Diseño de un flujo de investigación usando R/Bioconductor para el estudio del melanoma cutáneo

Palabras clave: RNASeq, R, Melanoma

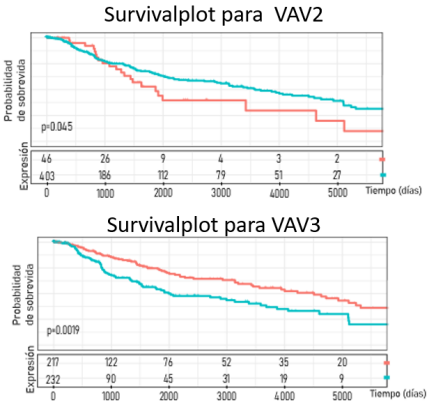
Abstract

El melanoma es la forma más severa de cáncer de piel, originado a partir de la transformación maligna de los melanocitos cutáneos. Con una incidencia global en aumento, es uno de los tipos de cáncer más metastásicos y con menores opciones de tratamiento.

Las Rho GTPasas son proteínas que controlan una amplia gama de procesos y cuya desregulación se asocia con fenómenos pro-oncogénicos. La activación de estas proteínas es controlada por los denominados “factores de intercambio de nucleótidos de guanina” (GEFs). La familia de proteínas Vav, son Rho GEFs que se compone de tres miembros: Vav1, Vav2 y Vav3. Datos previos de nuestro grupo demuestran que las proteínas Vav2 y Vav3 cumplen papeles importantes durante el la el desarrollo de cáncer de piel. De un modo sorprendente, los roles que estas proteínas desempeñan en melanoma no sólo parecen ser no-redundantes sino antagónicos. Nuestra hipótesis propone un rol supresor de tumor para Vav3 en melanoma, que se contrapone al papel pro-tumoral observado para Vav2.

Para realizar la búsqueda y descarga de datos de expresión génica del repositorio TCGA, se utilizó el paquete R/Bioconductor ‘TCGAbiolinks’ versión 3.17 (Colaprico et al. 2015, Silva et al. 2016 y Mounir et al. 2019). Los datos incluidos en el presente estudio provienen de muestras únicas de personas con tumores primarios de melanoma (n = 460). Antes de iniciar el análisis, se llevo a cabo una limpieza y se eliminaron los archivos de pacientes que no presentaron información clínico-patológica (sobrevida general, edad, sexo, etnia, estadio tumoral) asociados a la matriz de expresión genica compuesta de un total de 20531 genes por persona. Para la manipulación y limpieza de la matriz de expresión se utilizó el paquete ‘tidyverse’ (Wickman & Grolemud, 2018).

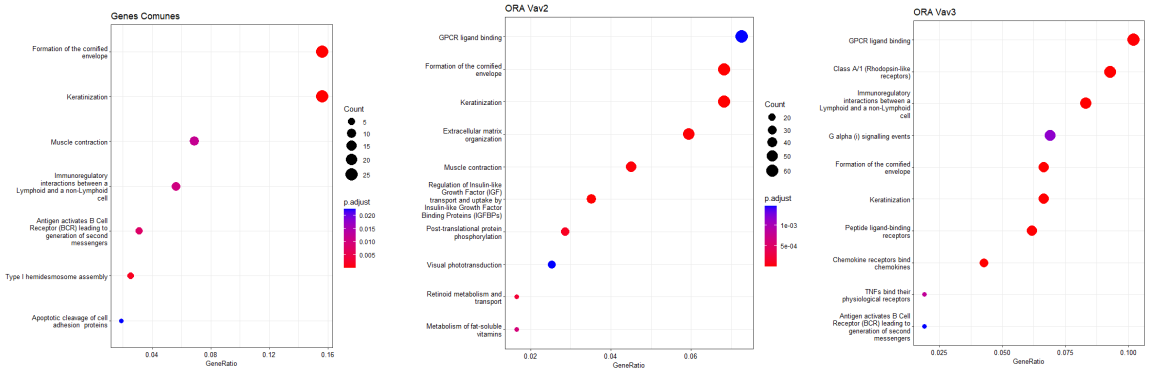
El análisis de los datos se realizó en el entorno R/Biocondutor versión 3.17 (Gentleman et al., 2004). Luego de la limpieza, las cuentas crudas (*raw counts*) fueron normalizadas mediante dos estrategias: cuentas por millón (CPM) y media recortada ponderada (TMM) según lo correspondiente a cada tipo de análisis. Ambas estrategias permiten comparar los valores de expresión y realizar el análisis de la expresión diferencial. Las CPM y TMM se calcularon a través de las funciones cpm y calcNormFactors del paquete ‘edgeR’ (Robinson & Oshlack, 2010).



Se utilizó un modelo univariado de riesgos proporcionales de Cox para analizar la relación entre la sobrevida y el nivel de expresión génica de cada proteína Vav. Este análisis consistió en ajustar un modelo de Cox por transcripto utilizando el paquete ‘survminer’ (Kassambara, 2016). Los valores de p se ajustaron mediante el método BH con un umbral del 5%. Utilizando la función surv_cutpoint se determinaron los valores de expresión génica de corte correspondientes a la expresión alta y baja para cada una de las proteínas Vav y se dividieron los pacientes en los siguientes grupos de comparación: personas con alta y baja expresión de Vav2, personas con alta y baja expresión de Vav3. Se utilizó la prueba de log-rank para comparar las curvas de supervivencia de Kaplan-Meier entre los grupos definidos de alta y baja expresión.

Diseño de un flujo de investigación usando R/Bioconductor para el estudio del melanoma cutáneo

El análisis de expresión diferencial entre pacientes con alta y baja expresión de Vav2 y Vav3 se llevó a cabo utilizando la función Exact Test del paquete EdgeR (Robinson et al., 2010; McCarthy et al., 2012 y Chen et al., 2016). Los DEGs para cada grupo de comparación se seleccionaron en función de valores de corte establecidos para su tasa de descubrimientos falsos (FDR) y el logaritmo del FC ("fold change"), los valores de corte fueron un FDR<0.01 y un logFC>1. Se compararon las listas de DEGs para cada par de confrontación alta/baja expresión mediante diagramas de Venn construidos con el paquete VennDiagram (Chen, 2011). De todos los DEGs, 37 genes fueron *upregulados* y 115 *downregulados* por ambas.



Con las listas de DEGs obtenidas mediante el análisis de expresión diferencial se realizó un análisis de sobrerrepresentación de genes (ORA) utilizando los paquetes 'reactomePA' versión v1.42.0 y 'enrichplot' (Yu & He, 2016 y Yu, 2023). Se generaron gráficos de tipo dotplot en los cuales se indican las vías más representadas por el set de genes dado (p<0.05, hasta 10 vías). En los gráficos hay tres referencias importantes: el generatio o proporción de genes de nuestro set en la vía; p-value de la significancia con un color característico según su valor; y el tamaño del punto en función del número de genes.

Aunque más estudios son necesarios, el presente abordaje permitió conocer los perfiles transcripcionales asociados a las altas y bajas expresiones de Vav2 y Vav3. Además se pudieron asociar los perfiles de DEGs de ambas Vav con procesos de queratinización, formación de capa córnea. Por otro lado, el perfil de Vav3 mostró asociación a vías de señalización relacionadas con procesos inmunitarios. Se asoció una mayor sobrevidda a valores altos de la expresión de Vav3 mientras que una peor sobrevida a los valores altos de Vav2. Esto coincide con nuestra hipótesis inicial.

El entorno de R resultó una herramienta muy potente para el acceso, limpieza, procesamiento e interpretación de datos. El pipeline generado en este trabajo puede servir de guía para el análisis de otros sets de datos.

El código utilizado, las salidas, el diseño y los gráficos se encuentran disponible en:https://github.com/ImoPupato/LatinR2023