Delgado-Navarro-Jose-Pablo-PEC1

Pablo Delgado

2025-03-27

Contents

Abstract

Este estudio explora el uso de muestras de Pap tests para la identificación de biomarcadores mediante análisis metabolómico, basado en el artículo Metabolomics of Papanicolaou Tests for the Discovery of Ovarian Cancer Biomarkers. Se utilizó el dataset ST003564 del Metabolomics Workbench, procesado e integrado en un objeto SummarizedExperiment en R. Se realizaron análisis exploratorios de las intensidades de metabolitos y su distribución por modo de ionización. Los resultados indican que las muestras de Pap contienen perfiles metabolómicos reproducibles, lo que respalda su potencial como herramienta no invasiva para la detección temprana de cáncer de ovario.

Introducción

El cáncer de ovario es una enfermedad con alta tasa de mortalidad debido a la ausencia de síntomas en etapas tempranas y la falta de biomarcadores específicos para su detección temprana. El estudio de Sah et., *al* investiga la posibilidad de utilizar muestras de Pap tests como fuente de biomarcadores, dado que ya se recolectan de manera rutinaria en controles ginecológicos.

Este informe reproduce el enfoque metabolómico de dicho estudio, empleando datos públicos del Metabolomics Workbench (ST003564) y utilizando herramientas bioinformáticas para evaluar la calidad de las muestras y la variabilidad de los metabolitos identificados.

Objetivos

General:

- Evaluar de manera independiente la viabilidad de utilizar datos metabolómicos de pruebas de Papanicolaou para la identificación de biomarcadores, mediante la replicación del estudio y el análisis de datos disponibles públicamente.

Específicos:

- Utilizar el dataset ST003564 y analizar su estructura mediante SummarizedExperiment.
- Evaluar la distribución de metabolitos y la variabilidad entre muestras mediante análisis exploratorios.
- Comparar los resultados con los hallazgos reportados en el artículo científico.

Métodos

Selección del Dataset

Se eligió el dataset ST003564 del Metabolomics Workbench por su alta calidad y relevancia en el contexto de la detección temprana de biomarcadores. Este conjunto de datos contiene mediciones metabolómicas obtenidas a partir de Pap tests, realizadas mediante UHPLC-MS en 29 muestras de pellets celulares residuales de pruebas de Papanicolaou de mujeres mayores de 50 años sin indicios de enfermedad. La disponibilidad de datos en modo de ionización positiva y negativa lo hace especialmente adecuado para evaluar la viabilidad de este enfoque en un contexto de salud pública.

Procesamiento en R

Se utilizó RStudio y herramientas de Bioconductor para el procesamiento de datos, siguiendo los siguientes pasos:

Importación y Preprocesamiento:

Los archivos de resultados en formato `.txt` fueron importados utilizando `read.table()`. La columna que combinaba el valor de m/z y el tiempo de retención (denominada `mz_rt`) se separó en dos variables distintas, `mz` y `rt`, empleando la función `separate()` del paquete `tidyr`.

Incorporación de Metadatos:

Se añadió una variable denominada `IonMode` a cada archivo, para identificar si los datos correspondían al modo de ionización positivo o negativo. Posteriormente, se generaron dos conjuntos de datos:

- Una matriz de intensidades que agrupa los datos numéricos (excluyendo las columnas `mz`, `rt` e `IonMode`).
- Objetos de metadatos:
  - **colData:** Contiene la información de las muestras, como el nombre de la muestra y su tipo (por ejemplo, “Blanco”, “Control de Calidad” o “Muestra”).
  - **rowData:** Incluye la información de cada característica (los metabolitos), con las variables `mz`, `rt` y `IonMode`.

Integración en un Objeto SummarizedExperiment:

Toda la información anterior se combinó en un objeto de la clase `SummarizedExperiment`, lo cual facilita el análisis integrado de datos y metadatos, permitiendo la realización de análisis exploratorios y multivariados de manera más eficiente. Posteriormente se incluyeron los metadatos relacionados al experimento y obtención de los datos a partir de la información disponible en la base de datos y en el artículo científico, por medio de la función `metadata(PEC1)`.

Una vez creado el objeto, este fué exportado en el archivo ***Delgado-Navarro-Jose-Pablo-PEC1.rda*** para el fácil acceso y manipulación por parte de terceros. Dicho archivo se encuentra disponible en el repositorio de la correspondiente PEC detallado en las referencias, junto con los metadata dataframe generados a partir del código ejecutado.

Herramientas Estadísticas y Análisis:

Se emplearon técnicas multivariadas y de visualización para evaluar el dataset:

- Histograma: Se generó un histograma de las intensidades (transformadas a escala logarítmica) para observar la distribución y detectar posibles necesidades de normalización.
- PCA: Se realizó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y visualizar patrones de agrupación entre muestras, eliminando controles (blancos y QC) para enfocarse en las muestras biológicas.
- Clustering: Se calculó una matriz de correlación (coeficiente de Spearman) y se generó un heatmap para identificar agrupamientos homogéneos y posibles outliers.

Resultados

Exploración de Datos

Tabla 1. Resumen del objecto PEC1 (`SummarizedExperiment`)

```
## class: SummarizedExperiment
## dim: 11051 40
## metadata(9): dataset_name source ... analysis_software comments
## assays(1): counts
## rownames: NULL
## rowData names(3): mz rt IonMode
## colnames(40): X6 X10 ... QC4 QC5
## colData names(2): SampleName SampleType
```

Tabla 2. Metadata del experimento

```
## $dataset_name
## [1] "ST003564 - Metabolomic analysis of Pap tests"
##
## $source
## [1] "https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST00356
##
## $description
## [1] "Análisis metabolómico de muestras de pruebas de Papanicolaou de mujeres mayores de 50 a
```

Tabla 3. Datos asociada a las columnas

```
## DataFrame with 4 rows and 2 columns
##      SampleName  SampleType
##     <character> <character>
## X6           X6     Muestra
## X10         X10     Muestra
## X14         X14     Muestra
## X18         X18     Muestra
```

Tabla 4. Datos asociada a las filas

```
## DataFrame with 4 rows and 3 columns
##                  mz        rt     IonMode
##           <numeric> <numeric> <character>
## [11048,]  1150.559     2.106    Negative
## [11049,]  1349.694     0.746    Negative
## [11050,]  1017.390     2.496    Negative
## [11051,]   436.052     2.022    Negative
```

Distribución de Intensidades de Peak Areas

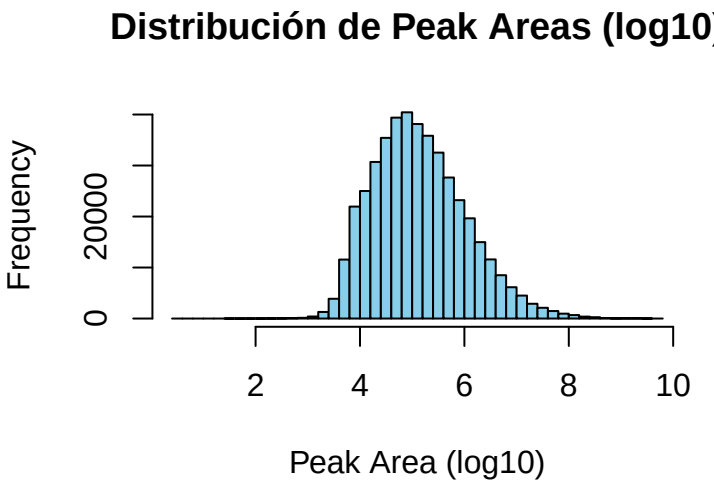

Figura 1. Histográma de distribución de muestras. El histograma muestra una distribución logarítmica, lo que sugiere que los datos metabolómicos siguen una distribución con tendencia normal.

Análisis PCA & Clustering de Muestras

- **Análisis PCA**

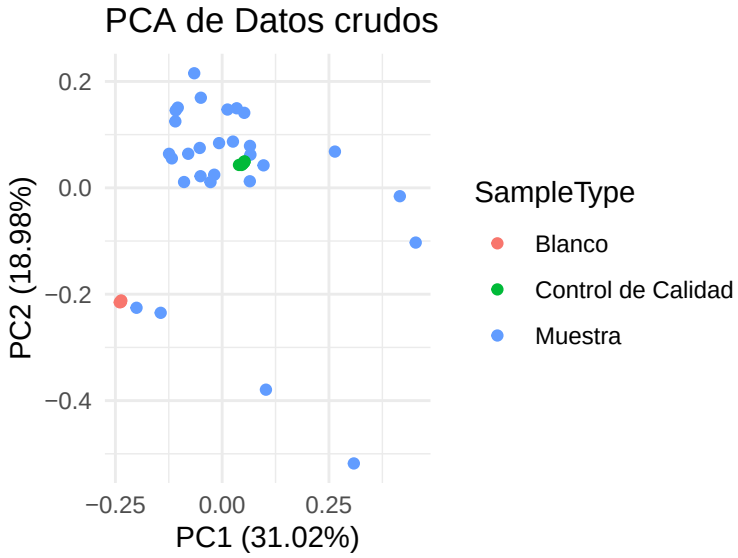

Figura 2. Análisis PCA. El PCA revela patrones de agrupación según el tipo de muestra, blancos y QC, por lo que idealmente para este propósito deberiamos trabajar solo con las muestras.

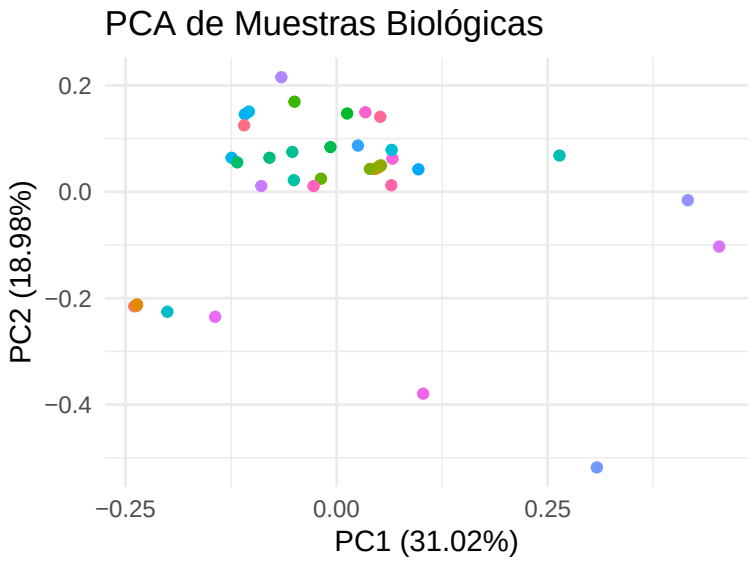

Figura 3. Análisis PCA con solamente muestras biológicas. El PCA revela patrones de agrupación de ciertas muestras biológicas, esto indica que comparten perfiles metabólicos similares.

Tabla 5. Resumen de los 3 primeros `Componente Principales`

```
##                             PC1      PC2      PC3
## Standard deviation     58.54817 45.79831 30.56836
## Proportion of Variance  0.31019  0.18980  0.08456
## Cumulative Proportion   0.31019  0.49999  0.58454
```

- **Clustering de Muestras**

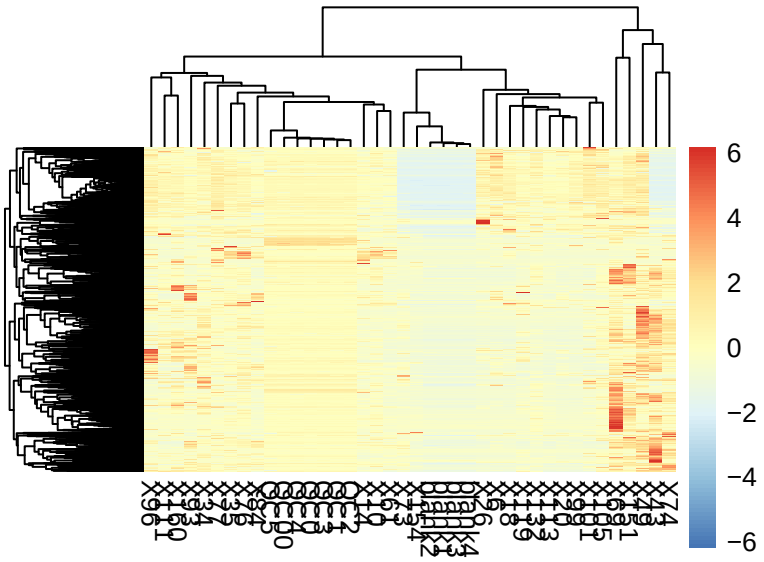

Figura 5. Heatmap de correlaciones entre muestras. En el heatmap se observan correlaciones entre distintas muestras (rojo más intenso), dichas agrupaciones indican posibles grupos homogéneos.

Discusión

Los resultados obtenidos en este análisis son consistentes con el estudio original de Sah *et al.*, en el cual se encontraron perfiles metabolómicos diferenciados en las muestras de Pap tests.

Análisis Exploratorio y Comparación de Métodos

El análisis exploratorio realizado sobre el dataset ST003564 permitió evaluar la distribución de intensidades, la variabilidad de las muestras y la coherencia de los datos de metabolómica en muestras de Pap tests.

Uno de los aspectos clave de este estudio fue el uso de la clase `SummarizedExperiment` en R para estructurar los datos de manera eficiente. Para comprender su relevancia, se comparó con `ExpressionSet`, una clase ampliamente utilizada en análisis transcriptómicos.

Diferencias entre SummarizedExperiment y ExpressionSet

Estructura y Flexibilidad

- `SummarizedExperiment` permite almacenar múltiples matrices de datos (por ejemplo, conteos sin procesar, datos normalizados, transformaciones logarítmicas), mientras que `ExpressionSet` está limitado a una única matriz de datos de expresión (Morgan *et al.*, 2020).

- `SummarizedExperiment` usa una estructura basada en `SimpleList` para manejar datos de diferentes formatos y tecnologías, mientras que `ExpressionSet` fue diseñado principalmente para datos de microarrays (Huber *et al.*, 2015).

Manejo de Metadatos

- En `SummarizedExperiment`, los metadatos de muestras se almacenan en `colData`, mientras que en `ExpressionSet` se encuentran en `phenoData` (Morgan *et al.*, 2020).

- `SummarizedExperiment` usa `rowData` para describir las características (genes, metabolitos), mientras que `ExpressionSet` emplea `featureData` (Huber *et al.*, 2015).

Compatibilidad y Uso en Bioconductor

- `SummarizedExperiment` es la opción recomendada para análisis modernos en Bioconductor, especialmente para datos de RNA-Seq, metabolómica y proteómica (Lawrence *et al.*, 2013).

- `ExpressionSet` sigue siendo utilizado en algunos análisis de microarrays, pero ha sido reemplazado en gran parte por `SummarizedExperiment` en flujos de trabajo recientes (Huber *et al.*, 2015).

Esta comparación evidencia que `SummarizedExperiment` ofrece una mayor versatilidad y capacidad de integración en estudios metabolómicos modernos, lo que lo convierte en la mejor opción para estructurar el dataset de este análisis.

Distribución de Intensidades de Peak Areas:

La alta dispersión observada en los valores de peak areas indica que algunos metabolitos son mucho más abundantes que otros, lo cual es típico en datos metabolómicos. Además nos dió un indicio visual de que los datos presentan una distribución con tendencia normal.

PCA & Clustering:

La separación y agrupación observada en el PCA y el Clustering presenta tendencias visuales de que las muestras comparten perfiles similares dentro de su grupo, lo cual indica que el dataset es consistente y adecuado para estudios de biomarcadores.

Comparación con el Artículo Original:

El estudio de referencia identificó ciertos lípidos clave como candidatos a biomarcadores. Aunque este análisis no se centra en la identificación de metabolitos específicos, los patrones observados en la distribución y agrupamiento de las muestras son similares a los reportados en el artículo.

Conclusiones

- Los datos metabolómicos de Pap tests presentan patrones de agrupamiento consistentes, lo que respalda su validez para estudios de biomarcadores.
- La estructura en SummarizedExperiment facilita la exploración y análisis de este tipo de dataset.
- Los resultados coinciden con hallazgos previos, lo que sugiere que el dataset puede ser útil en la detección de biomarcadores.

Referencias

- Github
  - El código, documentos y archivos utilizados para realizar esta PEC se encuentran en el repositorio: Pablo2996/Delgado-Navarro-Jose-Pablo-PEC1

Bioconductor. (2023). SummarizedExperiment: A container for high-throughput assays and associated meta-data. Bioconductor. Disponible en: https://bioconductor.org/packages/release/bioc/html/SummarizedExperiment.html

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., García, J. M., *et al.* (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 12(2), 115–121. https://doi.org/10.1038/nmeth.3252

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. (2013). Software for computing and annotating genomic ranges. PLoS Computational Biology, 9(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Metabolomics Workbench. (2023). Study ST003564: Metabolomics analysis of Pap test samples. Disponible en: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST003564

Sah, S.; Schwiebert, E.M.; Moore, S.G.; Liu, Y.; Gaul, D.A.; Boylan, K.L.M.; Skubitz, A.P.N.; Fernández, F.M. Metabolomics of Papanicolaou Tests for the Discovery of Ovarian Cancer Biomarkers.Metabolites 2024, 14, 600. https://doi.org/10.3390/metabo14110600

Anexos

El código crudo para la realización de este objeto de `SummarizedExperiment` se encuentra adjunto en el Github de la PEC como **Codigo_Delgado-Navarro-Jose-Pablo-PEC1**, junto con los archivos .txt necesarios para replicar su funcionalidad y detaller de los mismo.