Cálculo de DEGs
Datos bulk RNA-seq: TCGA_SKCM

Elena Eyre Sánchez, PhD

2024-11-23

Contents

1 Introducción y Objetivo

Este análisis es de especial interés para el TFM debido a que permitirá obtener listados de genes que cambien de manera significativa pero también de manera específica para cada tipo celular en el estudio en cuestión. Además, también permitirá obtener resultados de un análisis de pathways que aporte información sobre las vías afectadas en los tipos celulares de este tratamiento.

Los listados se guardarán para posteriores comparaciones entre tratamientos.

2 Paquetes y datos

En este análisis utilizo los datos del estudio TCGA-SKCM descargados de la base de datos TCGA: https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Melanoma%20(SKCM)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443.

Las muestras consisten en 470 muestras analizadas.

```
setwd("~/Desktop/ELENA_UOC/TFM")

gex <- read_delim("Datasets/TCGA-SKCM.htseq_fpkm.tsv",  delim = "\t",
                  escape_double = FALSE, trim_ws = TRUE)
bulk_metadata <- read_delim("Datasets/TCGA-SKCM.GDC_phenotype.tsv", delim = "\t",
                      escape_double = FALSE, trim_ws = TRUE)
#os_data <- read_delim("Datasets/TCGA-SKCM.survival.tsv", delim = "\t",
#                      escape_double = FALSE, trim_ws = TRUE)

# Debido a que los autores proporcionan los genes con la nomeclatura de ENSEMBL, lo convierto a símbolo
mart <- useDataset("hsapiens_gene_ensembl", useMart("ensembl"))
genes <- gex$Ensembl_ID
G_list <- getBM(filters= "ensembl_gene_id_version", attributes= c("ensembl_gene_id_version","hgnc_symbol

# Para usar los símbolos en lugar de nombres de ENSEMBL:
```

```r
gex2 <- merge(gex,G_list,by.x="Ensembl_ID",by.y="ensembl_gene_id_version")
gex2<-gex2[,-1]
gex2 <- aggregate(gex2, by = list(c(gex2$hgnc_symbol)), mean) # Agrego los posibles duplicados calculan
gex2 <- gex2[,-c(ncol(gex2))]
rownames(gex2) <- gex2$Group.1 # Los nombres de genes únicos sin duplicados sirven para dar nombre a la
Probes <- gex2$Group.1
gex2$Group.1 <- NULL # Elimino las columnas usadas para conseguir los nombres
bulk.mtx <- as.data.frame(gex2) # Los datos de expresión

# Convertir los datos de expresión del bulk RNA-seq a objeto ExpressionSet:
bulk.eset <- Biobase::ExpressionSet(assayData = as.matrix(as.data.frame(bulk.mtx)))
bulk.eset
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 15584 features, 472 samples
##   element names: exprs
## protocolData: none
## phenoData: none
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

Para poder ejecutar las proporciones celulares en el modelo lineal, cargo los datos guardados de la deconvolución. Con estos datos, genero una tabla con las fracciones y los confounders que estén disponibles (me serviá como atributo "y" del modelo lineal): supervivencia de los pacientes, género, tipo de tumor, o estadio del tumor.

```
## Rows: 472 Columns: 121
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (75): Sample, submitter_id.samples, batch_number, bcr, bcr_followup_barc...
## dbl (40): B.cells, Macrophages.M1, Macrophages.M2, Monocytes, Neutrophils, N...
## lgl  (6): withdrawn, releasable.project, days_to_sample_procurement.samples,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
## Warning: Unknown or uninitialised column: `loc_melan`.
## Unknown or uninitialised column: `loc_melan`.
## Unknown or uninitialised column: `loc_melan`.
```

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

3 Modelo lineal

Utilizando las fracciones celulares como confounders, en principio el modelo lineal buscaría los coeficientes correspondientes a éstas.

Con eBayes se podría calcular la diferenciación de las expresiones (fold change, etc).

```
## Coefficients not estimable: B.cells Neutrophils NK.cells T.cells.CD4 T.cells.CD8 Tregs Dendritic.cel
```

```
## Warning: Partial NA coefficients for 15584 probe(s)
```

```
## Warning: Zero sample variances detected, have been offset away from zero
```

```
## Warning in .ebayes(fit = fit, proportion = proportion, stdev.coef.lim =
```

```
## stdev.coef.lim, : Estimation of var.prior failed - set to default value
```

3.1 DEGs

Con topTable se puede extraer la información para cada tipo celular. Indicando la columna de tipo celular extraigo el cálculo estadístico específico. Cone esta información se puede definir los grupos expresados diferencialmente, y generar un volcano plot.

Esta función, topTable, permite también extraer un número determinado de resultados, cosa que me permite extraer los DEGs que utilizo para realizar un análisis de enriquimiento con el paquete `gprofiler2`.

- Células B: no hay resultados debido a la cantidad de ceros en esta fracción celular.

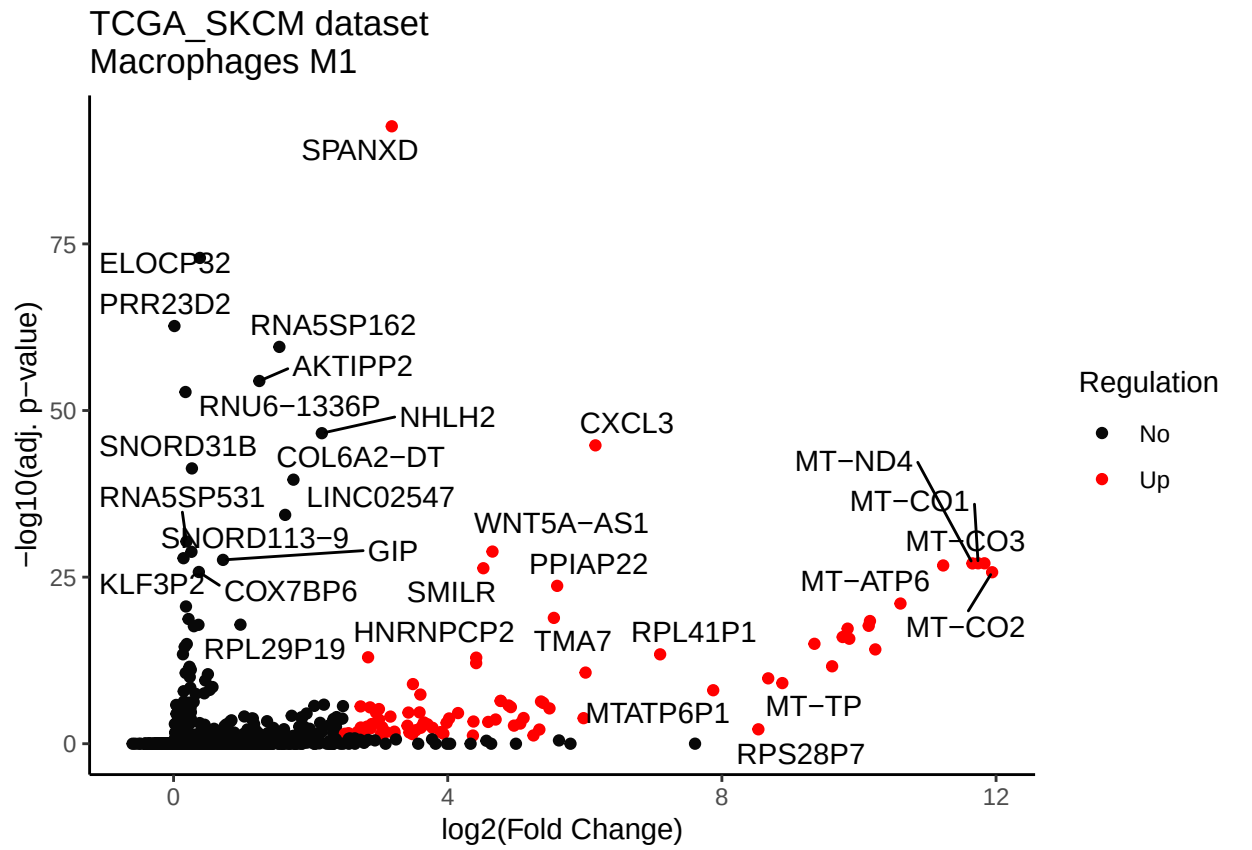
```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

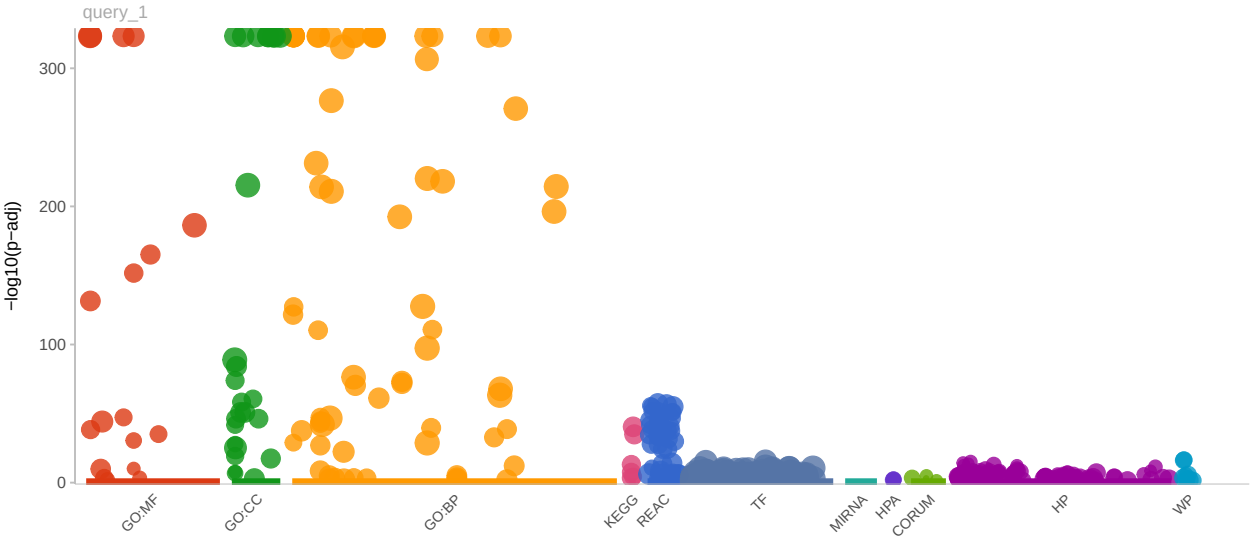
- Macrófagos M1:

```
##                 logFC      AveExpr        t      P.Value    adj.P.Val        B
## SPANXD     3.18370731 3.066387e-02 27.05591 1.445049e-97 2.251964e-93 210.4903
## ELOCP32    0.38434269 3.788825e-03 22.73463 1.629387e-77 1.269619e-73 165.1424
## PRR23D2    0.01266273 5.135633e-05 20.51952 4.052234e-67 2.105001e-63 141.5838
## RNA5SP162  1.54367297 1.234875e-02 19.82512 7.268994e-64 2.832000e-60 134.2047
## AKTIPP2    1.25131232 5.575803e-02 18.70711 1.200306e-58 3.741115e-55 122.3671
## RNU6-1336P 0.17407944 8.013907e-04 18.33445 6.457000e-57 1.677098e-53 118.4396
##            signif
## SPANXD         Up
## ELOCP32        No
## PRR23D2        No
## RNA5SP162      No
## AKTIPP2        No
## RNU6-1336P     No
```

```
## [1] "Número de genes: 84 up-regulated, 0 down-regulated"
```

```
## Warning: ggrepel: 90 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```





id	source	term_id	term_name	term_size	intersection_size		p_value
1	GO:BP	GO:0000244	spliceosomal tri−snRNP complex assembly	1211	1144		4.9e−324
2	GO:BP	GO:0000353	formation of quadruple SL/U4/U5/U6 snRNP	1142	1091		4.9e−324
3	GO:BP	GO:0000375	RNA splicing, via transesterification reactions	1735	1330		4.9e−324
4	GO:BP	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1731	1330		4.9e−324
5	GO:BP	GO:0000387	spliceosomal snRNP assembly	1237	1144		4.9e−324
6	GO:BP	GO:0000398	mRNA splicing, via spliceosome	1731	1330		4.9e−324

g:Profiler (biit.cs.ut.ee/gprofiler)

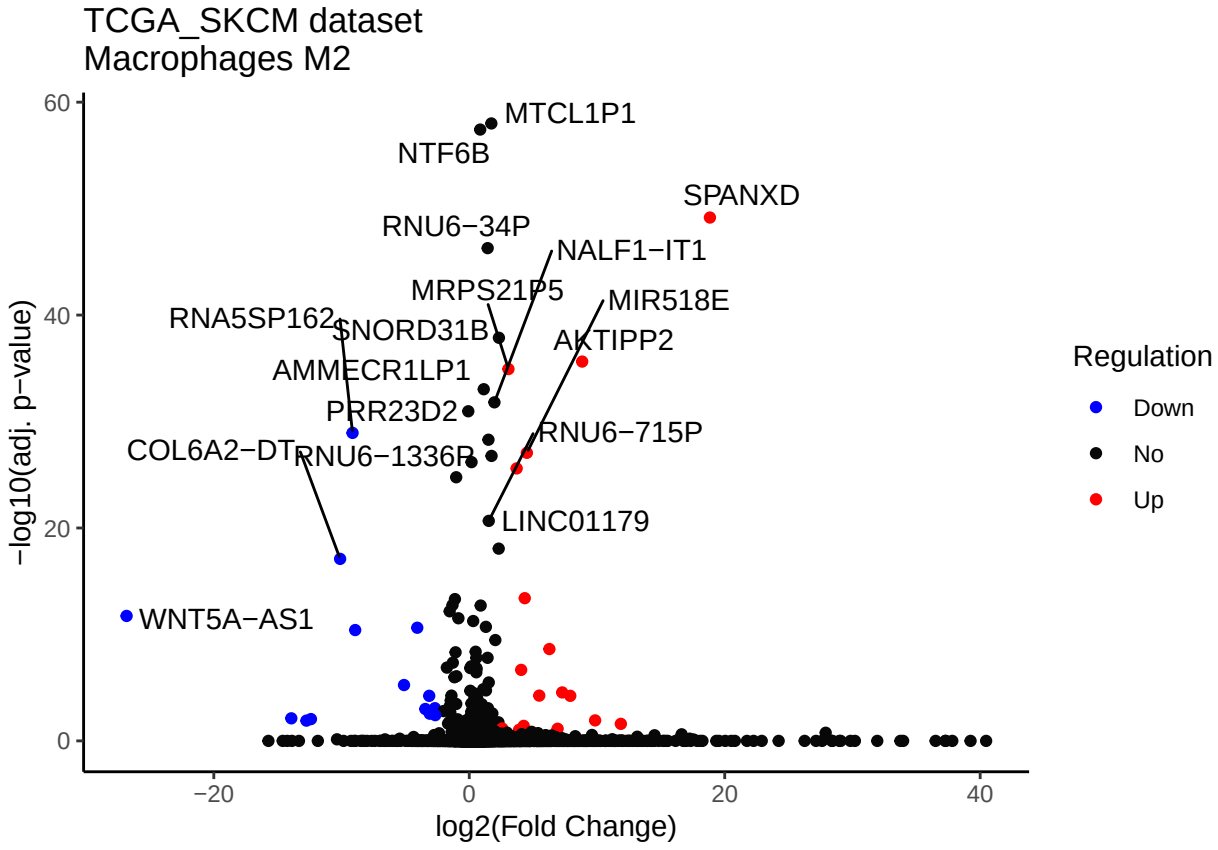
- Macrófagos M2:

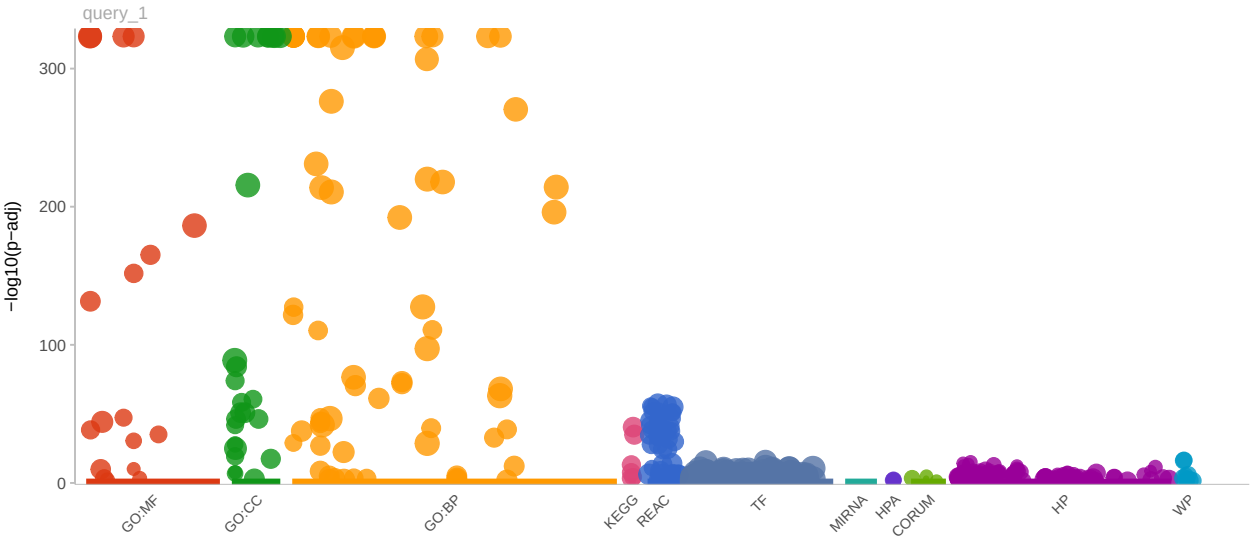
```
##                logFC      AveExpr         t      P.Value    adj.P.Val          B
```

```
## MTCL1P1   1.7302022 0.0022235116 19.62329 6.392643e-63 9.962295e-59 129.18620
## NTF6B     0.8564832 0.0008014510 19.43591 4.803453e-62 3.742850e-58 127.26442
## SPANXD   18.8434883 0.0306638670 17.61836 1.317692e-53 6.844973e-50 108.71659
## RNU6-34P  1.4411267 0.0008648348 16.96411 1.321848e-50 5.149921e-47 102.10588
## SNORD31B  2.3238382 0.0025417213 15.07070 4.363035e-42 1.359871e-38  83.30969
## AKTIPP2   8.8424492 0.0557580323 14.54416 8.951684e-40 2.325051e-36  78.20060
##          signif
## MTCL1P1      No
## NTF6B        No
## SPANXD       Up
## RNU6-34P     No
## SNORD31B     No
## AKTIPP2      Up
```

```
## [1] "Número de genes: 18 up-regulated, 14 down-regulated"
```

```
## Warning: ggrepel: 954 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```





id	source	term_id	term_name	term_size	intersection_size		p_value
.	GO:BP	GO:0000244	spliceosomal tri−snRNP complex assembly	1211	1144		4.9e−324
2	GO:BP	GO:0000353	formation of quadruple SL/U4/U5/U6 snRNP	1142	1091		4.9e−324
3	GO:BP	GO:0000375	RNA splicing, via transesterification reactions	1735	1330		4.9e−324
4	GO:BP	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1731	1330		4.9e−324
5	GO:BP	GO:0000387	spliceosomal snRNP assembly	1237	1144		4.9e−324
6	GO:BP	GO:0000398	mRNA splicing, via spliceosome	1731	1330		4.9e−324

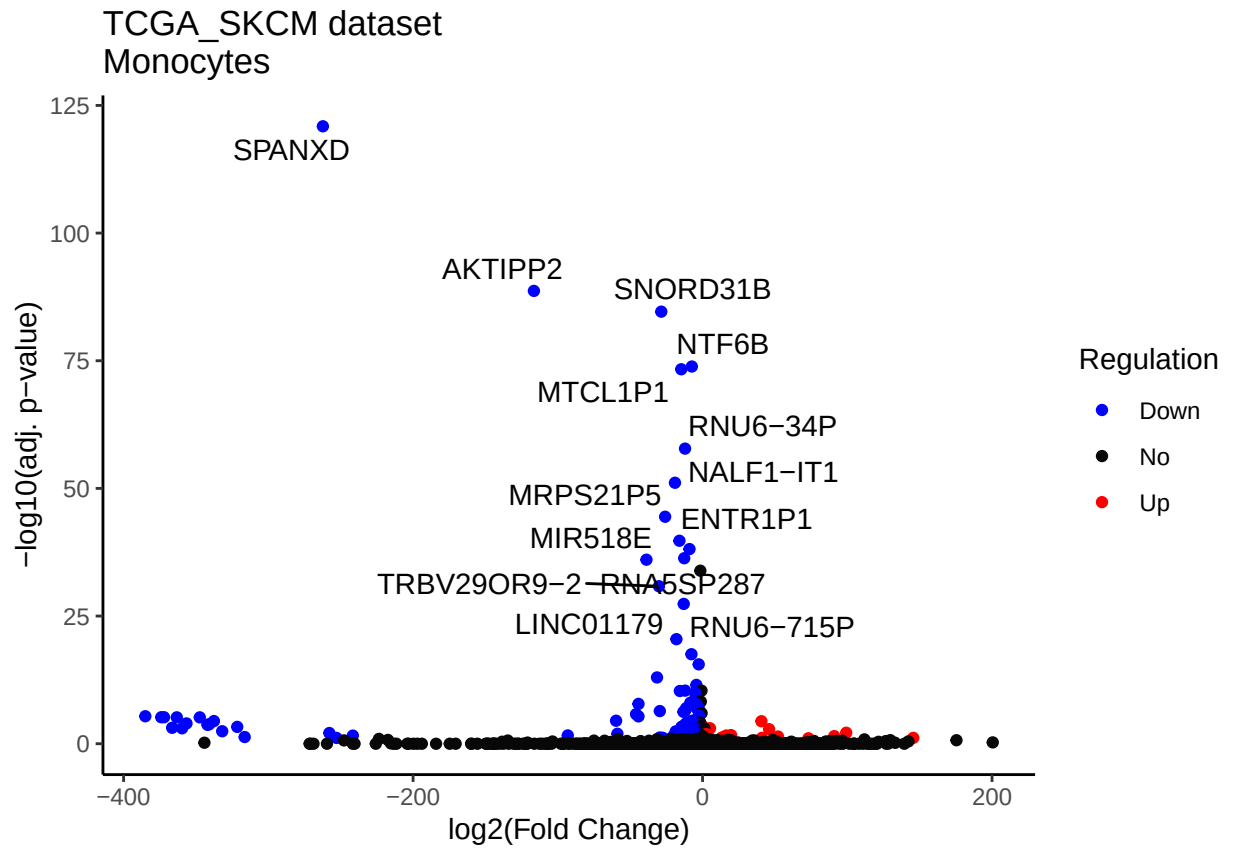
g:Profiler (biit.cs.ut.ee/gprofiler)

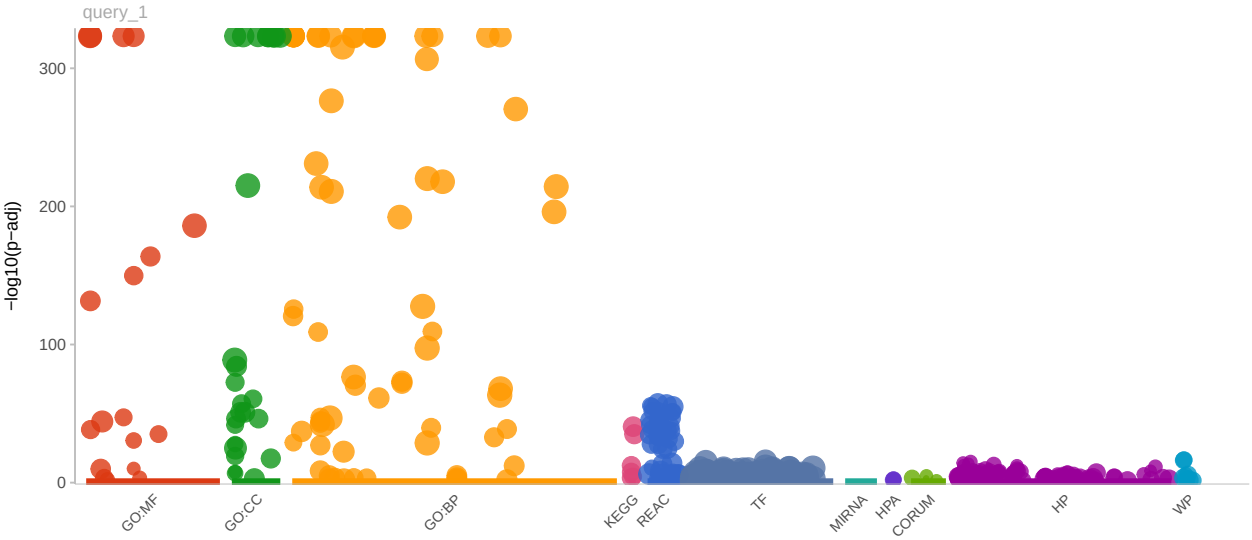
- Monocitos:

```
##                logFC      AveExpr         t       P.Value     adj.P.Val
## SPANXD   -262.295344 0.0306638670 -33.43484 7.838007e-126 1.221475e-121
## AKTIPP2  -116.504045 0.0557580323 -26.12532  2.705163e-93  2.107863e-89
## SNORD31B  -28.510437 0.0025417213 -25.20786  4.708849e-89  2.446090e-85
## NTF6B      -7.394839 0.0008014510 -22.87806  3.465411e-78  1.350124e-74
## MTCL1P1   -14.708145 0.0022235116 -22.74246  1.497232e-77  4.666573e-74
## RNU6-34P  -12.095727 0.0008648348 -19.41178  6.226511e-62  1.617232e-58
##                 B signif
## SPANXD   263.2739   Down
## AKTIPP2  194.3492   Down
## SNORD31B 185.2448   Down
## NTF6B    161.8054   Down
## MTCL1P1  160.4299   Down
## RNU6-34P 126.4822   Down
```

```
## [1] "Número de genes: 20 up-regulated, 113 down-regulated"
```

```
## Warning: ggrepel: 131 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```





id	source	term_id	term_name	term_size	intersection_size		p_value
1	GO:BP	GO:0000244	spliceosomal tri−snRNP complex assembly	1211	1144		4.9e−324
2	GO:BP	GO:0000353	formation of quadruple SL/U4/U5/U6 snRNP	1142	1091		4.9e−324
3	GO:BP	GO:0000375	RNA splicing, via transesterification reactions	1735	1329		4.9e−324
4	GO:BP	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1731	1329		4.9e−324
5	GO:BP	GO:0000387	spliceosomal snRNP assembly	1237	1144		4.9e−324
6	GO:BP	GO:0000398	mRNA splicing, via spliceosome	1731	1329		4.9e−324

g:Profiler (biit.cs.ut.ee/gprofiler)

- Neutrófilos: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##            logFC      AveExpr  t P.Value adj.P.Val  B
```

```
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

- Células NK: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

- Células T CD4: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

- Células T CD8: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

- Células T reg: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```

- Células dendríticas: no hay resultados debido a la cantidad de ceros en esta fracción celular.

```
##           logFC      AveExpr  t P.Value adj.P.Val  B
##              NA 0.1089672010 NA      NA        NA NA
## A2ML1-AS1    NA 0.0088749102 NA      NA        NA NA
## A2ML1-AS2    NA 0.0038936472 NA      NA        NA NA
## A3GALT2      NA 0.0755132317 NA      NA        NA NA
## AARS1P1      NA 0.0003316436 NA      NA        NA NA
## AARSD1P1     NA 0.0708700399 NA      NA        NA NA
```