PEC 1

Bruno Bel

2025-03-31

Contents

1. Selección Dataset **1**

Objetivo del estudio: 1

Archivos del Dataset 2

2. Creación de SummarizedExperminet / ExpressiónSet **2**

Diferencias entre las clases de objetos 5

3. Análisis exploratorio **6**

Ejemplo de análisis 8

1. Selección Dataset

Los datos seleccionados para la realización de está PEC 1, provienen del estudio *LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice* . Efectuado por Liu Haiyan en la Universidad de Florida.

Las procianidinas son un tipo de flavonoides, específicamente proantocianidinas, que se encuentran en muchas frutas, semillas y cortezas de plantas. Son conocidas por sus propiedades antioxidantes y sus posibles beneficios para la salud cardiovascular, la inflamación y la función cognitiva.

En este estudio se reclutaron 18 mujeres jóvenes (21-29 años) con un índice de masa corporal (IMC) normal (18.5-25). Se les proporcionó una lista de alimentos ricos en procianidinas (arándanos, manzanas, uvas, chocolate, etc.) y se les indicó evitarlos antes y durante el estudio.

El estudio consistió en dos fases en las que las participantes consumieron zumo de arándano o zumo de manzana durante tres días, con un período intermedio de dos semanas entre ambas. Se tomaron muestras de sangre y orina en distintos momentos para evaluar los efectos metabólicos.

Objetivo del estudio:

Investigar los cambios metabólicos provocados por los concentrados de procianidinas en arándanos y manzanas mediante un enfoque metabolómico basado en LC-MS (cromatografía líquida acoplada a espectrometría de masas).

Archivos del Dataset

Del repositorio podemos obtener disintos archivos en crudo para trabajar con ellos. En formato csv disponemos del archivo de metadata, donde se contiene la información de cada muestra y el grupo de estudio. El archivo de features donde se contiene los datos de los metabolitos analizados para las 45 muestras. Por último el fichero que relaciona cada metabolito con su nombre original y el ID de dos bases de dato bioquímicas.

2. Creación de SummarizedExperminet / ExpressiónSet

En primer lugar importamos los documentos del dataset de interés.

```
## Primeras observaciones del archivo: features
##                b1      b10      b11      b12      b13      b14      b15
## 443489     941000   757000   612000   858000   185000   671000  1140000
## 107754    8300000  6790000 20800000   320000  1290000  1580000  2340000
## 9543071      1500      890 16200000     1250      968      657      809
## 11011465   276000    35700   631000   369000   242000   472000  5320000
## 5281160    706000   121000 11600000   164000   424000   749000   267000
## 440341       6340    34100    31900     9440    92600     6740    14400
##                b16      b17       b2       b4       b6       b7       b8
## 443489     108000   383000   593000  7240000   494000   812000  1290000
## 107754    1180000  1260000 15000000   495000    58100  1350000  1860000
## 9543071       767      826     2810     1140     1010      635     1280
## 11011465    18000   243000   131000   158000   208000   228000   119000
## 5281160   3050000    99100   136000   452000    75600   132000   341000
## 440341       8180     8980     4610    10100     8180     5920     1950
##                b9       a1       a10      a11      a12      a13      a14
## 443489       66000   215000   310000   798000  1070000   228000   241000
## 107754      698000  1220000  6920000 18700000  1320000  1230000  1980000
## 9543071        664      644     1060     1500        0      818      660
## 11011465     58000    17700   394000  4230000  3740000   361000    63300
## 5281160     119000    51600    54900 26700000   323000   152000   208000
## 440341        1810     4350     1450        0    56200     3700     8360
##                a15      a16      a17       a2       a4       a6       a7
## 443489     1180000    15100   255000   411000   463000   242000  1010000
## 107754     6980000   716000   761000  2910000 11300000   689000  1350000
## 9543071        754      695      562      851      766      637      846
## 11011465  2090000    37400    13400   260000   347000   151000  1080000
## 5281160   1400000    76100    16500   374000   491000    81200  1000000
## 440341       2590     1390     1090    30400     2340     2930     7060
##                a8       a9       c1       c10      c11      c12      c13
## 443489      702000    44600   136000  1060000  1050000   464000  1460000
## 107754     1130000   479000   652000  2200000  7380000   187000  1430000
## 9543071          0      618      546      926   800000        0        0
## 11011465     5080     2140   266000   627000  3140000   127000   197000
## 5281160   7000000    22400  1500000   171000   331000   198000   110000
## 440341       2830        0        0     4460        0     3190    22900
##                c14      c15      c16      c17       c2       c4       c6
## 443489      636000  4510000   146000   400000   783000   213000   816000
## 107754     9730000 11200000  6660000  1830000 15100000   971000   574000
## 9543071        809     1380      982      625     1790      626      991
```

```
## 11011465   286000   545000    35800    23200   230000    59600    48100
## 5281160    178000   791000    44100    57100   150000    29500   126000
## 440341   49200000        0     6930     1730     2400     3450     2880
##                c7       c8       c9
## 443489     587000   319000   102000
## 107754    4590000  9730000   644000
## 9543071      1600      949      756
## 11011465    44000   576000    14200
## 5281160    646000   291000    58900
## 440341       2450    11200     3570
##
## Primeras observaciones del archivo: metaboliteNames
##                                              names  PubChem   KEGG
## 1                      10-Desacetyltaxuyunnanin C  5460449 C15538
## 2                         10-Hydroxydecanoic acid    74300 C02774
## 3                               10-Oxodecanoate_1 19734156 C02217
## 4 11beta,21-Dihydroxy-5beta-pregnane-3,20-dione 21145110 C05475
## 5                   1,1-Dichloroethylene epoxide   119521 C14857
## 6                         11-Hydroxycanthin-6-one   337601 C09212
##
## Primeras observaciones del archivo: metadata
##    ID Treatment
## 1  b1  Baseline
## 2 b10  Baseline
## 3 b11  Baseline
## 4 b12  Baseline
## 5 b13  Baseline
## 6 b14  Baseline
```

Instalación de Bioconductor

```
## Bioconductor version 3.20 (BiocManager 1.30.25), R 4.4.3 (2025-02-28 ucrt)

## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.4.3/library
##   packages:
##     cluster, foreign, MASS, Matrix, nlme

## Bioconductor version 3.20 (BiocManager 1.30.25), R 4.4.3 (2025-02-28 ucrt)

## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.4.3/library
##   packages:
##     cluster, foreign, MASS, Matrix, nlme

## Cargando paquete requerido: BiocGenerics

##
## Adjuntando el paquete: '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, saveRDS, 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")'.

## Cargando paquete requerido: MatrixGenerics

## Cargando paquete requerido: matrixStats

##
## Adjuntando el paquete: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

##
## Adjuntando el paquete: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## The following object is masked from 'package:Biobase':
##
##     rowMedians

## Cargando paquete requerido: GenomicRanges
```

```
## Cargando paquete requerido: stats4

## Cargando paquete requerido: S4Vectors

##
## Adjuntando el paquete: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Cargando paquete requerido: IRanges

##
## Adjuntando el paquete: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Cargando paquete requerido: GenomeInfoDb
```

En primer lugar debemos asegurar que las dimensiones de los objetos son comatibles y los nombres de las variables del objeto features, que son las muestras codificadas, son identicos al nombre en el archivo de metadata.

```
## [1] 1541   45

## [1] 45  1

## [1] TRUE
```

Luego bservamos que los nombres de las columnas de matriz de datos y las filas de metadata coinciden todas.

```
## Todos los códigos de metaboliteNames se encuentran en features.

## [1] TRUE
```

Una vez comprobado que los nombre coinciden procedemos a crear el objeto SummarizedExperiment.

Diferencias entre las clases de objetos

Mientras que el objeto de clase ExpressionSet está centrado en la expressión génica y microarrays los objetos de clase SummarizedExperiments son más genericos.

SummarizedExperiment permite almacenar diversas matrices de datos independientemente del tipo de ómica de estudio.

Además SummarizedExperiment contiene el tipo de objeto como GRanges, rowRanges o rowData para añadir más información sobre datos genómicos.

3. Análisis exploratorio

Observamos la información inicial del objeto creado. Así como las primeras observaciones del componente assay que coincide con el objeto features.

```
## [1] "Primera visualización del contenido del objeto SummarizedExperiment:"
```

```
## class: SummarizedExperiment
## dim: 1541 45
## metadata(0):
## assays(1): counts
## rownames(1541): 443489 107754 ... 53297445 11954209
## rowData names(3): names PubChem KEGG
## colnames(45): b1 b10 ... c8 c9
## colData names(1): Treatment
```

```
##
## Clase del objeto de interés: SummarizedExperiment
```

```
## Dimensiones del objeto SummarizedExperiment (número de metabolitos x número de muestras):
```

```
## [1] 1541   45
```

```
## Primeras 6 filas de la matriz con las medidas de los metabólitos:
```

```
##          b1         b10        b11        b12        b13        b14
## 443489   " 941000"  " 757000"  " 612000"  " 858000"  " 185000"  " 671000"
## 107754   " 8300000" " 6790000" "20800000" "  320000" " 1290000" " 1580000"
## 9543071  "    1500" "     890" "16200000" "    1250" "     968" "     657"
## 11011465 " 276000"  "  35700"  " 631000"  " 369000"  " 242000"  " 472000"
## 5281160  "  706000" "  121000" "11600000" "  164000" "  424000" "  749000"
## 440341   "    6340" "   34100" "   31900" "    9440" "   92600" "    6740"
##          b15        b16        b17        b2         b4         b6
## 443489   "1140000"  " 108000"  " 383000"  " 593000"  "7240000"  " 494000"
## 107754   " 2340000" " 1180000" " 1260000" "15000000" "  495000" "   58100"
## 9543071  "     809" "     767" "     826" "    2810" "    1140" "    1010"
## 11011465 "5320000"  "  18000"  " 243000"  " 131000"  " 158000"  " 208000"
## 5281160  "  267000" " 3050000" "   99100" "  136000" "  452000" "   75600"
## 440341   "   14400" "    8180" "    8980" "    4610" "   10100" "    8180"
##          b7         b8         b9         a1         a10        a11
## 443489   " 812000"  "1290000"  "  66000"  " 215000"  " 310000"  " 798000"
## 107754   " 1350000" " 1860000" "  698000" " 1220000" " 6920000" "18700000"
## 9543071  "     635" "    1280" "     664" "     644" "    1060" "    1500"
## 11011465 " 228000"  " 119000"  "  58000"  "  17700"  " 394000"  "4230000"
## 5281160  "  132000" "  341000" "  119000" "   51600" "   54900" "26700000"
## 440341   "    5920" "    1950" "    1810" "    4350" "    1450" "       0"
##          a12        a13        a14        a15        a16        a17
## 443489   "1070000"  " 228000"  " 241000"  "1180000"  "  15100"  " 255000"
## 107754   " 1320000" " 1230000" " 1980000" " 6980000" "  716000" "  761000"
## 9543071  "       0" "     818" "     660" "     754" "     695" "     562"
## 11011465 "3740000"  " 361000"  "  63300"  "2090000"  "  37400"  "  13400"
## 5281160  "  323000" "  152000" "  208000" " 1400000" "   76100" "   16500"
```

```
## 440341   "   56200" "    3700" "    8360" "    2590" "    1390" "    1090"
##          a2         a4         a6         a7         a8         a9
## 443489   " 411000"  " 463000"  " 242000"  "1010000"  " 702000"  "  44600"
## 107754   " 2910000" "11300000" "  689000" " 1350000" " 1130000" "  479000"
## 9543071  "     851" "     766" "     637" "     846" "       0" "     618"
## 11011465 " 260000"  " 347000"  " 151000"  "1080000"  "   5080"  "   2140"
## 5281160  "  374000" "  491000" "   81200" " 1000000" " 7000000" "   22400"
## 440341   "   30400" "    2340" "    2930" "    7060" "    2830" "       0"
##          c1         c10        c11        c12        c13        c14
## 443489   " 136000"  "1060000"  "1050000"  " 464000"  "1460000"  " 636000"
## 107754   "  652000" " 2200000" " 7380000" "  187000" " 1430000" " 9730000"
## 9543071  "     546" "     926" "  800000" "       0" "       0" "     809"
## 11011465 " 266000"  " 627000"  "3140000"  " 127000"  " 197000"  " 286000"
## 5281160  " 1500000" "  171000" "  331000" "  198000" "  110000" "  178000"
## 440341   "       0" "    4460" "       0" "    3190" "   22900" "49200000"
##          c15        c16        c17        c2         c4         c6
## 443489   "4510000"  " 146000"  " 400000"  " 783000"  " 213000"  " 816000"
## 107754   "11200000" " 6660000" " 1830000" "15100000" "  971000" "  574000"
## 9543071  "    1380" "     982" "     625" "    1790" "     626" "     991"
## 11011465 " 545000"  "  35800"  "  23200"  " 230000"  "  59600"  "  48100"
## 5281160  "  791000" "   44100" "   57100" "  150000" "   29500" "  126000"
## 440341   "       0" "    6930" "    1730" "    2400" "    3450" "    2880"
##          c7         c8         c9
## 443489   " 587000"  " 319000"  " 102000"
## 107754   " 4590000" " 9730000" "  644000"
## 9543071  "    1600" "     949" "     756"
## 11011465 "  44000"  " 576000"  "  14200"
## 5281160  "  646000" "  291000" "   58900"
## 440341   "    2450" "   11200" "    3570"
```

```
## Metadatos de los metabolitos (rowData):
```

```
## DataFrame with 1541 rows and 3 columns
##                            names     PubChem        KEGG
##                      <character> <character> <character>
## 443489   10-Deacetyl-2-debenz..      443489      C11899
## 107754   1,1-Diethyl-2-hydrox..      107754      C13773
## 9543071  1,2-Dihydroxynaphtha..     9543071      C16196
## 11011465 12-trans-Hydroxy juv..    11011465      C16508
## 5281160   14-Dihydroxycornestin     5281160      C08483
## ...                          ...         ...         ...
## 25271619 2,5-Diamino-6-(5-pho..    25271619      C18910
## 46878395 4-(4-Deoxy-beta-D-gl..    46878395      C04733
## 5460130  5-Amino-4-chloro-2-(..     5460130      C04798
## 53297445 6-[2,3-Dihydroxy-1-(..    53297445      C19590
## 11954209 Flavanone 7-O-[alpha..    11954209      C15579
```

```
## Metadatos de las muestras (colData):
```

```
## DataFrame with 45 rows and 1 column
##       Treatment
##     <character>
## b1     Baseline
```

```
## b10    Baseline
## b11    Baseline
## b12    Baseline
## b13    Baseline
## ...         ...
## c4    Cranberry
## c6    Cranberry
## c7    Cranberry
## c8    Cranberry
## c9    Cranberry
```

Con este resumen exploratorio conocemos los datos de los que disponemos, comprobamos la correcta creación del objeto clase SummarizedExperiment y que está compuesto por las medidas de los metabolitos, los metadatos de cada muestra y el grupo al que pertenece, así como la clasificación o distintos nombres asociados a cada metabolito medido.

Con esta información introductoria podriamos proceder a realizar el análisis de interés según estudio.

Ejemplo de análisis

Supongamos que queremos comparar la cantidad de cierto metabolito entre los grupos experimentales (tipo de zumo y control) para ver si este experimenta cambios significativos.

Definimos como metabolito de interés el códificado con ID 439541

```
## [1] "El metabolito de interés es:"

## DataFrame with 1 row and 3 columns
##              names     PubChem        KEGG
##        <character> <character> <character>
## 439541  2-AminoAMP      439541      C01655

##                    Df    Sum Sq   Mean Sq F value Pr(>F)
## grupo_experimental  2 2.994e+12 1.497e+12   4.196 0.0218 *
## Residuals          42 1.498e+13 3.568e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = valores_metabolitos ~ grupo_experimental)
##
## $grupo_experimental
##                         diff        lwr       upr     p adj
## Baseline-Apple     133113.3 -396768.46  662995.1 0.8152972
## Cranberry-Apple    601440.0   71558.21 1131321.8 0.0228769
## Cranberry-Baseline 468326.7  -61555.13  998208.5 0.0925436
```

Como ejemplo hemos podido comparar el nivel del metabolito con ID PubChem 439541 llamado 2-AminoAMP entre el estado Basal y después de tomar zumo de arandanos y manzana.

Al realizar el test anova podemos afirmar que hay diferencias entre los grupos con un nivel de significación del 0.05 ya que se contrasta con el p-valor inferior de 0.02.

Al realizar la prueba Post Hoc del test de Tukey vemos que la diferencia significativa se encuentra entre los grupos de árandanos y manzana.

Por lo que para este metabolito no podriamos afirmar que haya diferencia estadística entre el estado basal y después de tomar el zumo, pero si que hay diferencias según si es de árandanos o manzana.

Claramente esto resultaría contrario, hay que matizar que el p-valor para el contraste de el estado basal y zumo de manzana es de 0.8 mientras que el basal y árandanos de una magnitud menos. En caso de cambiar el nivel de significación o hipoteticamente ampliar la n, podríamos encontrar diferencias entre las medidas Basal y Árandanos.