                        ...         ...          ...
## TCGA-B4-5836-01A-11R-1672-07               T1b      8310/3 not reported
## TCGA-A3-3308-01A-02R-1325-07               T3b      8310/3 not reported
## TCGA-AK-3440-01A-02R-1277-07               T1a      8310/3 not reported
## TCGA-CJ-4871-01A-01R-1305-07               T3a      8310/3 not reported
## TCGA-AK-3460-01A-02R-1277-07               T1a      8310/3 not reported
##                              prior_malignancy
##                                   <character>
```

```
## TCGA-B0-5694-01A-11R-1541-07              yes
## TCGA-CJ-4637-01A-02R-1325-07               no
## TCGA-CZ-4860-01A-01R-1305-07               no
## TCGA-B0-4706-01A-01R-1503-07               no
## TCGA-B4-5844-01A-11R-1672-07               no
## ...                                       ...
## TCGA-B4-5836-01A-11R-1672-07               no
## TCGA-A3-3308-01A-02R-1325-07               no
## TCGA-AK-3440-01A-02R-1277-07               no
## TCGA-CJ-4871-01A-01R-1305-07               no
## TCGA-AK-3460-01A-02R-1277-07               no
##
##
## TCGA-B0-5694-01A-11R-1541-07
## TCGA-CJ-4637-01A-02R-1325-07
## TCGA-CZ-4860-01A-01R-1305-07
## TCGA-B0-4706-01A-01R-1503-07
## TCGA-B4-5844-01A-11R-1672-07
## ...
## TCGA-B4-5836-01A-11R-1672-07 Radiation Therapy, NOS:released:TCGA-B4-5836_treatment:...,Pharmaceutica
## TCGA-A3-3308-01A-02R-1325-07
## TCGA-AK-3440-01A-02R-1277-07 Radiation Therapy, NOS:released:TCGA-AK-3440_treatment:...,Pharmaceutica
## TCGA-CJ-4871-01A-01R-1305-07
## TCGA-AK-3460-01A-02R-1277-07 Radiation Therapy, NOS:released:TCGA-AK-3460_treatment:...,Pharmaceutica
##                              days_to_last_follow_up      primary_diagnosis
##                                           <integer>             <character>
## TCGA-B0-5694-01A-11R-1541-07                    NA Clear cell adenocarc..
## TCGA-CJ-4637-01A-02R-1325-07                    NA Clear cell adenocarc..
## TCGA-CZ-4860-01A-01R-1305-07                    NA Clear cell adenocarc..
## TCGA-B0-4706-01A-01R-1503-07                    26 Clear cell adenocarc..
## TCGA-B4-5844-01A-11R-1672-07                     7 Clear cell adenocarc..
## ...                                            ...                    ...
## TCGA-B4-5836-01A-11R-1672-07                   141 Clear cell adenocarc..
## TCGA-A3-3308-01A-02R-1325-07                    16 Clear cell adenocarc..
## TCGA-AK-3440-01A-02R-1277-07                  2865 Clear cell adenocarc..
## TCGA-CJ-4871-01A-01R-1305-07                  2423 Clear cell adenocarc..
## TCGA-AK-3460-01A-02R-1277-07                  2508 Clear cell adenocarc..
##                              synchronous_malignancy classification_of_tumor
##                                         <character>              <character>
## TCGA-B0-5694-01A-11R-1541-07          Not Reported             not reported
## TCGA-CJ-4637-01A-02R-1325-07                    No             not reported
## TCGA-CZ-4860-01A-01R-1305-07                    No             not reported
## TCGA-B0-4706-01A-01R-1503-07                    No             not reported
## TCGA-B4-5844-01A-11R-1672-07                    No             not reported
## ...                                            ...                      ...
## TCGA-B4-5836-01A-11R-1672-07                    No             not reported
## TCGA-A3-3308-01A-02R-1325-07                    No             not reported
## TCGA-AK-3440-01A-02R-1277-07                    No             not reported
## TCGA-CJ-4871-01A-01R-1305-07                    No             not reported
## TCGA-AK-3460-01A-02R-1277-07                    No             not reported
##                              days_to_diagnosis ajcc_pathologic_n
##                                      <integer>       <character>
## TCGA-B0-5694-01A-11R-1541-07                 0                N0
## TCGA-CJ-4637-01A-02R-1325-07                 0                NX
```

```
## TCGA-CZ-4860-01A-01R-1305-07                0               NX
## TCGA-B0-4706-01A-01R-1503-07                0               NX
## TCGA-B4-5844-01A-11R-1672-07                0               NO
## ...                                       ...              ...
## TCGA-B4-5836-01A-11R-1672-07                0               NO
## TCGA-A3-3308-01A-02R-1325-07                0               NO
## TCGA-AK-3440-01A-02R-1277-07                0               NX
## TCGA-CJ-4871-01A-01R-1305-07                0               NX
## TCGA-AK-3460-01A-02R-1277-07                0               NX
##                              last_known_disease_status
##                                            <character>
## TCGA-B0-5694-01A-11R-1541-07              not reported
## TCGA-CJ-4637-01A-02R-1325-07              not reported
## TCGA-CZ-4860-01A-01R-1305-07              not reported
## TCGA-B0-4706-01A-01R-1503-07              not reported
## TCGA-B4-5844-01A-11R-1672-07              not reported
## ...                                                ...
## TCGA-B4-5836-01A-11R-1672-07              not reported
## TCGA-A3-3308-01A-02R-1325-07              not reported
## TCGA-AK-3440-01A-02R-1277-07              not reported
## TCGA-CJ-4871-01A-01R-1305-07              not reported
## TCGA-AK-3460-01A-02R-1277-07              not reported
##                              progression_or_recurrence
##                                            <character>
## TCGA-B0-5694-01A-11R-1541-07              not reported
## TCGA-CJ-4637-01A-02R-1325-07              not reported
## TCGA-CZ-4860-01A-01R-1305-07              not reported
## TCGA-B0-4706-01A-01R-1503-07              not reported
## TCGA-B4-5844-01A-11R-1672-07              not reported
## ...                                                ...
## TCGA-B4-5836-01A-11R-1672-07              not reported
## TCGA-A3-3308-01A-02R-1325-07              not reported
## TCGA-AK-3440-01A-02R-1277-07              not reported
## TCGA-CJ-4871-01A-01R-1305-07              not reported
## TCGA-AK-3460-01A-02R-1277-07              not reported
##                              ajcc_staging_system_edition ajcc_clinical_m
##                                              <character>     <character>
## TCGA-B0-5694-01A-11R-1541-07                          NA              NA
## TCGA-CJ-4637-01A-02R-1325-07                          NA              NA
## TCGA-CZ-4860-01A-01R-1305-07                          NA              NA
## TCGA-B0-4706-01A-01R-1503-07                          NA              NA
## TCGA-B4-5844-01A-11R-1672-07                         7th              NA
## ...                                                  ...             ...
## TCGA-B4-5836-01A-11R-1672-07                         7th              NA
## TCGA-A3-3308-01A-02R-1325-07                          NA              NA
## TCGA-AK-3440-01A-02R-1277-07                         6th              NA
## TCGA-CJ-4871-01A-01R-1305-07                          NA              NA
## TCGA-AK-3460-01A-02R-1277-07                         6th              NA
##                              alcohol_history cigarettes_per_day
##                                  <character>          <numeric>
## TCGA-B0-5694-01A-11R-1541-07    Not Reported                 NA
## TCGA-CJ-4637-01A-02R-1325-07    Not Reported                 NA
## TCGA-CZ-4860-01A-01R-1305-07    Not Reported                 NA
## TCGA-B0-4706-01A-01R-1503-07    Not Reported                 NA
```

```
## TCGA-B4-5844-01A-11R-1672-07    Not Reported                NA
## ...                                      ...               ...
## TCGA-B4-5836-01A-11R-1672-07    Not Reported                NA
## TCGA-A3-3308-01A-02R-1325-07    Not Reported                NA
## TCGA-AK-3440-01A-02R-1277-07    Not Reported                NA
## TCGA-CJ-4871-01A-01R-1305-07    Not Reported                NA
## TCGA-AK-3460-01A-02R-1277-07    Not Reported                NA
##                                         exposure_id pack_years_smoked
##                                         <character>         <numeric>
## TCGA-B0-5694-01A-11R-1541-07 d86ca79c-954e-5c8a-8..                  NA
## TCGA-CJ-4637-01A-02R-1325-07 c8a989f5-83c9-5559-a..                  NA
## TCGA-CZ-4860-01A-01R-1305-07 f59f1a3e-b4f2-5063-8..                  NA
## TCGA-B0-4706-01A-01R-1503-07 f31ce773-a579-5a93-a..                  NA
## TCGA-B4-5844-01A-11R-1672-07 f22a82f2-2935-5b9b-b..                  NA
## ...                                            ...                 ...
## TCGA-B4-5836-01A-11R-1672-07 33db08a1-fce2-574a-b..                  NA
## TCGA-A3-3308-01A-02R-1325-07 c4461280-2070-5047-9..                  NA
## TCGA-AK-3440-01A-02R-1277-07 f2467123-d905-58d0-b..                  NA
## TCGA-CJ-4871-01A-01R-1305-07 44522f0b-f7d2-5740-b..                  NA
## TCGA-AK-3460-01A-02R-1277-07 0c03b000-82f4-5c27-a..                  NA
##                              vital_status age_at_index year_of_death
##                               <character>    <integer>     <integer>
## TCGA-B0-5694-01A-11R-1541-07         Dead           71          2009
## TCGA-CJ-4637-01A-02R-1325-07         Dead           52          2010
## TCGA-CZ-4860-01A-01R-1305-07         Dead           60          2005
## TCGA-B0-4706-01A-01R-1503-07         Dead           61          2007
## TCGA-B4-5844-01A-11R-1672-07        Alive           61            NA
## ...                                   ...          ...           ...
## TCGA-B4-5836-01A-11R-1672-07        Alive           61            NA
## TCGA-A3-3308-01A-02R-1325-07        Alive           77            NA
## TCGA-AK-3440-01A-02R-1277-07        Alive           58            NA
## TCGA-CJ-4871-01A-01R-1305-07        Alive           63            NA
## TCGA-AK-3460-01A-02R-1277-07        Alive           58            NA
##                                     race days_to_birth year_of_birth
##                              <character>     <integer>     <integer>
## TCGA-B0-5694-01A-11R-1541-07       white        -26060          1937
## TCGA-CJ-4637-01A-02R-1325-07       white        -19101          1952
## TCGA-CZ-4860-01A-01R-1305-07       white        -22115          1945
## TCGA-B0-4706-01A-01R-1503-07       white        -22331          1946
## TCGA-B4-5844-01A-11R-1672-07       white        -22409          1949
## ...                                  ...           ...           ...
## TCGA-B4-5836-01A-11R-1672-07       white        -22341          1949
## TCGA-A3-3308-01A-02R-1325-07       white        -28287          1929
## TCGA-AK-3440-01A-02R-1277-07       white        -21337          1947
## TCGA-CJ-4871-01A-01R-1305-07       white        -23195          1941
## TCGA-AK-3460-01A-02R-1277-07       white        -21281          1949
##                                       demographic_id      gender
##                                          <character> <character>
## TCGA-B0-5694-01A-11R-1541-07 b8114567-42c1-58d0-a..        male
## TCGA-CJ-4637-01A-02R-1325-07 d9f124e5-35df-5c43-8..      female
## TCGA-CZ-4860-01A-01R-1305-07 b46b362f-877a-577c-9..        male
## TCGA-B0-4706-01A-01R-1503-07 9ff2bd3e-0ad9-5c91-b..        male
## TCGA-B4-5844-01A-11R-1672-07 0865e6de-4b9f-5d10-a..      female
## ...                                             ...         ...
```

```
## TCGA-B4-5836-01A-11R-1672-07 96b5779b-6e2c-5946-8..      female
## TCGA-A3-3308-01A-02R-1325-07 c8eac0e5-d579-591d-8..      female
## TCGA-AK-3440-01A-02R-1277-07 52ed0983-006f-5eb9-8..        male
## TCGA-CJ-4871-01A-01R-1305-07 c5bb4d54-5965-587b-8..        male
## TCGA-AK-3460-01A-02R-1277-07 6dc802fc-4785-59b0-b..        male
##                                          ethnicity days_to_death
##                                        <character>     <integer>
## TCGA-B0-5694-01A-11R-1541-07 not hispanic or latino           480
## TCGA-CJ-4637-01A-02R-1325-07     hispanic or latino          2227
## TCGA-CZ-4860-01A-01R-1305-07           not reported           206
## TCGA-B0-4706-01A-01R-1503-07 not hispanic or latino            65
## TCGA-B4-5844-01A-11R-1672-07 not hispanic or latino            NA
## ...                                               ...           ...
## TCGA-B4-5836-01A-11R-1672-07 not hispanic or latino            NA
## TCGA-A3-3308-01A-02R-1325-07           not reported            NA
## TCGA-AK-3440-01A-02R-1277-07 not hispanic or latino            NA
## TCGA-CJ-4871-01A-01R-1305-07 not hispanic or latino            NA
## TCGA-AK-3460-01A-02R-1277-07 not hispanic or latino            NA
##                                bcr_patient_barcode  project_id
##                                        <character> <character>
## TCGA-B0-5694-01A-11R-1541-07    TCGA-B0-5694-01A   TCGA-KIRC
## TCGA-CJ-4637-01A-02R-1325-07    TCGA-CJ-4637-01A   TCGA-KIRC
## TCGA-CZ-4860-01A-01R-1305-07    TCGA-CZ-4860-01A   TCGA-KIRC
## TCGA-B0-4706-01A-01R-1503-07    TCGA-B0-4706-01A   TCGA-KIRC
## TCGA-B4-5844-01A-11R-1672-07    TCGA-B4-5844-01A   TCGA-KIRC
## ...                                          ...         ...
## TCGA-B4-5836-01A-11R-1672-07    TCGA-B4-5836-01A   TCGA-KIRC
## TCGA-A3-3308-01A-02R-1325-07    TCGA-A3-3308-01A   TCGA-KIRC
## TCGA-AK-3440-01A-02R-1277-07    TCGA-AK-3440-01A   TCGA-KIRC
## TCGA-CJ-4871-01A-01R-1305-07    TCGA-CJ-4871-01A   TCGA-KIRC
## TCGA-AK-3460-01A-02R-1277-07    TCGA-AK-3460-01A   TCGA-KIRC
##                                       disease_type releasable  released
##                                             <list>  <logical> <logical>
## TCGA-B0-5694-01A-11R-1541-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-CJ-4637-01A-02R-1325-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-CZ-4860-01A-01R-1305-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-B0-4706-01A-01R-1503-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-B4-5844-01A-11R-1672-07 Adenomas and Adenoca..       TRUE      TRUE
## ...                                             ...        ...       ...
## TCGA-B4-5836-01A-11R-1672-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-A3-3308-01A-02R-1325-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-AK-3440-01A-02R-1277-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-CJ-4871-01A-01R-1305-07 Adenomas and Adenoca..       TRUE      TRUE
## TCGA-AK-3460-01A-02R-1277-07 Adenomas and Adenoca..       TRUE      TRUE
##                                                 name primary_site years_smoked
##                                               <character>       <list>    <numeric>
## TCGA-B0-5694-01A-11R-1541-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-CJ-4637-01A-02R-1325-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-CZ-4860-01A-01R-1305-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-B0-4706-01A-01R-1503-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-B4-5844-01A-11R-1672-07 Kidney Renal Clear C..       Kidney           NA
## ...                                                  ...          ...          ...
## TCGA-B4-5836-01A-11R-1672-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-A3-3308-01A-02R-1325-07 Kidney Renal Clear C..       Kidney           NA
```

```
## TCGA-AK-3440-01A-02R-1277-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-CJ-4871-01A-01R-1305-07 Kidney Renal Clear C..       Kidney           NA
## TCGA-AK-3460-01A-02R-1277-07 Kidney Renal Clear C..       Kidney           NA
##                              paper_patient paper_mRNA_cluster
##                                   <factor>          <integer>
## TCGA-B0-5694-01A-11R-1541-07  TCGA-B0-5694                  2
## TCGA-CJ-4637-01A-02R-1325-07  TCGA-CJ-4637                  3
## TCGA-CZ-4860-01A-01R-1305-07  NA                           NA
## TCGA-B0-4706-01A-01R-1503-07  TCGA-B0-4706                  2
## TCGA-B4-5844-01A-11R-1672-07  NA                           NA
## ...                                    ...               ...
## TCGA-B4-5836-01A-11R-1672-07  NA                           NA
## TCGA-A3-3308-01A-02R-1325-07  TCGA-A3-3308                  3
## TCGA-AK-3440-01A-02R-1277-07  NA                           NA
## TCGA-CJ-4871-01A-01R-1305-07  TCGA-CJ-4871                  3
## TCGA-AK-3460-01A-02R-1277-07  TCGA-AK-3460                  1
##                              paper_microRNA_cluster
##                                           <integer>
## TCGA-B0-5694-01A-11R-1541-07                      3
## TCGA-CJ-4637-01A-02R-1325-07                      2
## TCGA-CZ-4860-01A-01R-1305-07                     NA
## TCGA-B0-4706-01A-01R-1503-07                      4
## TCGA-B4-5844-01A-11R-1672-07                     NA
## ...                                             ...
## TCGA-B4-5836-01A-11R-1672-07                     NA
## TCGA-A3-3308-01A-02R-1325-07                      4
## TCGA-AK-3440-01A-02R-1277-07                     NA
## TCGA-CJ-4871-01A-01R-1305-07                      1
## TCGA-AK-3460-01A-02R-1277-07                      3
```

Se puede acceder a estos metadatos usando `$`, lo que hace más sencillo sustraer un objeto entero dado un fenotipo. Por ejemplo, podemos extraer todas las muestras que tengan una etiqueta para el estado vital:

```
se[, se$vital_status == "Dead" | se$vital_status == "Alive"]
```

```
## class: RangedSummarizedExperiment
## dim: 19947 606
## metadata(1): data_release
## assays(2): raw_count scaled_estimate
## rownames(19947): A1BG|1 A2M|2 ... TMED7-TICAM2|100302736
##   LOC100303728|100303728
## rowData names(4): gene_id entrezgene ensembl_gene_id
##   transcript_id.transcript_id_TCGA-B0-5694-01A-11R-1541-07
## colnames(606): TCGA-B0-5694-01A-11R-1541-07
##   TCGA-CJ-4637-01A-02R-1325-07 ... TCGA-CJ-4871-01A-01R-1305-07
##   TCGA-AK-3460-01A-02R-1277-07
## colData names(70): barcode patient ... paper_mRNA_cluster
##   paper_microRNA_cluster
```

En este Dataframe, a pesar de ser del proyecto TCGA-KIRC, hay 606 muestras, cuando en el de datos clínicos había 537. Parece que todas las muestras de datos de Expresión génica tienen su etiqueta de `vital_status` correspondiente. Encontramos 202 muertos y 404 vivos.

Vamos a buscar Missing values para la etiqueta de Fase del tumor para lo cual utilizaremos el siguiente código:

```r
which(is.na(se$ajcc_pathologic_stage))
```

```
## [1] 128 290 460
```

```r
se$patient[c(128,190,460)]
```

```
## [1] "TCGA-BP-4798" "TCGA-BP-4338" "TCGA-MM-A563"
```

Los códigos de los pacientes que no tienen esta etiqueta disponible son los mismos que no la tenían para los datos clínicos, lo cual tiene sentido.

Análisis y visualización de datos TCGA

Seguiremos el Pipeline mostrado en: https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html#TCGAanalyze:_Analyze_data_from_TCGA

Preprocesado de los datos de expresión génica

```r
KIRCMatrix <- assay(se,"raw_count")

# Para los datos de expresión génica si es necesario hacer una gráfica de cajas y bigotes y  and AAIC p
KIRCnaseq_CorOutliers <- TCGAanalyze_Preprocessing(se)
```

El resultado se muestra abajo:

```r
KIRCMatrix[1:10,1:2]
```

```
##             TCGA-B0-5694-01A-11R-1541-07 TCGA-CJ-4637-01A-02R-1325-07
## A1BG|1                             36.00                         87.66
## A2M|2                           71105.61                      47586.92
## NAT1|9                            111.00                        295.00
## NAT2|10                             5.00                        126.00
## SERPINA3|12                       422.00                        643.00
## AADAC|13                            0.00                         77.00
## AAMP|14                          4536.00                       4774.00
## AANAT|15                            1.00                          0.00
## AARS|16                          4255.00                       6358.00
## ABAT|18                           703.00                        951.00
```

Análisis de expresión diferencial (DEA)

Realizar DEA (*Differential expression analysis*) para identificar genes expresados diferencialmente (DEGs) utilizando la función `TCGAanalyze_DEA`.

`TCGAAnalyze_DEA` utiliza las siguiente funciones de R:

1. `edgeR::DGEList` que convierte la matriz de conteos en un objeto edgeR.
2. `edgeR::estimateCommonDisp` se le asigna a cada gen el mismo estimador de dispersión

3. `edgeR::exactTest`realiza el test de comparación por parejas (*pair-wise test*) para la expresión diferenciaal entre dos grupos.
4. `edgeR::topTags` coge el resultado de `exactTest()`, ajusta los p-valores crudos utilizando la corrección FDR y devuelve los genes más diferencialmente expresados.

Después, filtramos el resultado de dataDEGs con `abs(LogFC) >= 1`, y utilizamos la función `TCGAanalyze_levelTab`para crear una tabla con DEGs (genes diferencialmente expresados), log Fol Change (FC), false discovery rate (FDR), el nivel de expresión génica para las muestras de la condición 1 y la condición 2 y el valor Delta.

```
#Downstream análisis usando datos de expresión génica de muestras dde TCGA de IlluminaHiSeq_RNASeqV2 co

library(TCGAbiolinks)
dataNorm <- TCGAanalyze_Normalization(tabDF = se, geneInfo =  geneInfo)
```

Normalización

```
## I Need about  150 seconds for this Complete Normalization Upper Quantile  [Processing 80k elements /s

## Step 1 of 4: newSeqExpressionSet ...

## Step 2 of 4: withinLaneNormalization ...

## Step 3 of 4: betweenLaneNormalization ...

## Step 4 of 4: exprs ...
```

`TCGAanalyze_Normalization` nos ha devuelto el archivo `dataNorm`, que se trata de una matriz bastante grande. Si lo observamos vemos que se trata de la matriz de conteos pero sin decimales.

```
dataNorm[1:10,1:3]
```

```
##          TCGA-B0-5694-01A-11R-1541-07 TCGA-CJ-4637-01A-02R-1325-07
## A1BG                               36                           88
## A2M                             71106                        47587
## NAT1                              111                          295
## NAT2                                5                          126
## SERPINA3                          422                          643
## AADAC                               0                           77
## AAMP                             4536                         4774
## AANAT                               1                            0
## AARS                             4255                         6358
## ABAT                              703                          951
##          TCGA-CZ-4860-01A-01R-1305-07
## A1BG                               60
## A2M                             39078
## NAT1                              208
## NAT2                                9
## SERPINA3                         1568
## AADAC                               4
## AAMP                            11681
## AANAT                               0
## AARS                            30018
## ABAT                              302
```

Filtrado Queremos una variabilidad mayor del percentil 75 y hacer el experimento con estos genes

```
#quantile filter of genes

dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm, method = "quantile", qnt.cut = 0.75)
```

El archivo resultante `dataFilt` se ha quedado en un tamaño de 23 Mb (2963946 ejementos), mientras que el archivo con los datos normalizados pesaba 91.7 Mb.

Separación de muestras Si queremos seguir con el DEA tendremos que separar las muestras según nuestra variable de interés. En nuestro caso, la variable de interés es vital_status.

```
samplesD <- subset(se$barcode, se$vital_status == "Dead")
samplesA <- subset(se$barcode, se$vital_status == "Alive")
```

```
# Diff.expr.analysis (DEA)
dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,samplesD],
                            mat2 = dataFilt[,samplesA],
                            Cond1type = "Dead",
                            Cond2type = "Alive",
                            fdr.cut = 0.01 ,
                            logFC.cut = 1,
                            method = "glmLRT")
```

Diff.expr.analysis (DEA)

```
## Batch correction skipped since no factors provided

## ----------------------- DEA -------------------------------

## there are Cond1 type Dead in  202 samples

## there are Cond2 type Alive in  404 samples

## there are  4965 features as miRNA or genes

## I Need about  100 seconds for this DEA. [Processing 30k elements /s]

## ----------------------- END DEA -------------------------------
```

Tras esto, podemos crear la tabla resumen:

```
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGs, "Dead", "Alive", dataFilt[,samplesD], dataFilt[,samp
```

```
dataDEGsFiltLevel
```

```
##            mRNA     logFC          FDR      Dead     Alive     Delta
## IGF2       IGF2 -1.553384 4.855987e-10 30784.376  9484.958 47819.951
## H19         H19 -1.215643 3.922022e-10 32273.149 14034.324 39232.618
## FGA         FGA -1.058576 8.634841e-04  8677.193  3769.448  9185.464
## HP           HP  1.557566 4.685048e-05  4652.525 13841.921  7246.617
## SAA1       SAA1 -1.324132 5.498045e-05  5043.455  2285.344  6678.203
## PPP1R1A PPP1R1A -1.074558 6.227843e-05  4402.223  1958.156  4730.443
## LBP         LBP -1.111236 7.420168e-04  4094.520  1889.441  4549.979
## MDK         MDK -1.037098 9.161921e-15  4171.124  2025.777  4325.863
## ALB         ALB  2.487878 8.271861e-12  1296.817  7953.696  3226.322
## SLC6A19 SLC6A19  1.184586 2.841919e-05  2361.856  5424.213  2797.822
```