MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y BIOESTADÍSTICA1 - Análisis de Datos Ómicos PEC1

Lisseth Vergaray Del Aguila

2024-11-06

Contents

1 Contenidos

1.1 Objetivo

El objetivo de la práctica es realizar el análisis de un conjunto de datos de metabolómica siguiendo un workflow de trabajo bioinformático estándar. Este análisis debe incluir la selección y descarga de datos en repositorios especializados hasta la organización, exploración y presentación de los mismos en un repositorio que será accesible en GitHub.

1.2 Selección de Datos

Se ha seleccionado el data set 2023-CIMCBTutorial, que está relacionado con un estudio de cáncer gástrico, publicado en el artículo Chan et al.(2016), in the British Journal of Cancer.

Este estudio utiliza datos metabolómicos para analizar variaciones en el perfil metabólico de muestras, lo que es útil para entender como se ve afectado el meabolismo celular con el cáncer gástrico.

Se utilizó la técnica 1H-RMN (Resonancia Magnética Nuclear) para adquirir los espectros de las muestras, y se pretende identificar patrones metabólicos característicos en tejidos o fluidos de pacientes con cáncer, pudiendo obtener así marcadores que podrían ser útiles en el diagnóstico o pronósitico del cáncer.

1.3 Descarga de Datos

Procedemos a la descarga y leida de archivo para su manejo.

```
## # A tibble: 6 x 153
##     Idx SampleID SampleType Class    M1      M2    M3    M4    M5    M6    M7
##   <dbl> <chr>    <chr>      <chr> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1 sample_1 QC         QC     90.1   492.  203.   35   164.   19.7  41
## 2     2 sample_2 Sample     GC     43     526.  130.   NA   694.  114.   37.9
## 3     3 sample_3 Sample     BN    214.  10703.  105.   46.8 483.  152.  110.
## 4     4 sample_4 Sample     HE     31.6    59.7  86.4  14    88.6  10.3 170.
## 5     5 sample_5 Sample     GC     81.9   259.  315.    8.7 243.   18.4 349.
## 6     6 sample_6 Sample     BN    197.    128.  862.   18.7 200.    4.7  37.3
## # i 142 more variables: M8 <dbl>, M9 <dbl>, M10 <dbl>, M11 <dbl>, M12 <dbl>,
## #   M13 <dbl>, M14 <dbl>, M15 <dbl>, M16 <dbl>, M17 <dbl>, M18 <dbl>,
## #   M19 <dbl>, M20 <dbl>, M21 <dbl>, M22 <dbl>, M23 <dbl>, M24 <dbl>,
## #   M25 <dbl>, M26 <dbl>, M27 <dbl>, M28 <dbl>, M29 <dbl>, M30 <dbl>,
## #   M31 <dbl>, M32 <dbl>, M33 <dbl>, M34 <dbl>, M35 <dbl>, M36 <dbl>,
## #   M37 <dbl>, M38 <dbl>, M39 <dbl>, M40 <dbl>, M41 <dbl>, M42 <dbl>,
## #   M43 <dbl>, M44 <dbl>, M45 <dbl>, M46 <dbl>, M47 <dbl>, M48 <dbl>, ...
```

1.4 Descripción del Dataset

Este dataset incluye datos de concentración de metabolitos obtenidos mediante técnicas de espectrometría en diversas muestras. Los datos están organizados de la siguiente manera:

Columnas de las Muestras:

SampleID: Identificador único de cada muestra.

SampleType: Tipo de muestra, que puede ser “QC” para control de calidad o “Sample” para las muestras de análisis.

Class: Condición o grupo al que pertenece cada muestra. Las abreviaciones usadas incluyen:BN, GC, HE, QC

Metabolitos:

Los metabolitos están organizados en columnas (M1, M2, ..., M149) en la hoja de datos, y representan diferentes compuestos metabolómicos medidos en cada muestra.

Perc_missing: Porcentaje de valores faltantes para cada metabolito, lo que da una idea de la calidad de los datos para cada uno.

QC_RSD: Desviación estándar relativa en las muestras de control de calidad, que indica la variabilidad de cada metabolito en estas muestras y es un indicador de su consistencia.

1.5 Creación del Contenedor SummarizedExperiment

El *SummarizedExperiment* es una clase de objeto en R que nos permite almacenar y manejar datos ómicos de manera estructurada. Esta clase es ampliamente utilizada en análisis de datos de alto rendimiento, como los de genómica, proteómica y metabolómica, esta clase nos permite integrar tanto los datos cuantitativos como la información descriptiva de las muestras y las características en este caso los metabolitos. La clase fue diseñada para simplificar la manipulación y el análisis de estos datos, facilitando el trabajo con múltiples matrices de datos y metadatos.

Es útil porque nos ofrece una estructura integrada, combinando en un solo contenedor, la matriz de los datos y los metadatos de las filas y columnas. Al tener los datos y metadatos organizados en un solo contenedor, SummarizedExperiment permite realizar análisis de manera eficiente, pues se pueden acceder fácilmente a todas las muestras de un grupo experimental específico o a todas las concentraciones de un metabolito en particular sin necesidad de buscar en varias tablas.

También nos ofrece el poder trabajar con otros paquetes y herramientas bioninformáticas en R que están hechos para trabajar con este objeto.

1.5.1 Funcionamiento de SummarizedExperiment?

SummarizedExperiment organiza estos elementos de una manera que facilita su manipulación y análisis.Si queremos acceder a la matriz de datos principal se hace a través de assay(rse), o a los metadatos de filas rowData(rse), y finalmente a los metadatos de la columna con colData(rse).

Esto proporciona una coherencia en el análisis pues se asegura que los datos. Con todas estas características en su funcionamiento se pueden aplicar métodos estadísticos y visualización de gráficos basados en los datos.

1.5.2 Preparación de la matrices de datos y los metadatos

Para poder construir el objeto deberemos extraer la información, cogeremos los datos de la sheet Data para construir la matriz de datos y los datos de la sheet Peak para construir la matriz de metadatos.

Matriz de datos principal (data_matrix): Contiene los datos de expresión con 149 metabolitos medidos (variables) en las filas y 140 muestras en las columnas. Obtenemos esta matriz extrayendo todas las filas y las columnas 5 a 153 (que corresponden a los metabolitos M1-M149) de main_data. Luego, transponemos esta matriz para que los metabolitos queden en las filas y las muestras en las columnas.

Información de las columnas: Proporciona metadatos sobre las muestras, como tipo de muestra o clase.

Información de las filas: Contiene metadatos sobre las variables medidas, en este caso, los metabolitos.

Este enfoque nos permite estructurar los datos adecuadamente para utilizarlos en el objeto SummarizedExperiment, facilitando posteriores análisis bioinformáticos.

```
## DataFrame with 149 rows and 3 columns
##                       Label Perc_missing    QC_RSD
##                 <character>    <numeric> <numeric>
## M1        1_3-Dimethylurate    11.428571  32.20800
## M2   1_6-Anhydro- -D-gluc..     0.714286  31.17803
## M3     1_7-Dimethylxanthine     5.000000  34.99060
## M4     1-Methylnicotinamide     8.571429  12.80420
## M5           2-Aminoadipate     1.428571   9.37266
## ...                     ...          ...       ...
## M145                 uarm1     23.57143   41.4070
## M146                 uarm2      4.28571   34.4582
## M147              -Alanine      1.42857   27.6235
## M148      -Methylhistidine      1.42857   16.5619
## M149      -Methylhistidine      0.00000    8.3518
```

```
## DataFrame with 140 rows and 2 columns
##             SampleType       Class
##            <character> <character>
## sample_1            QC          QC
## sample_2        Sample          GC
## sample_3        Sample          BN
## sample_4        Sample          HE
## sample_5        Sample          GC
## ...                ...         ...
## sample_136          QC          QC
## sample_137      Sample          GC
## sample_138      Sample          BN
## sample_139      Sample          HE
## sample_140          QC          QC
```

1.5.3 Creación del Objeto SummarizedExperiment

Ahora procedemos a la creación de nuestro objeto, como se puede ver, se ha construido correctamente

Es es el resultado de construir nuestro objeto, junto con otros datos que ahora son fáciles de extraer.

```
# Mostramos el obejto y los metadatos del mismo.
rse
```

```
## class: SummarizedExperiment
## dim: 149 140
## metadata(0):
## assays(1): counts
## rownames(149): M1 M2 ... M148 M149
## rowData names(3): Label Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

```
dim(rse)
```

```
## [1] 149 140
```

```
head(assay(rse),1)
```

```
##    sample_1 sample_2 sample_3 sample_4 sample_5 sample_6 sample_7 sample_8
## M1     90.1       43    214.3     31.6     81.9    196.9     45.5       91
##    sample_9 sample_10 sample_11 sample_12 sample_13 sample_14 sample_15
## M1    480.6      62.2      36.5      93.5        NA      52.1        NA
##    sample_16 sample_17 sample_18 sample_19 sample_20 sample_21 sample_22
## M1        NA      62.5       0.9     154.3     216.3     420.3        NA
##    sample_23 sample_24 sample_25 sample_26 sample_27 sample_28 sample_29
## M1       0.5     410.8     133.2       355      80.2        NA     171.2
##    sample_30 sample_31 sample_32 sample_33 sample_34 sample_35 sample_36
## M1     135.8      33.6      78.2      19.2     148.5      68.5     190.1
##    sample_37 sample_38 sample_39 sample_40 sample_41 sample_42 sample_43
## M1     124.1       5.4      17.1        NA      51.2       4.1      77.9
##    sample_44 sample_45 sample_46 sample_47 sample_48 sample_49 sample_50
## M1        21      14.8        84      74.2        NA      21.4        NA
##    sample_51 sample_52 sample_53 sample_54 sample_55 sample_56 sample_57
## M1     487.5       1.1      21.8        53     114.7        NA     136.3
##    sample_58 sample_59 sample_60 sample_61 sample_62 sample_63 sample_64
## M1      58.2       5.5      98.6      47.1        NA      83.7     111.9
##    sample_65 sample_66 sample_67 sample_68 sample_69 sample_70 sample_71
## M1         6     104.9      14.4      26.6      69.4      48.5       9.8
##    sample_72 sample_73 sample_74 sample_75 sample_76 sample_77 sample_78
## M1        NA      66.6     192.5       8.3      34.1      58.5      37.6
##    sample_79 sample_80 sample_81 sample_82 sample_83 sample_84 sample_85
## M1      56.9      33.4        NA     123.7      19.6     148.1      46.1
##    sample_86 sample_87 sample_88 sample_89 sample_90 sample_91 sample_92
## M1     221.2       126     115.3      37.2        35     102.3       170
##    sample_93 sample_94 sample_95 sample_96 sample_97 sample_98 sample_99
## M1       7.9      20.3     273.7      41.5     138.2     909.9      49.1
##    sample_100 sample_101 sample_102 sample_103 sample_104 sample_105 sample_106
## M1       31.3       45.8      189.1       23.9         31      145.8      181.7
##    sample_107 sample_108 sample_109 sample_110 sample_111 sample_112 sample_113
## M1        0.7       31.9      123.9       25.9         NA         NA        0.9
##    sample_114 sample_115 sample_116 sample_117 sample_118 sample_119 sample_120
## M1       51.4       41.3       86.7      316.4      152.3      215.8      205.1
##    sample_121 sample_122 sample_123 sample_124 sample_125 sample_126 sample_127
## M1       48.1        7.1       27.2      113.4       50.8        8.6         NA
##    sample_128 sample_129 sample_130 sample_131 sample_132 sample_133 sample_134
## M1        0.4         NA      163.1       25.3       89.9       12.3      133.9
##    sample_135 sample_136 sample_137 sample_138 sample_139 sample_140
## M1        7.5       97.6      405.3       45.4       30.7       99.8
```

1.5.4 Añadiendo metadata del dataset al objeto SummarizedExperiment

Con la información proporcionada en el repositorio del dataset podemos añadir los metadatos a nuestro objeto. Creamos una lista almacenando la información mostrada en description.md.

```
## $content
##  [1] "Dataset used in the CIMBC tutorial on [\"Basic Metabolomics Data Analysis Workflow\"](https://
##  [2] ""
```

```
##  [3] "The tutorial describes the data as follows:"
##  [4] ""
##  [5] "- The study used in this tutorial has been previously published as an open access article Chan
##  [6] ""
##  [7] "- The deconvolved and annotated data file have been deposited at the Metabolomics Workbench da
##  [8] ""
##  [9] "- The data can be accessed directly via its project DOI:10.21228/M8B10B "
## [10] ""
## [11] "- 1H-NMR spectra were acquired at Canada's National High Field Nuclear Magnetic Resonance Centi
## [12] ""
## [13] "- Spectral deconvolution and metabolite annotation was performed using the Chenomx NMR Suite v
## [14] ""
## [15] "**Unfortunately, the Raw NMR data is unavailable.**"
##
## $title
## [1] "1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer"
##
## $description
## [1] "Data from a gastric cancer study in article Chan et al. (2016), in the British Journal of Cancer
##
## $project_DOI
## [1] "ID del proyecto PR000699."
##
## $pub_year
## [1] "2016"
```

1.6 Limpieza y normalización de los datos

1.6.1 Descartar NAs

Los datos metabolómicos contienen valores faltantes. Es importante decidir cómo manejarlos, en el tutorial seguido se toma como umbral el 20%, por lo tanto descartaremos los metabolitos que superen este umbral.

```
## class: SummarizedExperiment
## dim: 140 140
## metadata(5): content title description project_DOI pub_year
## assays(1): counts
## rownames(140): M1 M2 ... M148 M149
## rowData names(3): Label Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

Se puede confirmar que se han eliminado 9 filas que no cumplían con el umbral establecido.

1.6.2 Descartar metabolitos cuyo QC_RSD sea alto

A continuación, descartaremos aquellos metabolitos que presenten una puntuación de calidad (QC_RSD) superior a 20. Este criterio se basa en que la puntuación QC_RSD, que representa la desviación estándar relativa, indica qué tan lejos están los valores del metabolito con respecto a la media. Un valor alto de QC_RSD sugiere una mayor variabilidad, lo que podría afectar la fiabilidad de los datos para ese metabolito.

```
## class: SummarizedExperiment
```

```
## dim: 52 140
## metadata(5): content title description project_DOI pub_year
## assays(1): counts
## rownames(52): M4 M5 ... M148 M149
## rowData names(3): Label Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

52 filas (metabolitos) y 140 columnas (muestras), lo que indica que solo los metabolitos que cumplieron con ambos criterios (Perc_missing <= 20 y QC_RSD <= 20) permanecen en el dataset.

1.6.3 Normalización de los datos

Es común que los datos presenten variaciones significativas en las concentraciones de metabolitos debido a factores técnicos y biológicos. Estas variaciones pueden surgir de diferencias en el procesamiento de muestras, la sensibilidad de los instrumentos y la amplitud de los valores de concentración entre distintos metabolitos. Sin una normalización adecuada, estas variaciones pueden introducir sesgos y dificultar la interpretación de los resultados, afectando la comparación entre muestras y metabolitos.

Utilizaremos la transformación Logarítmica para reducir la dispersión y manejar datos con rangos amplios de valores. Esto es particularmente útil cuando existen metabolitos con concentraciones muy dispares.

```
## class: SummarizedExperiment
## dim: 52 140
## metadata(5): content title description project_DOI pub_year
## assays(2): counts normalized_counts
## rownames(52): M4 M5 ... M148 M149
## rowData names(3): Label Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

1.6.4 Guardando Contenedor

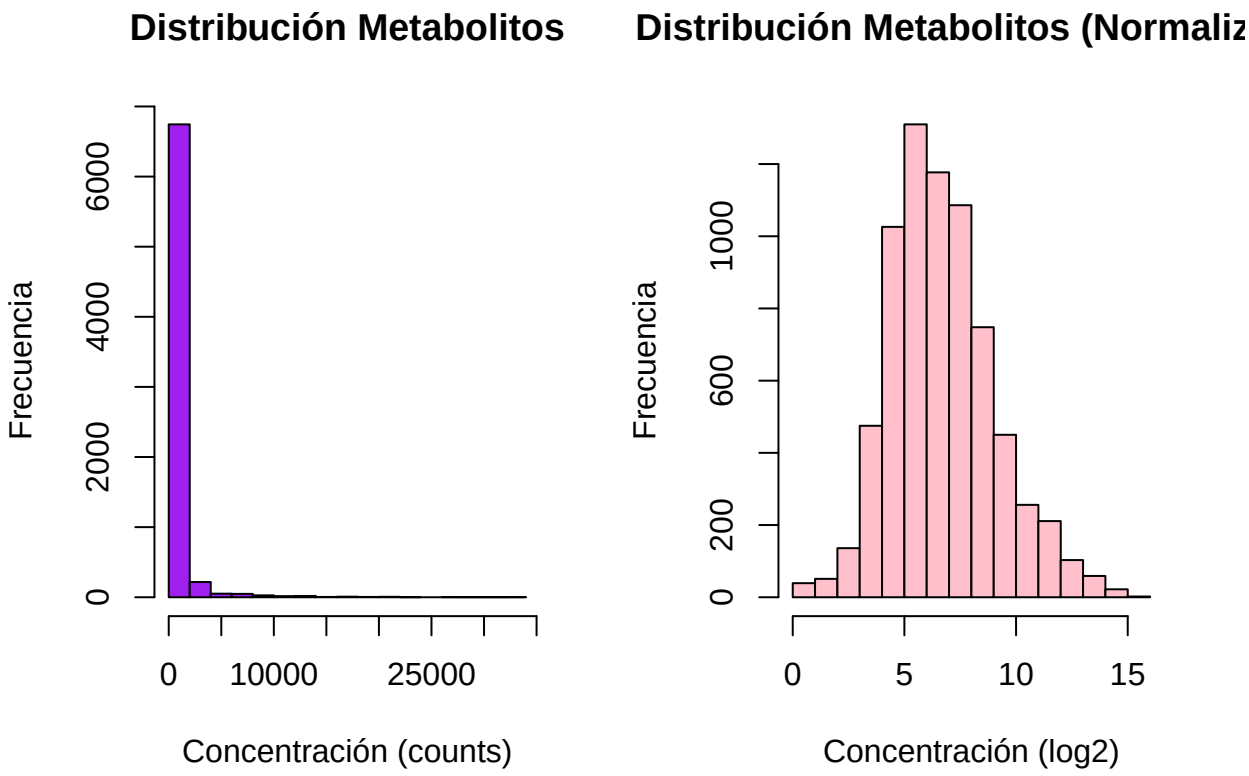
Una vez tenemos el objeto final (rse_final) generamos nuestro contenedor.

```
# Guardar el objeto SummarizedExperiment en un archivo .Rda
save(rse_final, file = "datos_metabolomica.Rda")
```

Tenemos ahora nuestro objeto con dos assays, uno sin normalizar y otro normalizado, se puede continuar con el análisis, ya que ahora tendremos únicamente en los metabolitos de alta calidad según los criterios estándar.

1.7 Análisis de los datos

Una vez los datos están limpios y normalizado podemos realizar un análisis básico, empezamos por comparar los datos sin normalizar con los normalizados.



La comparación de los dos histogramas muestra claramente que la normalización logarítmica ha mejorado la distribución de los datos de metabolitos, haciéndolos mucho más adecuados para el análisis estadístico y visualización.

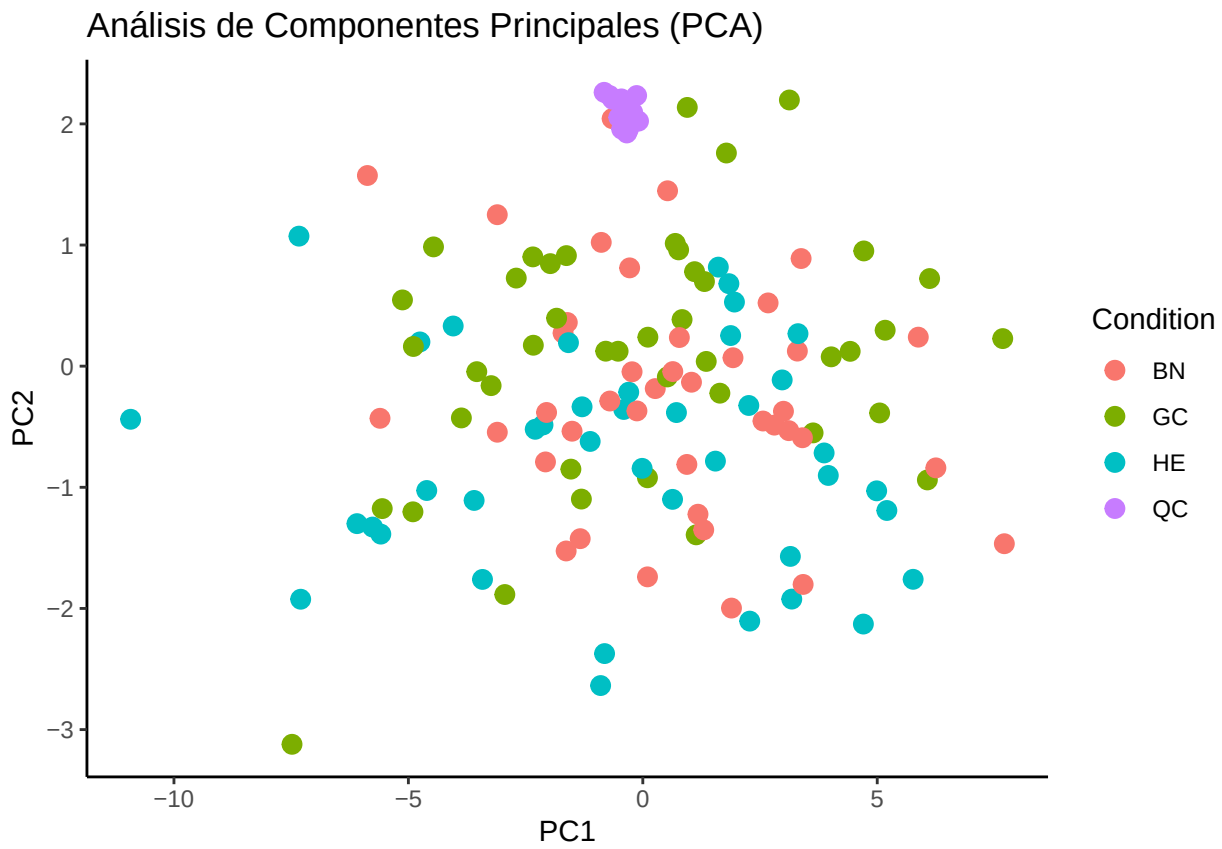
En el histograma de la izquierda, los datos originales muestran una gran concentración de valores en el extremo izquierdo del gráfico, con unos pocos valores extremadamente altos. Esta distribución sesgada sugiere que algunos metabolitos presentan concentraciones muy elevadas en comparación con otros, lo que genera un rango amplio y una distribución altamente asimétrica. Este tipo de distribución dificulta el análisis y puede influir negativamente en las comparaciones estadísticas, ya que los valores extremos pueden tener un efecto desproporcionado.

Tras la normalización logarítmica, la distribución de los datos en el histograma de la derecha es mucho más simétrica y se asemeja a una distribución normal, con la mayoría de los valores concentrados en un rango manejable.

1.7.1 Análisis de Componentes Principales (PCA)

El PCA es muy útil para reducir la dimensionalidad y observar patrones generales, como si las muestras se agrupan por condición.

```
## [1] 129
```



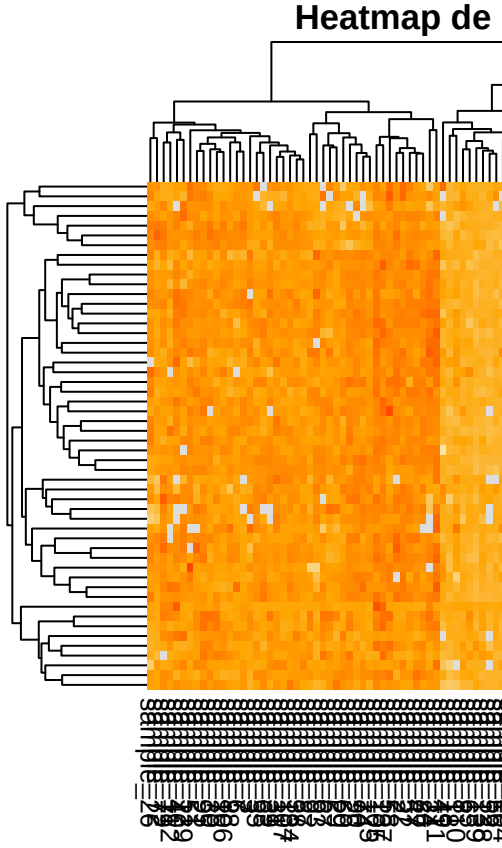
Observamos que los puntos morados (QC) están muy agrupados en la parte superior, lo que indica que las muestras de control de calidad (QC) son consistentes entre sí, esto sugiere que el control de calidad fue exitoso y que las medidas son reproducibles para estas muestras.

Las otras condiciones están más dispersas, lo que sugiere una mayor variabilidad metabólica entre las muestras en esas condiciones.

No hay una separación clara entre las condiciones BN, GC, y HE, lo que sugiere que sus perfiles metabólicos son similares en general o que las diferencias entre estos grupos no son suficientemente grandes como para que el PCA los separe claramente en el espacio de PC1 y PC2.

1.7.2 Clustered Heatmap de Metabolitos

Exploremos las relaciones entre metabolitos mediante un gráfico de correlación. Esto es útil para observar



posibles agrupaciones o relaciones que puedan existir entre diferentes metabolitos.
El agrupamiento de ciertos metabolitos también indica posibles conjuntos de metabolitos con comportamientos similares.

1.8 Resumen Final

Este estudio demuestra cómo los análisis de datos metabolómicos pueden proporcionar información valiosa sobre los perfiles metabólicos y las relaciones entre distintas condiciones experimentales. Los resultados obtenidos sugieren la existencia de patrones metabólicos consistentes en los controles de calidad y posibles subgrupos entre las muestras de distintas condiciones. Sin embargo, se requiere un análisis adicional para identificar de manera concluyente los metabolitos que podrían actuar como biomarcadores diferenciadores entre los estados de interés.

1.9 Repositorio GitHub

https://github.com/Lisscheese/Vergaray-DelAguila-Lisseth-PEC1

1.10 Referencias

1. SummarizedExperiment Manual. Disponible en: http://new.bioconductor.org/packages/release/bioc/manuals/SummarizedExperiment/man/SummarizedExperiment.pdf.

2. Tutorial de flujo de trabajo para metabolómica (MetabWorkflowTutorial). Disponible en: https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html.