Uso de R para análisis clínicos-transcriptómicos

Palabras clave: Análisis de datos clínicos, Expresión Diferencial, Cáncer de pulmón

Autores: Lauretta Paolo*, Formoso Karina*, Formoso Jesica, Medina Vanina

*Estos autores contribuyeron de manera equitativa a este trabajo.

Introducción

El cáncer de pulmón es la principal causa de muerte por cáncer a nivel mundial, siendo el carcinoma de pulmón de células no pequeñas el subtipo más frecuente. En la actualidad, la modulación del sistema inmune representa la alternativa terapéutica más eficaz. Sin embargo, a pesar de los avances terapéuticos, las tasas de supervivencia siguen siendo bajas. Evidencias recientes sugieren un papel relevante del sistema histaminérgico en la biología tumoral. En particular, se han detectado niveles elevados del receptor de histamina H3 (H3R/HRH3) en muestras de pacientes, asociados a una menor supervivencia.
En este trabajo presentamos un abordaje reproducible basado en R para integrar datos experimentales y clínico-transcriptómicos, con foco en el adenocarcinoma de pulmón y en el sistema histaminérgico.

Metodología

Desde cBioPortal —plataforma pública que integra datos oncológicos— y usando el paquete **cbioportalR** (Whiting, 2024) seleccionamos estudios de adenocarcinoma de pulmón, el estudio del consorcio OncoSG y Lung Cancer Consortium de Singapore publicado en Nature Genetics (2020) incluye 305 muestras de adenocarcinoma de pulmón de individuos de ascendencia del Este Asiático, con tumores–normales pareados, datos clínicos y datos de transcriptómica (RNA-seq, normalizado en z-scores RSEM). De este recurso empleamos específicamente la matriz de niveles de expresión y la tabla de datos clínicos (dimensión: 305 observaciones y 20 variables) para integrar expresión génica con variables y desenlaces clínicos.

En primer lugar, realizamos la limpieza y preparación de datos con el ecosistema **tidyverse** —en particular **dplyr** (Wickham H, 2025), **tidyr** (Wickham H, 2025), **tibble** (Müller, K., & Wickham, H. 2025) y **stringr** (Wickham H, 2023). A partir de las tablas crudas, seleccionamos las columnas de expresión, armonizamos identificadores, y eliminamos NA. Para evitar duplicaciones, removimos pacientes repetidos y símbolos génicos redundantes, colapsando valores por promedio; además filtramos columnas sin variación (sd = 0) o con datos insuficientes. Integramos la matriz de correlación con las variables clínicas, específicamente las poblaciones inmunes en el tejido tumoral, mediante su cuantificación relativa utilizando las IMsig (inmune signature), estas son un conjunto de firmas génicas inmunes derivados directamente de datos transcriptómicos, incluyen marcadores robustos para células inmunes. Estas firmas permiten estimar de forma relativa la abundancia de esas poblaciones. Tras evaluar supuestos de distribución (shapiro.test) y la morfología de las distribuciones (asimetría y curtosis) con **e1071** (Meyer D, 2024), estimamos las correlaciones de interés principalmente con el coeficiente de Spearman y ajustamos valores p por FDR (Benjamini–Hochberg). Construimos una matriz de correlaciones y la visualizamos con **ggplot2** (Wickham H, 2025) y etiquetado no superpuesto con **ggrepel** (Slowikowski K, 2024) (Figura 1), generando los scatterplots y guardando las figuras automáticamente, además usamos **stringr** (Wickham H, 2023) para limpiar y estandarizar los nombres de los archivos. Finalmente, exportamos diagnósticos y resultados y volcamos listas auxiliares (p. ej., genes HRH) a Excel con **openxlsx** (Schauberger P, 2023)

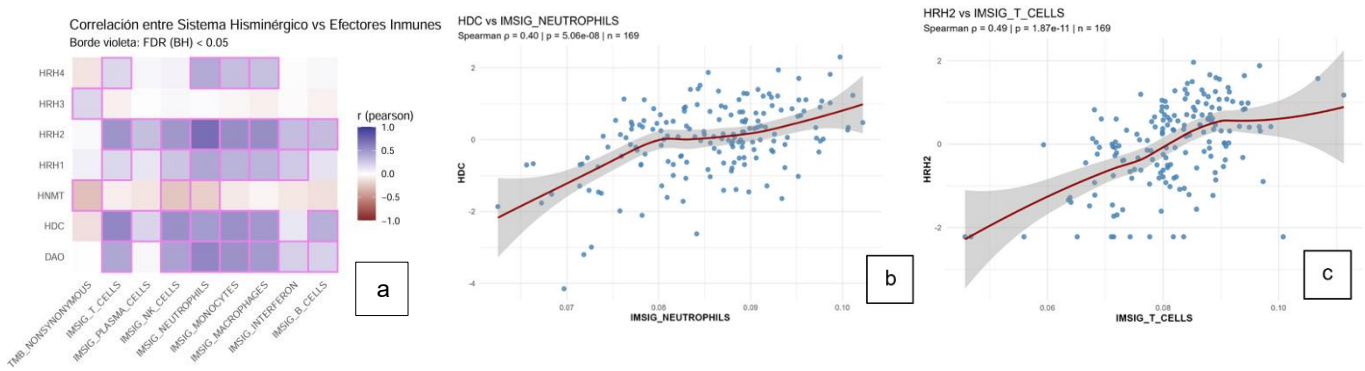

Figura 1: (a) Matriz de correlación de variables clínicas y expresión del sistema histaminérgico. (b) Eje X: Niveles de expresión de ARN de HDC* vs Eje Y: Abundancia relativa (IMsig) de neutrófilos en tejido tumoral (c) Eje X: Niveles de expresión de ARN de HRH2** vs Eje Y: Abundancia relativa (IMsig) linfocitos T en tejido tumoral. *Histidina descarboxilasa **Receptor de Histamina 2

Aunque detectamos correlaciones del sistema histaminérgico, la baja expresión tumoral de HRH3 en esta población del Este asiático nos llevó a evaluar otro estudio. Nuevamente usando el paquete cbioportalR seleccionamos un estudio que representa un conjunto de datos de pacientes relevados del "LUAD TCGA Pan-Cancer Atlas 2018", que nuclea información clínica y molecular de adenocarcinoma de pulmón proveniente de TCGA (The Cancer Genome Atlas), un consorcio del NCI/NHGRI con datos clínicos estandarizados y perfiles multi-ómicos. En este caso obtuvimos tres bases de datos, una de expresión génica (dimensión: 20531 observaciones y 512 variables), una de datos clínicos de los pacientes (dimensión: 566 observaciones y 38 variables), y una de datos clínicos de las muestras (dimensión: 566 observaciones y 19 variables). En este análisis se realizó una limpieza de datos como se explicó anteriormente. Luego estratificamos a los pacientes en dos grupos biológicos según su expresión de HRH3 (Grupo1: HRH3 ≤ −0.5; Grupo 2 > -0.5).

Para evaluar correlaciones de expresión génica y características clínicas utilizamos solo el Grupo 2 y variables clínicas definidas. Realizamos correlaciones de Spearman, y controlamos por comparaciones múltiples con FDR. Acto seguido, nos interesó evaluar la expresión diferencial de genes entre los pacientes del Grupo 1 y Grupo 2. Para ello realizamos análisis diferencial entre los grupos con el paquete limma (Ritchie Me, 2015) tras filtrar genes con varianza > 0 y ausencia de NA/Inf. Para comunicar resultados generamos un volcano plot con ggplot2 (Wickham H, 2025) y etiquetado no superpuesto con ggrepel, destacando genes up/down bajo criterios $|\log_2 FC| \geq 1$ y FDR < 0,05. Para mostrar solo los genes que mostraron estar modificados significativamente realizamos un heatmap con pheatmap (Kolde R, 2025), escalando por gen, donde la primera línea corresponde a nuestro gen de interés (HRH3) (Figura 2).

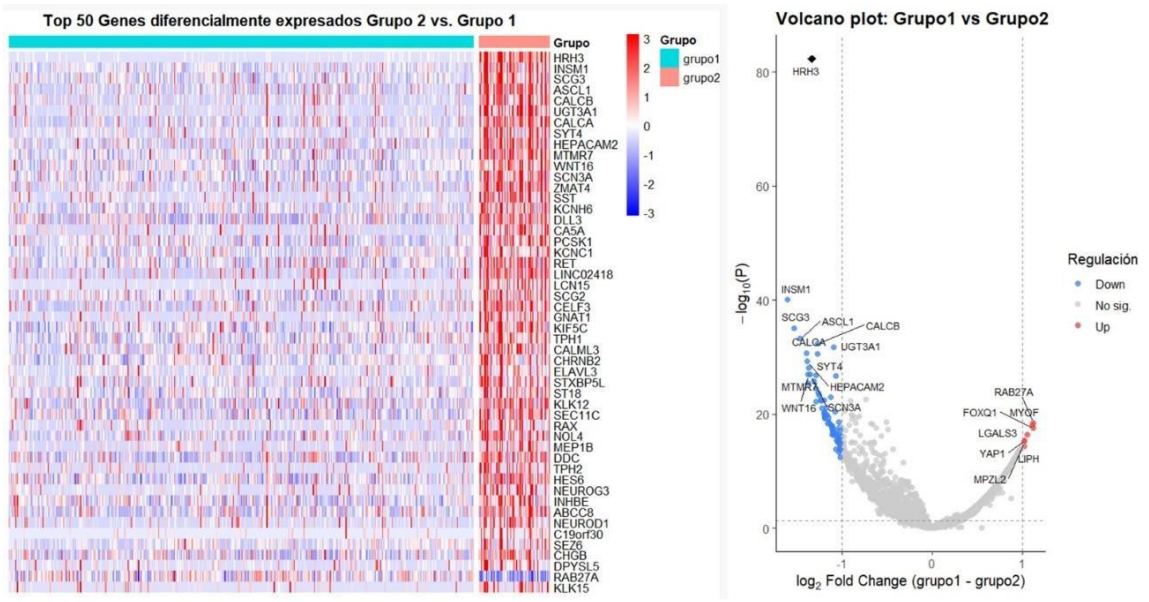

Figura 2: Heatmap con los genes diferencialmente expresados (Panel izquierdo). Volcano plot representando el log2foldchange vs. -log10(P).

Resultados

Estos hallazgos demuestran el rol significativo del sistema histaminérgico como un importante regulador de la respuesta inmune; por lo tanto, la regulación de este sistema representa un enfoque terapéutico prometedor para este tipo de neoplasias. Análisis traslacionales a gran escala como este, nos ofrecen información invaluable para comprender mejor las enfermedades y buscar nuevos tratamientos. A su vez, la capacidad de correlacionar gran cantidad de variables de forma automática brinda a la ciencia una herramienta esperanzadora para la búsqueda de nuevas soluciones.

Referencias:

Whiting, K (2024). _cbioportalR: Browse and Query Clinical and Genomic Data from: cBioPortal.
Wickham H, François R, Henry L, Müller K, Vaughan D (2025). dplyr: A Grammar of Data Manipulation. Wickham H, Vaughan D, Girlich M (2025). tidyr: Tidy Messy Data.
Müller, K., & Wickham, H. (2025). tibble: Simple Data Frames.
Wickham H (2023). stringr: Simple, Consistent Wrappers for Common String Operations.
Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies."
Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Slowikowski K (2024). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'.
Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2024). e1071.
Kolde R (2025). pheatmap: Pretty Heatmaps.
Schauberger P, Walker A (2023). openxlsx: Read, Write and Edit xlsx Files.