Deconvolución de datos de mealoma con Bisque

Primera sección

Elena Eyre Sánchez, PhD

2024-10-12

Contents

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 sc RNA-seq.

Bisque requere datos de expresión en formato ExpressionSet del paquete `Biobase`.

2.1 Bulk RNA-seq: GSE54467

Los datos de expresión de seqüenciación de Bulk RNA recogidos de muestras de dos condiciones clínicas diferentes, por ejemplo, sano y enfermo. Estos serían los datos que queremos deconvolucionar.

Ene ste estudio uso los datos del estudio GSE54467 descargados mediante la función `getGEO`.

Bulk RNA-seq data can be converted from a matrix (columns are samples, rows are genes) to an ExpressionSet as follows:

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

#gset <- getGEO("GSE65904", GSEMatrix =TRUE, getGPL=FALSE)
#if (length(gset) > 1) idx <- grep("GPL10558", attr(gset, "names")) else idx <- 1
#gset <- gset[[idx]]
#table(gset$characteristics_ch1) # gender
#table(gset$characteristics_ch1.2) # Tumor stage
#table(gset$characteristics_ch1.3) # Tissue
#table(gset$characteristics_ch1.4) # distant metastasis free survival
#table(gset$characteristics_ch1.5) # distant metastasis free survival (death/alive)
#table(gset$characteristics_ch1.6) # disease specific survival
#table(gset$characteristics_ch1.7) # disease specific survival (death/alive)

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

# Convert the object to a list
x <- illuminaHumanv4SYMBOL
# Get the probe identifiers that are mapped to a gene symbol
mapped_probes <- mappedkeys(x)
# Convert to a list
xx <- as.list(x[mapped_probes])
my_genes <- as.data.frame(unlist(xx[(rownames(gset@assayData$exprs))]))
my_genes$gene <- rownames(my_genes)

# clinical conditions
#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

#ex <- exprs(gset)
# log2 transform
#qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
#LogC <- (qx[5] > 100) ||
#          (qx[6]-qx[1] > 50 && qx[2] > 0)
#if (LogC) { ex[which(ex <= 0)] <- NaN
#  ex <- log2(ex) }

bulk_metadata <- as.data.frame(gset@phenoData@data)
#gset@annotation # GPL6884
#bulk_gex <- as.data.frame(gset@assayData$exprs)

dim(gset@assayData$exprs) # 26085    79
```

```
## [1] 26085    79
```

```
bulk.mtx <- as.data.frame(gset@assayData$exprs)
bulk.mtx$gene <- rownames(bulk.mtx)
bulk.mtx <- inner_join(my_genes, bulk.mtx, by = "gene")
bulk.mtx$gene <- NULL
colnames(bulk.mtx)[1] <- "symbols"
bulk.mtx <- aggregate(bulk.mtx, by = list(c(bulk.mtx$symbols)), mean)
rownames(bulk.mtx) <- bulk.mtx$Group.1
bulk.mtx <- bulk.mtx[,-c(1:2)]

bulk.eset <- Biobase::ExpressionSet(assayData = as.matrix(as.data.frame(bulk.mtx)))
```

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

Single-cell data requires additional information in the ExpressionSet, specificially cell type labels and individual labels. Individual labels indicate which individual each cell originated from. To add this information, Biobase requires it to be stored in a data frame format. Assuming we have character vectors of cell type labels (`cell.type.labels`) and individual labels (`individual.labels`), we can convert scRNA-seq data (with counts also in matrix format) as follows:

```
GSE72056_melanoma_single_cell_revised_v2 <- read_delim("Datasets/GSE72056_melanoma_single_cell_revised_
    delim = "\t", escape_double = FALSE,
    trim_ws = TRUE)

sc_metadata <- as.data.frame(t(GSE72056_melanoma_single_cell_revised_v2[1:3,]))
colnames(sc_metadata) <- sc_metadata[1,]
sc_metadata <- sc_metadata[-1,]
dim(sc_metadata) #
```

```
## [1] 4645    3
```

```
colnames(sc_metadata)[2] <- "malignant"
sc_metadata$malignant <- sapply(sc_metadata$malignant, as.numeric)
sc_metadata <- sc_metadata[sc_metadata$malignant == 1,] # Here we keep only non-malignan cells
colnames(sc_metadata)[3] <- "non_malignant"
#sc_metadata$T_cell <- (ifelse(sc_metadata$non_malignant == 1, "T_cell", "Other_cells"))
#sc_metadata$B_cell <- (ifelse(sc_metadata$non_malignant == 2, "B_cell", "Other_cells"))
#sc_metadata$M_cell <- (ifelse(sc_metadata$non_malignant == 3, "Macrophage", "Other_cells"))
#sc_metadata$E_cell <- (ifelse(sc_metadata$non_malignant == 4, "Endo_cell", "Other_cells"))
#sc_metadata$CAF_cell <- (ifelse(sc_metadata$non_malignant == 5, "CAF", "Other_cells"))
#sc_metadata$NK_cell <- (ifelse(sc_metadata$non_malignant == 6, "NK", "Other_cells"))
sc_metadata$SampleID <- rownames(sc_metadata)
sc_metadata$non_malignant <- sapply(sc_metadata$non_malignant, as.numeric)
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"
```

```
sc_gset <- getGEO("GSE72056", GSEMatrix =TRUE, getGPL=FALSE)
if (length(gset) > 1) idx <- grep("GPL6884", attr(gset, "names")) else idx <- 1
seger.sce <- gset[[idx]]

# individual.ids and cell.types should be in the same order as in sample.ids
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)
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)

sc_gex <-GSE72056_melanoma_single_cell_revised_v2[GSE72056_melanoma_single_cell_revised_v2$Cell %in% rou
#rownames(sc_gex) <- Probes
sc_gex2 <- sc_gex[rowSums(sc_gex[,-1]) != 0,]
sc_gex2$Probes <- sc_gex2$Cell #
sc_gex2$Cell <- NULL
#dim(sc_gex2) #
sc_gex2 <- aggregate(sc_gex2[,-ncol(sc_gex2)], by= list(c(sc_gex2$Probes)), mean)
#dim(sc_gex2) # 22844  4646
rownames(sc_gex2) <- sc_gex2$Group.1
sc_gex3 <- sc_gex2[,colnames(sc_gex2) %in% sc_metadata$SampleID]

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

Note that if you have samples with both single-cell and bulk RNA-seq data, their IDs should be found in both `sc.eset$SubjectName` and `sampleNames(bulk.eset)` .

3.1 Reference-based decomposition

By default, Bisque uses all genes for decomposition. However, you may supply a list of genes (such as marker genes) to be used with the `markers` parameter. Also, since we have samples with both bulk and single-cell RNA-seq data, we set the `use.overlap` parameter to `TRUE`. If there are no overlapping samples, you can set this parameter to `FALSE` (we expect performance to be better if overlapping samples are available).

Here's how to call the reference-based decomposition method:

```
#rownames(bulk.eset@assayData$exprs)
res <- BisqueRNA::ReferenceBasedDecomposition(bulk.eset, sc.eset, markers=NULL, use.overlap=FALSE)
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
```

```
##     IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
## 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.
```

A list is returned with decomposition estimates in slot `bulk.props`.

```
ref.based.estimates <- t(res$bulk.props)
knitr::kable(ref.based.estimates, digits=2)
```

	B_cell	CAF	Endo_cell	Macrophage	NK	Other_cells	T_cell
GSM1315904	0.14	0.00	0.03	0.11	0.01	0.04	0.67
GSM1315905	0.07	0.02	0.06	0.00	0.00	0.25	0.59
GSM1315906	0.00	0.06	0.06	0.00	0.00	0.34	0.54
GSM1315907	0.20	0.00	0.00	0.13	0.01	0.00	0.66
GSM1315908	0.18	0.08	0.07	0.00	0.00	0.12	0.55
GSM1315909	0.00	0.05	0.01	0.00	0.00	0.48	0.47
GSM1315910	0.12	0.00	0.00	0.11	0.01	0.00	0.76
GSM1315911	0.04	0.09	0.07	0.00	0.00	0.27	0.52
GSM1315912	0.07	0.04	0.09	0.00	0.00	0.29	0.51
GSM1315913	0.17	0.01	0.07	0.14	0.08	0.00	0.55
GSM1315914	0.06	0.03	0.07	0.00	0.00	0.39	0.45
GSM1315915	0.20	0.00	0.01	0.17	0.04	0.00	0.57
GSM1315916	0.00	0.03	0.02	0.00	0.00	0.49	0.46
GSM1315917	0.12	0.00	0.01	0.00	0.00	0.35	0.53
GSM1315918	0.33	0.00	0.00	0.00	0.00	0.00	0.67
GSM1315919	0.14	0.00	0.04	0.01	0.00	0.28	0.53
GSM1315920	0.01	0.03	0.04	0.00	0.00	0.57	0.36
GSM1315921	0.16	0.00	0.00	0.00	0.02	0.00	0.82
GSM1315922	0.20	0.00	0.00	0.04	0.03	0.00	0.73

	B_cell	CAF	Endo_cell	Macrophage	NK	Other_cells	T_cell
GSM1315923	0.06	0.04	0.00	0.06	0.01	0.12	0.70
GSM1315924	0.07	0.00	0.00	0.12	0.06	0.00	0.75
GSM1315925	0.16	0.02	0.00	0.08	0.02	0.29	0.43
GSM1315926	0.03	0.03	0.01	0.00	0.00	0.52	0.41
GSM1315927	0.15	0.00	0.00	0.05	0.01	0.13	0.66
GSM1315928	0.00	0.01	0.04	0.00	0.00	0.55	0.41
GSM1315929	0.06	0.00	0.00	0.14	0.00	0.33	0.47
GSM1315930	0.09	0.03	0.08	0.08	0.10	0.00	0.62
GSM1315931	0.16	0.00	0.04	0.17	0.08	0.00	0.55
GSM1315932	0.07	0.02	0.04	0.00	0.00	0.39	0.48
GSM1315933	0.22	0.01	0.02	0.14	0.02	0.00	0.60
GSM1315934	0.22	0.00	0.00	0.00	0.00	0.18	0.60
GSM1315935	0.09	0.12	0.04	0.00	0.00	0.22	0.54
GSM1315936	0.15	0.01	0.02	0.00	0.00	0.27	0.54
GSM1315937	0.00	0.08	0.06	0.00	0.00	0.41	0.44
GSM1315938	0.22	0.06	0.09	0.00	0.00	0.03	0.60
GSM1315939	0.06	0.00	0.00	0.00	0.00	0.47	0.47
GSM1315940	0.29	0.00	0.00	0.04	0.05	0.00	0.63
GSM1315941	0.28	0.00	0.00	0.01	0.00	0.00	0.71
GSM1315942	0.08	0.06	0.00	0.07	0.00	0.30	0.50
GSM1315943	0.35	0.00	0.00	0.00	0.00	0.00	0.64
GSM1315944	0.19	0.00	0.00	0.14	0.04	0.00	0.63
GSM1315945	0.06	0.02	0.00	0.10	0.00	0.30	0.51
GSM1315946	0.27	0.00	0.00	0.00	0.00	0.00	0.73
GSM1315947	0.06	0.04	0.05	0.06	0.02	0.21	0.57
GSM1315948	0.04	0.04	0.03	0.00	0.00	0.45	0.44
GSM1315949	0.35	0.00	0.00	0.07	0.02	0.00	0.56
GSM1315950	0.12	0.14	0.01	0.04	0.03	0.16	0.49
GSM1315951	0.17	0.00	0.10	0.00	0.00	0.00	0.72
GSM1315952	0.00	0.05	0.02	0.00	0.00	0.52	0.40
GSM1315953	0.13	0.00	0.00	0.08	0.02	0.00	0.77
GSM1315954	0.07	0.04	0.04	0.04	0.01	0.27	0.53
GSM1315955	0.11	0.00	0.00	0.09	0.01	0.15	0.63
GSM1315956	0.19	0.00	0.00	0.09	0.02	0.00	0.70
GSM1315957	0.34	0.00	0.00	0.00	0.00	0.00	0.66
GSM1315958	0.14	0.08	0.03	0.03	0.02	0.28	0.43
GSM1315959	0.31	0.05	0.00	0.00	0.00	0.04	0.60
GSM1315960	0.20	0.00	0.00	0.12	0.04	0.00	0.64
GSM1315961	0.21	0.00	0.00	0.10	0.02	0.03	0.63
GSM1315962	0.16	0.00	0.00	0.14	0.03	0.00	0.67
GSM1315963	0.22	0.05	0.06	0.08	0.08	0.00	0.51
GSM1315964	0.15	0.07	0.04	0.13	0.04	0.12	0.45
GSM1315965	0.27	0.01	0.00	0.00	0.00	0.05	0.67
GSM1315966	0.20	0.00	0.00	0.08	0.00	0.00	0.72
GSM1315967	0.02	0.04	0.00	0.02	0.04	0.19	0.69
GSM1315968	0.23	0.00	0.00	0.02	0.02	0.00	0.73
GSM1315969	0.13	0.00	0.00	0.08	0.04	0.12	0.64
GSM1315970	0.03	0.00	0.01	0.00	0.00	0.33	0.62
GSM1315971	0.30	0.00	0.00	0.00	0.00	0.00	0.70
GSM1315972	0.21	0.00	0.00	0.10	0.03	0.00	0.66
GSM1315973	0.13	0.03	0.07	0.04	0.01	0.06	0.66
GSM1315974	0.08	0.01	0.03	0.02	0.04	0.13	0.70

	B_cell	CAF	Endo_cell	Macrophage	NK	Other_cells	T_cell
GSM1315975	0.16	0.12	0.10	0.00	0.00	0.00	0.62
GSM1315976	0.06	0.00	0.00	0.20	0.08	0.00	0.67
GSM1315977	0.13	0.00	0.00	0.02	0.03	0.30	0.53
GSM1315978	0.05	0.00	0.01	0.00	0.00	0.51	0.44
GSM1315979	0.07	0.04	0.06	0.00	0.00	0.35	0.49
GSM1315980	0.09	0.15	0.08	0.00	0.00	0.10	0.58
GSM1315981	0.01	0.01	0.02	0.00	0.00	0.38	0.59
GSM1315982	0.15	0.00	0.00	0.13	0.03	0.00	0.70