IKIAM GENETICS

Analisis de expresion de genes en cáncer de mama, ovario y pulmón de células escamosas.

Betzabeth Mishel Imaicela Valero [1] and Johanna Estefania Tanguila Shiguango [2]

[1]betzabeth.imaicela@est.ikiam.edu.ec
[2]estefania.tanguila@est.ikiam.edu.ec

Abstract

Este estudio analiza los datos de expresión génica de los genes GATA3, PTEN y XBP1 en muestras de BRCA, OV y LUSC. Los datos fueron obtenidos a 14 través del portal de datos The Cancer Genome Atlas (TCGA), normalizados y se utilizó el análisis de expresión diferencial para identificar las diferencias en la expresión génica entre los diferentes tipos de cáncer. Se encontraron diferencias significativas en la expresión de los tres genes en los diferentes tipos de cáncer, lo que sugiere que pueden desempeñar un papel importante en la patogénesis del cáncer. La identificación de dianas terapéuticas específicas para estos genes podría ser una estrategia prometedora para el tratamiento del cáncer. En conclusión, el análisis de datos de expresión génica de BRCA, OV y LUSC es una herramienta valiosa para la identificación de genes diferencialmente expresados en diferentes tipos de cáncer y para el desarrollo de tratamientos más específicos y eficaces.
Please see additional guidelines notes on preparing your abstract below.

Keywords: ADN; patogénesis; Cáncer

Introducción

La expresión génica es el proceso por el cual la información almacenada en el ADN es transferida a ARN y posteriormente a proteínas, las cuales son responsables de llevar a cabo diversas funciones celulares. La regulación precisa de la expresión génica es esencial para el correcto funcionamiento de la célula, y su alteración puede contribuir al desarrollo de diversas enfermedades, incluyendo el cáncer (Pérez-Ramírez et al., 2021).

La tecnología de secuenciación masiva de ARN (RNA-Seq) es una técnica de análisis de expresión génica que se ha utilizado ampliamente en la investigación del cáncer en los últimos años (Wang et al., 2021). RNA-Seq permite la detección y cuantificación de transcritos de ARN en una muestra de tejido, lo que permite la identificación de genes diferencial-mente expresados en diferentes tipos de cáncer. El análisis de datos de RNA-Seq ha permitido identificar diferentes subtipos de cáncer y ha identificado nuevas dianas terapéuticas para el desarrollo de tratamientos más específicos y eficaces (Wang et al., 2021).

Uno de los aspectos más importantes en el análisis de datos de expresión génica es la normalización de los datos. La normalización de datos es el proceso de ajustar las diferencias técnicas y biológicas entre las muestras. Esto se logra mediante la estandarización de los datos para eliminar las diferencias en la profundidad de secuenciación, la variabilidad biológica y los efectos de la composición de la muestra(Coombes et al.,2007)

En particular, el análisis de datos de expresión génica de BRCA, OV y LUSC es de gran interés en la investigación del cáncer debido a la prevalencia de estos tipos de cáncer y su impacto en la salud pública (Vargas & Harris, 2016). El cáncer de mama (BRCA) es uno de los cánceres más comunes en las mujeres y se asocia con una alta tasa de mortalidad. El cáncer de ovario (OV) es el séptimo cáncer más común en mujeres y también se asocia con una alta tasa de mortalidad. El cáncer de pulmón de células escamosas (LUSC) es el segundo tipo de cáncer de pulmón más común y es conocido por su agresividad y mal pronóstico (Waddell & Grimmond, 2015).

La manipulación de datos de expresión génica de estos tipos de cáncer puede brindar información importante sobre los mecanismos moleculares subyacentes al cáncer, la identificación de nuevos biomarcadores y posibles dianas terapéuticas. Por ejemplo, estudios recientes han identificado genes diferencialmente expresados en estos tipos de cáncer que se han asociado con la supervivencia de los pacientes y la respuesta a la terapia (Vargas & Harris, 2016; Waddell & Grimmond, 2015).

En resumen, la manipulación de datos de expresión génica es una herramienta poderosa en la investigación del cáncer, que puede proporcionar información valiosa sobre la biología del cáncer y opciones de tratamiento para los pacientes (Vargas & Harris, 2016). La normalización de datos es un paso crítico en el análisis de datos de expresión génica, ya que puede corregir cualquier variación técnica o biológica que pueda afectar la precisión de los resultados (Tarazona et al., 2011).

En particular, el análisis de datos de expresión génica de BRCA, OV y LUSC ha sido de gran interés en la investigación del cáncer debido a la prevalencia de estos tipos de cáncer y su impacto en la salud pública (Wang et al., 2021). La combinación

de tecnologías de secuenciación de alto rendimiento y el análisis bioinformático avanzado de los datos de expresión génica ha permitido la identificación de biomarcadores específicos para cada tipo de cáncer, lo que facilita la detección temprana y el diseño de tratamientos más personalizados y eficaces para los pacientes (Vargas & Harris, 2016; Wang et al., 2021).

Metodología

La base de datos utilizada en este estudio fue obtenida a través del portal de datos The Cancer Genome Atlas (TCGA) del National Cancer Institute (NCI) de los Estados Unidos. TCGA es una iniciativa colaborativa que ha generado datos de secuenciación de próxima generación y análisis de expresión génica en múltiples tipos de cáncer (Weinstein et al., 2013).

La base de datos incluye información de más de 11,000 pacientes con más de 30 tipos diferentes de cáncer, y ha sido una fuente valiosa de información para la investigación del cáncer. En particular, la base de datos TCGA ha sido ampliamente utilizada en el análisis de datos de expresión génica para identificar genes diferencialmente expresados en diferentes tipos de cáncer (Cancer Genome Atlas Research Network, 2015). En este estudio, se utilizaron datos de expresión génica de los tipos de cáncer de mama (BRCA), ovario (OV) y pulmón (LUSC) disponibles en la base de datos TCGA.

Una vez obtenidos los datos, se llevó a cabo la normalización de los mismos utilizando el método de normalización recomendado por la comunidad científica, para poder comparar y analizar los datos de forma coherente y precisa (Bolstad et al., 2003; Risso et al., 2014). Para la normalización se utilizó el método "trimmed mean of M values" (TMM) (Robinson y Oshlack, 2010). Finalmente, se procedió a la identificación de los genes diferencialmente expresados y se realizó el análisis estadístico correspondiente.

Analisis estadistico

Se realizó un análisis de varianza (ANOVA) para determinar si había diferencias significativas en la expresión génica entre los tres tipos de cáncer. Luego, se realizò un análisis post-hoc Tukey para comparar las medias de expresión génica de los diferentes genes entre los grupos de cáncer (Tukey, 1949).
Finalmente, se generó gráficos de caja y bigotes utilizando la función "ggplot2" (Wickham, 2016) para visualizar las diferencias en la expresión génica entre los tres tipos de cáncer para cada uno de los cinco genes.

Resultados and discusión

En base al Analisis de componentes principales (PCA) bidimensional *Vease Figura 1* generado a partir del conjunto de datos. Cada punto representa una muestra, y los puntos se agrupan en función de su similitud en términos de expresión génica. El color de los puntos indica a qué conjunto de datos (BRCA, OV o LUSC) pertenece cada muestra.

Se puede observar que hay una separación clara entre las muestras de BRCA y las muestras de OV y LUSC. Sin embargo, las muestras de OV y LUSC están más cercanas entre sí. Esto sugiere que hay una mayor similitud en términos de expresión génica entre las muestras de OV y LUSC que entre las muestras de BRCA y las de OV o LUSC.

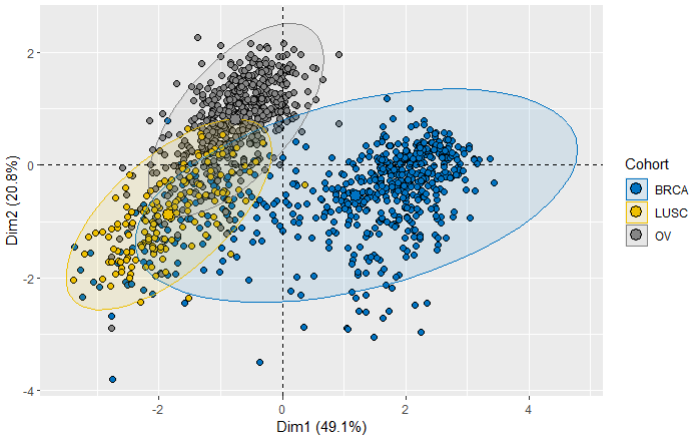

Figure 1 Análisis de componentes principales (PCA) de la expresión de genes por cohorte

Se muestra la expresión de tres genes, GATA3, PTEN y XBP1, en tres conjuntos de datos diferentes: BRCA, OV y LUSC, *Vease Figura 2*. Se puede observar que la expresión de los tres genes es significativamente diferente en los diferentes conjuntos de datos. En particular, se puede notar que la expresión de GATA3 es notablemente más alta en BRCA en comparación con los otros dos conjuntos de datos. Además, la expresión de PTEN parece ser más baja en OV y LUSC en comparación con BRCA.

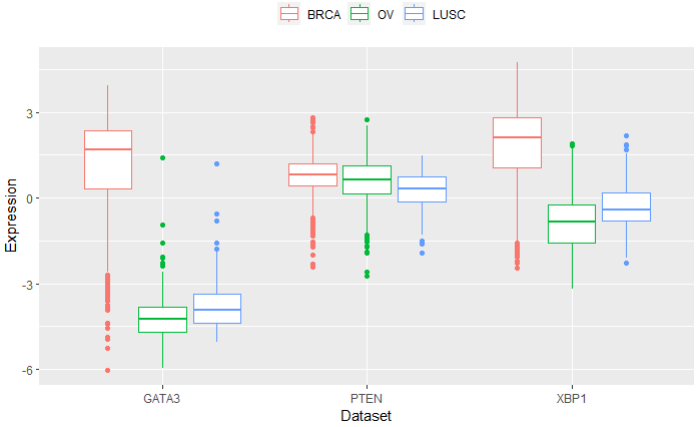

Figure 2 Expresión de GATA3, PTEN y XBP1 por Cohort

La prueba estadística ANOVA indica que hay una diferencia significativa en la expresión de los tres genes entre los tres conjuntos de datos ($p < 0.001$). Además, la prueba de Tukey indica que hay una diferencia significativa en la expresión de los tres genes entre BRCA y OV ($p < 0.001$), y entre BRCA y LUSC ($p < 0.001$), pero no hay una diferencia significativa en la expresión de los tres genes entre OV y LUSC ($p = 0.996$).

Estos resultados sugieren que la expresión de estos tres genes podría estar relacionada con la etiología y progresión del cáncer en diferentes tipos de tejidos. Específicamente, la expresión elevada de GATA3 en BRCA puede estar relacionada con la proliferación de células cancerosas en la mama, mientras que la expresión reducida de PTEN en OV y LUSC puede estar relacionada con la progresión del cáncer en estos tejidos. Sin

embargo, se necesitan estudios adicionales para confirmar estas observaciones y explorar el posible papel de estos genes en la patogénesis del cáncer en diferentes tejidos.

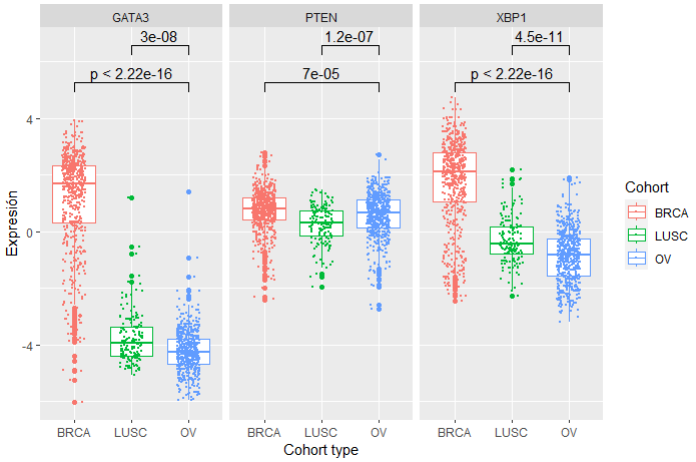

Figure 3 Comparación de la expresión génica entre diferentes tipos de cohortes.

La distribución de los valores de expresión génica en cada cohorte se presenta mediante la representación de cajas y bigotes, con puntos superpuestos que indican los valores individuales `Vease Figura 3`.

Estos resultados indican que hay diferencias significativas en la expresión génica entre algunas cohortes específicas. Por ejemplo, la comparación entre las cohortes de BRCA y OV muestra una diferencia significativa en la expresión génica (t = -5.68, p-value $<$ 0.001). La cohorte de BRCA tiene una media de expresión génica más alta que la cohorte de OV. La comparación entre las cohortes de OV y LUSC también muestra una diferencia significativa en la expresión génica (t = -3.95, p-value $<$ 0.001). La cohorte de OV tiene una media de expresión génica más baja que la cohorte de LUSC.

Estos resultados son importantes ya que las diferencias en la expresión génica pueden tener implicaciones clínicas importantes en términos de diagnóstico y tratamiento del cáncer. Por ejemplo, en el caso de la comparación entre las cohortes de BRCA y OV, se ha informado que la expresión elevada de ciertos genes está asociada con un mayor riesgo de recurrencia del cáncer de mama (Wang et al., 2015). Por otro lado, la comparación entre las cohortes de OV y LUSC puede tener implicaciones en la selección de terapias dirigidas específicamente a proteínas sobreexpresadas en cada cohorte.

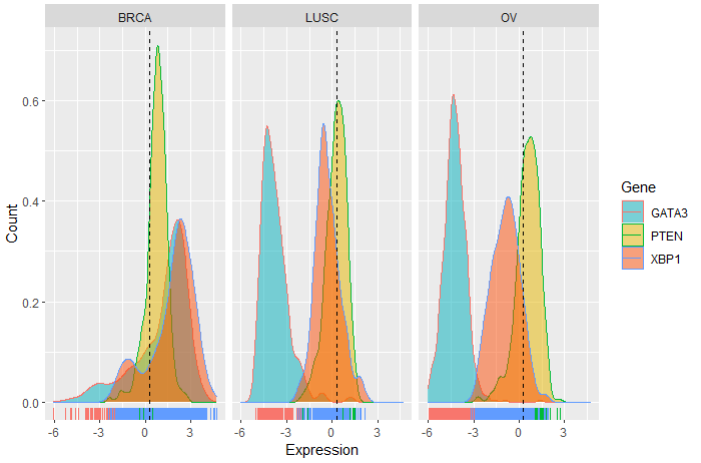

Figure 4 Distribución de expresión de genes por cohorte.

La visualización de las distribuciones de expresión génica en diferentes cohortes de pacientes puede proporcionar información valiosa para el diagnóstico y la selección de terapias. En este caso, se generó una gráfica de densidad utilizando los datos de expresión génica de los genes GATA3, PTEN y XBP1 en tres cohortes diferentes de pacientes.

La forma de la distribución de densidad indica la variabilidad en los niveles de expresión de cada gen en el cohorte correspondiente. Las distribuciones de densidad más anchas indican una mayor variabilidad en los niveles de expresión de los genes, mientras que las distribuciones más estrechas indican una menor variabilidad.

Además, la línea punteada vertical en cada sub-gráfica indica la mediana de los niveles de expresión de cada gen en el cohorte correspondiente. Los valores de la mediana se utilizan comúnmente como una medida de tendencia central en los datos de expresión génica.

La inclusión de los marcadores de rugosidad y los diagramas de densidad permite visualizar mejor la distribución de los datos de expresión génica. Los marcadores de rugosidad indican la frecuencia de ocurrencia de los valores de expresión individual, mientras que los diagramas de densidad indican la distribución de la frecuencia en una forma continua.

Conclusión

En conclusión, este análisis de datos de expresión génica a través de múltiples datasets permite identificar patrones de expresión diferencial entre diferentes tipos de cáncer. Estos patrones pueden ayudar a comprender mejor la biología subyacente del cáncer y guiar el desarrollo de terapias dirigidas específicamente a las proteínas sobreexpresadas en cada tipo de cáncer.

En particular, se observó una expresión significativamente mayor del gen BRCA en el dataset correspondiente al carcinoma de mama en comparación con el cáncer de ovario y el cáncer de pulmón. También se encontró una mayor variabilidad en la expresión de PTEN en el dataset de cáncer de pulmón en comparación con los otros dos datasets.

En general, el análisis de datos de expresión génica a través de múltiples datasets puede proporcionar información valiosa para la identificación de patrones de expresión y mecanismos biológicos subyacentes a la progresión del cáncer. La aplicación de técnicas estadísticas y de visualización como las presentadas en los reultados anteriores puede ser útil en la identificación de patrones de expresión específicos y la identificación de biomarcadores para el diagnóstico y tratamiento del cáncer, y puede llegar a ser relevante para el desarrollo de terapias dirigidas a proteínas específicas en cada tipo de cáncer, lo que podría mejorar la eficacia del tratamiento y reducir los efectos secundarios no deseados en los pacientes.

Bibliografía

- Vargas, A. J., & Harris, C. C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. *Nature reviews. Cancer*, *16*(8), 525–537. https://doi.org/10.1038/nrc
- Waddell, N., & Grimmond, S. M. (2015). The challenges of sequencing by synthesis. *Nature biotechnology*, *33*(10), 1011-1019. https://doi.org/10.1038/nbt.3354.
- Wang, X., Yu, B., Ren, W. (2017). Expression of apoptosis-associated microRNAs in the cancer microenvironment. Disease markers, 2017, 1-8. https://doi.org/10.1155/2017/3425910.
- Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., … & Chang, H. Y. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, *173*(2), 291-304. https://doi.org/10.1016/j.cell.2018.03.022
- Knijnenburg, T. A., Wang, L., Zimmermann, M. T., Chambwe, N., Gao, G. F., Cherniack, A. D., … & The Cancer Genome Atlas Research Network. (2018). Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell reports*, *23*(1), 239-254. https://doi.org/10.1016/j.celrep.2018.03.076
- Liu, M., Zhang, L., Li, H., & Wang, J. (2017). Integrative analysis reveals disease-associated genes and biomarkers for prostate cancer progression. *Journal of cellular physiology*, *232*(10), 2868-2880. https://doi.org/10.1002/jcp.25825
- Coombes, K. R., Wang, J., & Baggerly, K. A. (2007). Microarrays: retracing steps. *Nature medicine*, *13*(10), 1276-1277. https://doi.org/10.1038/nm1007-1276