Deconvolución de datos de mealoma con Bisque: Datos bulk RNA-seq: GSE54467

Datos scRNA-seq: GSE72056

Elena Eyre Sánchez, PhD

2024-10-13

Contents

1 Introducción y Objetivo **1**

2 Paquetes y datos **1**
2.1 Bulk RNA-seq 2

3 scRNA-seq data **3**
3.1 Análisis de deconvolución 4

1 Introducción y Objetivo

2 Paquetes y datos

El paquete usado para este análisis es Bisque, el cual está diseñado para estimar proporciones celulares en datos bulk RNA-seq mediante el uso de datos scRNA-seq como referencia, cuando los datos bulk y scRNA-seq se generan con muestras con diferentes condiciones clínicas.

Repositorio GitHub de Bisque: https://github.com/cozygene/bisque

```
knitr::opts_chunk$set(warning=FALSE)
package_to_load <- c("readr", "dplyr", "ggplot2", "tidyr", "dplyr","RColorBrewer",
                     "Biobase","BisqueRNA", "gplots")
for (package in package_to_load) {
  require(package, character.only = T); packageVersion(package)
}
extra_to_load <- c("knitr","stringr", "stringi", "ggrepel","ggpubr", "ggbreak","reshape2","ggfortify","
for (package in extra_to_load) {
  require(package, character.only = T); packageVersion(package)
}
rm(package_to_load, extra_to_load)
```

#Datos

Hay dos tipos de input data: bulk RNA-seq y scRNA-seq. Los datos bulk RNA-seq son los que queremos deconvolucionar, y los datos scRNA-seq servirán como referencia de las poblaciones celulares a consultar.

`Bisque` requiere datos de expresión en formato ExpressionSet del paquete `Biobase`, así que previamente a aplicar `bisque` se necesita preparar los datos.

2.1 Bulk RNA-seq

En este análisis utilizo los datos del estudio GSE54467 descargados mediante la función `getGEO` des de la base de datos GEO, del NCBI: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54467.

Las muestras consisten en biopsias tumorales de estadio III. Este estudio es de especial interés para el TFM debido a que los autores también proporcionan metadata con la edad, el género, y la supervivencia de los pacientes, cosa que permitirá estudiar posibles correlaciones con las poblaciones obtenidas de la deconvolución.

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

gset <- getGEO("GSE54467", GSEMatrix =TRUE, getGPL=FALSE)
if (length(gset) > 1) idx <- grep("GPL6884", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# Expploración de las condiciones clínicas de los datos
#table(gset$characteristics_ch1) # Age at primary diagnosis
#table(gset$characteristics_ch1.1) # gender
#table(gset$characteristics_ch1.2) # Age at sample banked
#table(gset$characteristics_ch1.3) # Survival from stage iii fumor banked
#table(gset$characteristics_ch1.4) # Survival from primary melanoma
#table(gset$characteristics_ch1.5) # Patient last status (OS)
#table(gset$characteristics_ch1.6) # number of primary melanomas
#table(gset$characteristics_ch1.7) # stage at primary diagnosis

# Debido a que los autores proporcionan los genes con la nomeclatura de Illumina, lo convierto a símbol
x <- illuminaHumanv4SYMBOL # cargado con el paquete illuminaHumanv4.db
mapped_probes <- mappedkeys(x) # Para sacar los símbolos
xx <- as.list(x[mapped_probes]) # Lo paso a listado
my_genes <- as.data.frame(unlist(xx[(rownames(gset@assayData$exprs))])) # Lo convierto en tabla para po
my_genes$gene <- rownames(my_genes)

bulk_metadata <- as.data.frame(gset@phenoData@data) # Paso la metadata disponible a una tabla

#dim(gset@assayData$exprs) # 26085 genes    79 muestras
# Para usar los símbolos en lugar de nombres de ilumina, extraigo los datos de expresión:
bulk.mtx <- as.data.frame(gset@assayData$exprs) # Los datos de expresión
bulk.mtx$gene <- rownames(bulk.mtx) # La columna que usaré para integrar
bulk.mtx <- inner_join(my_genes, bulk.mtx, by = "gene") # Integración de ambas tablas
bulk.mtx$gene <- NULL # Elimino la columna con nombres de Illumina
colnames(bulk.mtx)[1] <- "symbols" # Nombro la columna de símbolos de los genes

# Agrego los posibles duplicados calculando la media:
bulk.mtx <- aggregate(bulk.mtx, by = list(c(bulk.mtx$symbols)), mean) # Agregación
rownames(bulk.mtx) <- bulk.mtx$Group.1 # Los nombres de genes únicos sin duplicados sirven para dar nom
bulk.mtx <- bulk.mtx[,-c(1:2)] # Elimino las columnas usadas para conseguir los nombres

# 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: 15354 features, 79 samples
##   element names: exprs
## protocolData: none
## phenoData: none
```

```
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

3 scRNA-seq data

Los datos de expresión single-cell RNA de secuenciación (scRNA-seq) se recogen de muestras con una única condición, por ejemplo, sanos. Los tipos celulares del scRNA-seq son pre-determinados. Estos sirven como una referencia para estimar las proporciones del tipo celular de los datos bulk.

Para este análisis he escogido los datos procedentes del estudio GSE72056, que se encuentran células tumorales y no tumorales. Por ello, uno de los pasos a dar es la selección de las células no tumorales, entre las que se encuentran las células inmunitarias, que son de nuestro interés.

De nuevo, también procesaré los datos para obtener un objeto ExpressionSet, que en este caso también contendrá información sobre los tipos celulares con los que se identifica cada célula del estudio.

```
# Carga del fichero txt, después de descargar el archivo zip de GEO y despemaquetarlo:
GSE72056_melanoma_single_cell_revised_v2 <- read_delim("Datasets/GSE72056_melanoma_single_cell_revised_
    delim = "\t", escape_double = FALSE, trim_ws = TRUE)

# Proceso la información de la metadata, pues está en las primeras 3 líneas, con nombres de columnas mu
sc_metadata <- as.data.frame(t(GSE72056_melanoma_single_cell_revised_v2[1:3,]))
colnames(sc_metadata) <- sc_metadata[1,] # Los nombres de las columnas estan en la primera línea
sc_metadata <- sc_metadata[-1,] # Elimino línea de los nombres de columnas
colnames(sc_metadata)[2] <- "malignant" # Simplifico el nombre de la columna
sc_metadata$malignant <- sapply(sc_metadata$malignant, as.numeric) # Para poder seleccionar el valor qu
sc_metadata <- sc_metadata[sc_metadata$malignant == 1,] # Selecciono sólo las células no tumorales
colnames(sc_metadata)[3] <- "non_malignant" # Simplifico el nombre de la columna
sc_metadata$SampleID <- rownames(sc_metadata) # Genero una columna para identificar las células en el o
sc_metadata$non_malignant <- sapply(sc_metadata$non_malignant, as.numeric) # Para poder convertir los n
sc_metadata$Cell_type<- as.factor(if_else(sc_metadata$non_malignant == 1, "T_cell",
                                  ifelse(sc_metadata$non_malignant == 2, "B_cell",
                                         ifelse(sc_metadata$non_malignant == 3, "Macrophage",
                                                ifelse(sc_metadata$non_malignant == 4, "Endo_cell",
                                                       ifelse(sc_metadata$non_malignant == 5, "CAF",
                                                              ifelse(sc_metadata$non_malignant == 6, "NK"

# Preparo los apartados asociados a la metadata, y a los tipos celulares que bisque necesita saber, que
sc.pheno <- data.frame(check.names=F, check.rows=F, stringsAsFactors=F, row.names=sc_metadata$SampleID,
                       SubjectName=sc_metadata$SampleID, cellType=sc_metadata$Cell_type) # Phenodata pa
sc.meta <- data.frame(labelDescription=c("SampleID","Cell_type"),row.names=c("SampleID","Cell_type")) #
sc.pdata <- new("AnnotatedDataFrame",data=sc.pheno,varMetadata=sc.meta) # Este parámetro contiene ambos

# Procesamiento de los niveles de expresión de las células del scRNA-seq, que irán en el objeto Express
sc_gex <-GSE72056_melanoma_single_cell_revised_v2[GSE72056_melanoma_single_cell_revised_v2$Cell %in% ro
sc_gex2 <- sc_gex[rowSums(sc_gex[,-1]) != 0,] # También simplifico el análisis eliminando los genes que
sc_gex2$Probes <- sc_gex2$Cell # Genero una nueva columna de los símbolos de los genes para que no sea
sc_gex2$Cell <- NULL # Elimino la columna con nombre confuso
sc_gex2 <- aggregate(sc_gex2[,-ncol(sc_gex2)], by= list(c(sc_gex2$Probes)), mean) # Agrego los posibles
#dim(sc_gex2) # 22844  4646
rownames(sc_gex2) <- sc_gex2$Group.1 # Los nombres de genes únicos sin duplicados sirven para dar nombr
sc_gex3 <- sc_gex2[,colnames(sc_gex2) %in% sc_metadata$SampleID] # Me quedo con las columnas que aparec

# Convertir los datos de expresión del scRNA-seq a objeto ExpressionSet:
```

```
sc.eset <- Biobase::ExpressionSet(assayData=as.matrix(sc_gex3), phenoData=sc.pdata)
sc.eset
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 14074 features, 3256 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: Cy72_CD45_H02_S758_comb CY58_1_CD45_B02_S974_comb ...
##     CY75_1_CD45_CD8_7__S274_comb (3256 total)
##   varLabels: SubjectName cellType
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

3.1 Análisis de deconvolución

Por defecto, Bisque utiliza todos los genes para decomposición.

Hay que indicar que no hay solapamiento de muestras por tratarse de muestras diferentes en el bulk RNA-seq y el scRNA-seq, pues Bisque podría realizar un análisis usando las mismas muestras si lo quisiéramos: `use.overlap=FALSE`.

```
res <- BisqueRNA::ReferenceBasedDecomposition(bulk.eset, sc.eset, markers=NULL, use.overlap=FALSE)
```

```
## Decomposing into 7 cell types.

## Using 14074 genes in both bulk and single-cell expression.

## Converting single-cell counts to CPM and filtering zero variance genes.

## Filtered 15 zero variance genes.

## Converting bulk counts to CPM and filtering unexpressed genes.

## Filtered 0 unexpressed genes.

## Generating single-cell based reference from 3256 cells.

## Inferring bulk transformation from single-cell alone.

## Applying transformation to bulk samples and decomposing.
```

Encontramos las proporciones del bulk RNA-seq en el apartado `bulk.props`, el qual puedo integrar en la metadata que ya tenía y almacenar en un archivo para posteriores análisis.

```
ref.based.estimates <- as.data.frame(t(res$bulk.props))
ref.based.estimates$geo_accession <- rownames(ref.based.estimates)
ref.based.estimates <- inner_join(ref.based.estimates, bulk_metadata, by = "geo_accession")
knitr::kable(head(ref.based.estimates[,1:7]), digits=2, caption = "Sección de las primeras muestras com
```

Table 1: Sección de las primeras muestras como ejemplo del resultado

B_cell	CAF	Endo_cell	Macrophage	NK	Other_cells	T_cell
0.14	0.00	0.03	0.11	0.01	0.04	0.67
0.07	0.02	0.06	0.00	0.00	0.25	0.59
0.00	0.06	0.06	0.00	0.00	0.34	0.54
0.20	0.00	0.00	0.13	0.01	0.00	0.66

B_cell	CAF	Endo_cell	Macrophage	NK	Other_cells	T_cell
0.18	0.08	0.07	0.00	0.00	0.12	0.55
0.00	0.05	0.01	0.00	0.00	0.48	0.47

```
write.csv(ref.based.estimates,"./bisque_GSE54467.csv", row.names = FALSE)
```