UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE BIOTECNOLOGÍA

APLICACIÓN DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL Y MODELOS BAYESIANOS JERÁRQUICOS AL CASO DE ESTUDIO: GEN DE FUSIÓN RUNX1-JAK2

Trabajo de titulación previo a la obtención del título de Ingeniera Biotecnóloga

AUTORAS: TATIANA LIZBETH CHACÓN JIMBO
VANESSA YAMILÉ MOGROVEJO ARCENTALES
TUTORA: DRA. INÉS PATRICIA MALO CEVALLOS, Ph.D.

Cuenca - Ecuador

2023

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotras, Tatiana Lizbeth Chacón Jimbo con documento de identificación Nº 0106078041 y Vanessa Yamilé Mogrovejo con documento de identificación N° 0105774152; manifestamos que:

Somos las autoras y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Cuenca, 10 de octubre del 2023

Atentamente,

Tatiana Lizbeth Chacón Jimbo

0106078041

Vanessa Yamilé Mogrovejo Arcentales

0105774152

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotras, Tatiana Lizbeth Chacón Jimbo con documento de identificación Nº 0106078041 y Vanessa Yamilé Mogrovejo con documento de identificación N° 0105774152, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Trabajo experimental: "Aplicación del análisis de expresión diferencial y modelos bayesianos jerárquicos al caso de estudio: gen de fusión RUNX1-JAK2", mismo que se ha desarrollado para optar el título de: Ingeniera Biotecnóloga, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 10 de octubre del 2023

Atentamente,

Tatiana Lizbeth Chacón Jimbo

0106078041

Vanessa Yamilé Mogrovejo Arcentales

0105774152

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Inés Patricia Malo Cevallos con documento de identificación N° 0102291044, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: APLICACIÓN DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL Y MODELOS BAYESIANOS JERÁRQUICOS AL CASO DE ESTUDIO: GEN DE FUSIÓN RUNX1-JAK2, realizado por Tatiana Lizbeth Chacón Jimbo con documento de identificación N° 0106078041 y por Vanessa Yamilé Mogrovejo Arcentales con documento de identificación Nº 0105774152, obteniendo como resultado final el trabajo de titulación bajo la opción de Trabajo experimental que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 10 de octubre del 2023

Atentamente,

Dra. Inés Patricia Malo Cevallos, Ph.D.

0102291044

DEDICATORIA

Dedicatoria de Tatiana Lizbeth Chacón Jimbo

Dedicó este trabajo a mi madre, Priscila Jimbo, a mi padre, Patricio Chacón, a mi padrastro, Fernando Andrade, a mi abuelita, Paz Rodríguez, a mi tía, Rocío Jimbo, a mi hermana y hermanastro, Gaby y Lucho. Por ser ese apoyo incondicional en cada proceso académico. A mis perros, Peluche y Sofy, que han sido mis acompañantes en cada desvelada.

Dedicó a mi enamorado, Joseph Pesantez, por haberme dado la mano en cada momento de mi vida, en especial cada vez que me estresaba y me daba por vencida en algún ámbito académico. A la familia de mi enamorado, por ser ese apoyo en cada momento.

A mi tutora la Dr. Inés Malo, al Ing. Edmond Geráud y a todos mis profesores que me han formado académica y personalmente. Y a mi compañera de tesis que ha estado incondicionalmente en cada momento de mi carrera universitaria.

Dedicatoria de Vanessa Yamilé Mogrovejo

Dedico este trabajo de tesis a todas las personas que formaron parte especial de mi vida y trayectoria universitaria. En especial a mis padres y mi hermana, que a pesar de los malentendidos, siempre me apoyaron. A mi abuelita Blanca y abuelo Segundo, por ser el soporte ante cualquier situación. A mis tíos Alex y Mónica, que siempre me brindaron una mano y que me ayudaron a salir de varios problemas.

A mi perro Bruno, por su fidelidad, por la compañía en las tantas desveladas y por demostrarme su cariño incondicional sin necesidad de hablar.

A mi querida amiga Tatiana, con quien siempre conté en cada una de mis aventuras y también con la que establecí uno de los más lindos lazos de amistad, mi incondicional.

A la vida misma, que me demostró que la vida se compone de momentos felices y tristes, pero que, es ahí donde nosotros somos los únicos responsables de forjar nuestro camino y decidir si queremos darle un rumbo mejor.

AGRADECIMIENTO

Quiero agradecer a todas las personas que formaron parte de mi desarrollo académico, docentes, familiares y amigos. Agradezco a nuestra tutora de tesis, Dra. Inés Malo, por haberme apoyado tanto académicamente como personalmente, además, agradezco al Ing. Edmond Géraud por habernos apoyado con sus conocimientos en todo momento.

Expresó mi agradecimiento a mi familia en especial a mi mamá por haberme apoyado en toda mi formación académica, a mi abuelita por estar pendiente de mi en cada momento de mi carrera, a mi papá por ser un apoyo, a mi padrastro por sus consejos y mi hermana y hermanastro.

También agradezco a mi enamorado, Joseph Pesantez, por estar cada momento a mi lado, apoyándome en cada paso de mi carrera y a su familia por su apoyo. Por último, agradezco a mi compañera de tesis, Yamilé, por estar ahí a mi lado en cada momento y sobre todo ser una amiga de verdad, a mis amigos de ASU Ayudantías y a los integrantes del IEEE que nos han dado consejos y un ambiente confortante para realizar la tesis.

Tatiana Lizbeth Chacón Jimbo

Quiero agradecer de forma muy especial, a todas las personas que formaron parte significativa en el desarrollo de este trabajo de tesis, en especial a mi estimado tutor de tesis, Edmond Geráud, mismo que nos impulsó desde un inicio en aplicar un tema novedoso y que gracias a su empeño, conocimientos y paciencia constante, se permitió el éxito de este trabajo. A la Dr. Inés por su apoyo a lo largo de la carrera.

Además, deseo expresar mi agradecimiento a mi compañera de tesis, Tati, ya que no solo demostró ser una verdadera amiga, sino que también, me apoyó de forma moral. No quiero dejar de lado a mi familia, quienes fueron de bastante respaldo económico, a lo largo de este camino, en especial a mis padres, abuela Blanca y mis tíos Alex y Mónica, por aconsejarme,

ayudarme y apoyarme. Del mismo modo, expreso mi total gratitud a mi pareja, quien siempre estuvo en los buenos y malos momentos.

De igual manera, a todos los amigos con los cuales construí una amistad en la universidad, los chicos de ASU Ayudantías, a Miguel Samaniego, Matías Cuenca y Vinicio Ordoñez, que siempre me incluyeron en sus eventos y con quienes disfruté mucho de las experiencias culturales. Así como de los que forman parte de la oficina de IEEE, Joseph, Tracy, Japón, Paula y Guñi, por acogerme en sus instalaciones y levantar mi ánimo con sus risas y bromas, ante los momentos más difíciles, en especial a mis queridos amigos Sebastián Bedoya y Francisco Mendieta, quienes fueron los más graciosos, pacientes y leales. Y con todo el corazón, a mi "JE", Pedro, Brian y Klever, con ellos formamos una ONG y hemos participado de proyectos, talleres e incluso conocido personas de gran trayectoria profesional.

Yamilé Mogrovejo Arcentales

ÍNDICE DE CONTENIDO

DEDICATORIA V

Dedicatoria de Tatiana Lizbeth Chacón Jimbo V

Dedicatoria de Vanessa Yamilé Mogrovejo V

AGRADECIMIENTO VI

LISTA DE ABREVIATURAS XVII

RESUMEN XVIII

ABSTRACT XIX

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN 20

1.1. Introducción 20

1.2. Antecedentes 20

1.3. Planteamiento del problema 27

1.4. Justificación de la investigación 28

1.5. Limitaciones 29

1.6. Objetivos de la investigación 29

1.6.1. Objetivo general 29

1.6.2. Objetivos específicos 30

1.7. Hipótesis 30

CAPÍTULO II: MARCO TEÓRICO 31

2.1. Leucemia linfoblástica aguda precursora de células B (LLA-B) 31

2.1.1. Diagnóstico de LLA-B 31

2.1.2. Patogenia de la LLA-B: gen de fusión RUNX1-JAK2 31

2.2. Secuenciación del transcriptoma: RNA-seq 33
2.3. Expresión diferencial de los genes (DEG).....34
2.4. Lenguaje de programación R.....34
2.4.1. Definición de paquete (package) 35
2.4.2. Definición de función 35
2.5. Flujo de trabajo 35
2.5.1. Post-procesado de datos.....35
2.5.2. Lectura de datos 35
2.5.3. Filtrado de genes.....36
2.5.4. Normalización de datos.....36
2.5.5. Control de calidad previo al análisis de expresión diferencial 37
2.6. Paquete DESeq2.....37
2.7. Paquete BADER 38
CAPÍTULO III: MARCO METODOLÓGICO.....42
3.1. Descripción del diseño general 42
3.1.1. Diseño de investigación.....42
3.1.2. Nivel de investigación.....42
3.2. Población y obtención de muestras.....42
3.3. Variables 43
3.4. Recogida de datos 43
3.5. Softwares y paquetes de R usados para el análisis 43

3.5.1. Programa de R....43
3.5.2. Bioconductor....44
3.6. Flujo de trabajo de análisis de datos....45
3.6.1. Lectura de los datos....46
3.6.2. Preparación de los datos....46
3.6.3. Análisis de expresión diferencial con DESeq2....46
3.6.3.1. Normalización de los datos....46
3.6.3.2. Análisis de componentes principales (PCA)....47
3.6.3.3. Análisis de expresión diferencial....47
3.6.4. Análisis de expresión diferencial con BADER....48
3.6.5. Análisis de resultados....49
3.6.5.1. Volcano plot....49
3.6.5.2. GSEA....50
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN....51
4.1. Análisis de datos....51
4.2. Presentación de los datos....51
4.2.1. Resultados del análisis de componentes principales....51
4.2.2. Resultados del análisis de expresión diferencial con DESeq2....52
4.2.2.1. Resultado de genes sobreexpresados de DESeq2 con GSEA....55
4.2.2.2. Resultado de genes infraexpresados de DESeq2 con GSEA....57
4.2.3. Resultados del análisis de expresión diferencial con BADER....59

4.2.3.1. Resultados de genes sobreexpresados de BADER con GSEA 60

4.2.3.2. Resultados de genes infraexpresados de BADER con GSEA 62

4.3. Discusión..... 64

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES..... 68

5.1. Conclusiones..... 68

5.2. Recomendaciones 69

REFERENCIAS BIBLIOGRÁFICAS..... 70

APÉNDICES Y ANEXOS 79

ÍNDICE DE TABLAS

Tabla 1. Cuadro comparativo que emplea las metodologías de RNA-seq en base a los paquetes estadísticos de expresión diferencial de datos 25

Tabla 2. Número de muestras utilizadas 51

Tabla 3. DEGs sobreexpresados determinados con DESeq2 54

Tabla 4. DEGs infraexpresados determinados con DESeq2 55

Tabla 5. Nombre de los genes diferencialmente expresados obtenidos por BADER 60

ÍNDICE DE FIGURAS

Figura 1. Estructura del modelo bayesiano jerárquico de la actividad funcional en la mutación. 41

Figura 2. Logo de R 44

Figura 3. Logo de RStudio 44

Figura 4. Logo de Bioconductor 45

Figura 5. Flujo de trabajo con el cual se trabajó en el análisis DEG con los paquetes DESeq2 y BADER 45

Figura 6. Varianza de los componentes principales (PCA) 52

Figura 7. Gráfico de Scree Plot con los porcentajes de la proporción de las varianzas 52

Figura 8. Volcano Plot del análisis de expresión diferencial realizada con DESeq2 53

Figura 9. BarPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con DESeq2 56

Figura 10. CnetPlot de los genes sobreexpresados con sus vías de señalización determinado con DESeq2 57

Figura 11. BarPlot de los genes infraexpresados según p value y sus vías de señalización determinado con DESeq2 58

Figura 12. CnetPlot de los genes infraexpresados con sus vías de señalización determinado con DESeq2 58

Figura 13. Volcano Plot del análisis de expresión diferencial realizada con BADER 59

Figura 14. BarPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con BADER 61

Figura 15. CnetPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con BADER 62

Figura 16. BarPlot de los genes infraexpresados según p value y sus vías de señalización determinado con BADER 63

Figura 17. CnetPlot de los genes infraexpresados con sus vías de señalización determinado con BADER 64

ÍNDICE DE ECUACIONES

Ecuación 1. Teorema de Bayes 39

ÍNDICE DE ANEXOS

Anexo 1. Librerías utilizadas en R 79

Anexo 2. Código en R 80

LISTA DE ABREVIATURAS

ADME	Absorción, Distribución, Metabolismo y Excreción
RNA-seq	Secuenciación de ARN - *RNA-sequencing*
BADER	Análisis bayesiano de expresión diferencial en datos de RNA-seq - *Bayesian Analysis of Differential Expression in RNA-seq DATA*
DE	Expresión diferencial – *Differential expression*
DESeq2	Análisis de expression génica diferencial basado en la distribución binomial negativa- *Differential gene expression analysis based on the negative binomial distribution*
DGE	Genes diferencialmente expresados – *Differential gene expression*
FDR	Tasa de descubrimiento falso - *False Discovery Rate*
GSEA	Análisis de enriquecimiento de conjunto de genes - *Gene set enrichment analysis*
LLA-B	Leucemia Linfoblástica Aguda precursora de células B – *B-cell precursor acute lymphoblastic leukemia*
MCMC	Muestreo de Markov Chain Monte Carlo
NGS	Secuenciación de nueva generación - *Next Generation Sequencing*
PCA	Análisis de componentes principales - *Principal Component Analysis*
RNA	Ácido ribonucleico - *Ribonucleic acid*
RJ	RUNX1-JAK2
WT	Tipo salvaje – *Wild type*

RESUMEN

La leucemia linfoblástica aguda precursora de células B (LLA-B) es una enfermedad compleja con una patogénesis multifactorial que involucra varias vías de señalización y genes. Se sabe que la fusión del gen RUNX1-JAK2 está asociada con la LLA-B y puede influir en su variabilidad. Existen varios métodos para analizar la expresión génica, incluyendo DESeq2 y BADER, pero no está claro cuál de estos métodos proporciona los resultados más precisos o completos. Esta tesis se centra en el uso de DESeq2 y BADER para analizar la expresión génica en la LLA-B, con un enfoque particular en los genes y las vías de señalización asociados con la fusión del gen RUNX1-JAK2. Se llevaron a cabo análisis de expresión génica utilizando DESeq2 y BADER, seguidos de análisis de enriquecimiento de vías mediante GSEA. A través del análisis se descubrió que los métodos DESeq2 y BADER, producen diferentes resultados. DESeq2 encontró 60 genes con cambios significativos en su actividad, mientras que BADER encontró 23. Algunos de los genes y vías identificadas se relacionan con funciones críticas de las células como la señalización celular, el crecimiento y desarrollo, y la muerte celular programada, lo que sugiere que podrían desempeñar un papel en la LLA-B. Los resultados también sugieren que BADER podría ser útil para identificar procesos ocultos que podrían contribuir a la enfermedad. Esto abre la posibilidad de futuras investigaciones para entender mejor la complejidad de la LLA-B y utilizar múltiples métodos de análisis de expresión diferencial para obtener una visión más completa de los cambios en la expresión génica.

Palabras claves: Expresión diferencial, DESeq2, BADER, modelos bayesianos jerárquicos.

ABSTRACT

B-cell precursor acute lymphoblastic leukemia (B-ALL) is a complex disease with a multifactorial pathogenesis involving several signaling pathways and genes. The RUNX1-JAK2 gene fusion is known to be associated with B-ALL and may influence its variability. Several methods exist to analyze gene expression, including DESeq2 and BADER, but it is unclear which of these methods provides the most accurate or complete results. This thesis focuses on the use of DESeq2 and BADER to analyze gene expression in B-ALL, with a particular focus on genes and signaling pathways associated with the RUNX1-JAK2 gene fusion. Gene expression analyses were performed using DESeq2 and BADER, followed by pathway enrichment analyses using GSEA. Through analysis, it was found that the DESeq2 and BADER methods produced different results. DESeq2 found 60 genes with significant changes in their activity, while BADER found 22. Some of the genes and pathways identified relate to critical cell functions, such as cell signaling, growth and development, and programmed cell death, suggesting that they may play a role in B-ALL. The results also suggest that BADER could be useful in identifying hidden processes that could contribute to the disease. This opens the possibility for future research to better understand the complexity of B-ALL and to use multiple methods of differential expression analysis to gain a more complete picture of changes in gene expression.

Keywords: Differential expression, DESeq2, BADER, hierarchical Bayesian models.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1. Introducción

En los últimos años, el estudio del cáncer ha presentado dificultades por sus microambientes complejos y heterogeneidades celulares, en este contexto se han desarrollado varias técnicas de secuenciación como la secuenciación del RNA (RNA-seq) que ofrece información para el tratamiento del cáncer, los procesos involucrados, las vías de señalización y los biomarcadores (Guo et al., 2021; Hong et al., 2020). Por lo que, uno de los casos de estudio en la genómica es el gen de fusión RUNX1-JAK2 que se asocia a la leucemia linfoblástica aguda.

1.2. Antecedentes

De acuerdo con el estudio de Fortschegger et al. (2021), la leucemia linfoblástica aguda precursora de células B (LLA-B) es una forma de cáncer pediátrico común y genéticamente heterogénea que afecta a los receptores de citoquinas o quinasas como ABL1, ABL2, PDGFRB, CSF1R, JAK2, EPOR y CRLF2. Estas proteínas de fusión pueden generar una mutación oncogénica de doble efecto al inducir las vías de señalización proliferativas y actuar con otro socio de fusión, el factor de transcripción de desarrollo, que bloquea la diferenciación. En este caso se fusionan los genes RUNX1-JAK2, creando un gen integrado, que activa dos vías de señalización importantes: JAK-STAT y MYC. De esta forma se podría comprender cómo actúan las neoplasias hematopoyéticas que implican la fusión de RUNX1-JAK2, que representa al menos el 20% de todos los casos de LLA-B.

En este contexto, apoyándose de las bases moleculares, se introduce a la expresión diferencial de genes, como una herramienta esencial para medir la calidad y cantidad de la expresión génica en base a condiciones experimentales específicas mediante técnicas como microarrays y la más idónea, RNA-seq, que ha reemplazado a la secuenciación de Sanger y que ahora solo sirve como un método de comprobación de la técnica mencionada. Está última

permite medir la manifestación de genes individuales en determinadas condiciones, en comparación con las técnicas de nueva generación que dan la apertura del análisis de varios fragmentos, con su masiva secuenciación adaptada a plataformas especializadas (Miao et al., 2021; Sundaram et al. 2017). Debido a la ingente cantidad de datos que se obtienen mediante las tecnologías de secuenciación de nueva generación (NGS, por sus siglas en inglés), se requiere una aproximación computacional y estadística (Canzoneri et al., 2019). En este trabajo se utiliza una aproximación bioestadística, puesto que se cuentan ya con los datos procesados.

Los datos, en primera instancia, se obtienen del secuenciador en formato *fastq*, el cual es un formato en el que se almacenan secuencias nucleicas de pocas pares de bases, alrededor de 70-150 (Canzoneri et al., 2019; Masip, 2019). Por lo tanto, al secuenciar el transcriptoma humano, teniendo en cuenta, que consta de más de 140000 transcritos, cada lectura puede no encajar o encajar más de una vez en un transcrito (Soriano, 2023; Valdespino-Gómez, 2013). El hecho de encajar las lecturas en un genoma de referencia se le conoce como el proceso del alineamiento. Una vez dicho proceso se ha realizado, se procede a obtener una matriz de números enteros. Los números enteros nos indican cuántas veces un transcrito ha sido alineado, a esto se le conoce como la matriz de contajes (Díez, 2022; Soriano, 2023). De la cual se parte el siguiente trabajo.

Con tal analizar una matriz que contiene todo el transcriptoma, un abordaje bioestadístico es necesario. Para realizar dicho análisis se utilizó el software estadístico R, y el repositorio de librerías para la Bioinformática conocido como *Bioconductor*. El lenguaje de programación R, se basa en librerías, las cuales son un conjunto de funciones específicas para un análisis en específico (Masip, 2019).

Ahora bien, el presente proyecto, intenta comparar el método de análisis de expresión diferencial mediante DESeq2 y BADER. Cabe recalcar que existen diversas librerías estándar para dicho análisis como Limma o EdgeR. No obstante, la literatura es consistente en que la

librería que mejor se adapta a este tipo de datos es DESeq2 (Schurch et al., 2016). Sin embargo, recientemente, en el año en curso, se publicó un nuevo paquete llamado BADER (Neudecker y Katzfuss, 2023). Dicho paquete aproxima el problema de una forma distinta la cual se discutirá más adelante en el presente documento.

Dado el hecho que los datos transcriptómicos poseen características especiales en el sentido de que no son continuos, sino más bien discretos, el principal problema radica en determinar la distribución probabilista subyacente de estos datos. Es decir, cuando se realiza una medición de algún experimento o se recopilan datos, dichas mediciones normalmente provienen de una distribución conocida. Esta distribución se puede obtener mediante la función acumulativa de distribución, que toma cualquier valor y lo transforma en una probabilidad, dando una forma específica de distribución con ciertas propiedades que explican los datos observados. Por lo tanto, cada metodología como Limma-voom, DESeq2, BADER, EdgeR, etc., intentan explicar la verdadera procedencia de los datos (Amezquita, et al., 2020; Liu et al, 2021; Schurch, 2016).

En realidad, los paquetes más populares, como DESeq2, EdgeR y Limma, emplean una aproximación bayesiana. El teorema de bayes se basa en creencias, lo que significa que si consideramos que existe una cierta realidad explicada por los datos, conocida como "posterior", esta creencia o realidad es proporcional a los datos, los cuales explican en cierta parte la realidad o creencia. A esto se le conoce como la "verosimilitud de los datos" (Liu et al., 2021; Ortiz, 2018). Sin embargo, la probabilidad o verosimilitud por sí sola no es suficiente para explicar la creencia o realidad. Es necesario multiplicarla por la realidad que se quiere observar. En otras palabras, la multiplicación de la realidad en el numerador por la verosimilitud de que los datos realmente expliquen la creencia representa cómo las suposiciones de la verdad influyen en la plausibilidad en la actualización del conocimiento

sobre el hecho que se investiga. Existe un término constante el cual divide a la fórmula que simplemente asegura que la salida del algoritmo sea una probabilidad (Ortiz, 2018).

Uno de los principales desafíos de los métodos bayesianos radica en la suposición de la realidad que actualizará el conocimiento sobre la plausibilidad de los datos al explicar la creencia que se quiere observar. Dicha realidad generalmente se asume que sigue una distribución específica, la cual es desconocida. En la metodología bayesiana, se postula que esta realidad sigue una distribución particular y se actualiza hasta alcanzar un punto de convergencia (Díez, 2022). Sin embargo, en los métodos más populares, la creencia o realidad sobre la expresión diferencial de los genes se deriva directamente de los datos en sí. Esto supone un cierto sesgo, ya que depende en gran medida de los propios datos. En otras palabras, la interpretación de los datos depende tanto de los datos mismos como del contexto en el que se están analizando (Limma et al., 2014; Ortiz; 2018; Robinson et al., 2010).

Los métodos mencionados utilizan la aproximación bayesiana por un problema que se plantean con los contajes. Al graficar el histograma o la densidad de expresión de todos los genes en una muestra, se puede observar una distribución muy parecida a la de Poisson. Dicha distribución nos dice el número de cuentas que existen hasta que hay el evento que queremos observar. Sin embargo, existe lo que se llama una sobredispersión de los datos. En términos técnicos, esto significa que existe un problema de sobredispersión. Esto conlleva a que la varianza sea mayor a la media, contrariamente a lo que esperaríamos en una distribución de Poisson, en el que la tasa de descubrimiento es igual a la media la varianza. En un gráfico, ya no es decreciente y la varianza ya no es igual a la media, es mayor, dicha sobredispersión de los datos se pueden notar como un bulto en el histograma, en lugar de ser decreciente (Chamorro, 2019; Ortiz, 2018).

Por lo que, un enfoque bayesiano aborda este problema de sobredispersión, cada uno de una manera ligeramente diferente. Por ejemplo, en una distribución normal, los parámetros

que la rigen son la media y la varianza, al igual que en muchas distribuciones conocidas (Universidad de las Palmas de Gran Canaria, s.f.). Sin embargo, los datos de RNA-seq, se debe de considerar a parte el parámetro de la sobredispersión, siendo así que los métodos difieren en cómo tratarla. No obstante, al tratar millones de lecturas que se han alineado a diversos transcritos a lo largo del exoma, las observaciones que se han mencionado anteriormente, es necesario transformarlas a una escala logarítmica. De hecho dicha transformación es esencial para entender el concepto de la significancia biológica. Puesto al ser un análisis diferencial se observan las diferencias en forma logarítmica, por lo que en realidad, al transformar dichas diferencias de nuevo al exponente original, no se observará una diferencia entre medias sino una razón entre un caso y un control. Es decir, se observará, cuánto un transcrito estará expresándose sobre otro transcrito, por eso en los resultados del análisis de expresión diferencial, al final se observa una sobreexpresión o infraexpresión respecto el mismo transcrito en una condición (Chamorro, 2019).

La relación de la media con la varianza no es directamente proporcional, es más es mayor la varianza a la media, y esto se observa cuando se computa una gráfica de los contajes en forma logarítmica, se computa la media y en el eje de las ordenadas la desviación estándar la varianza (Stupnikov et al., 2021). Este fenómeno es la dispersión así pues los diferentes paquetes lo tratarán de forma distinta pero similar. Similar en el sentido que aproximan dicha relación media varianza en lugar de tomarla en un concepto más general, como se haría en un modelo bayesiano jerárquico (Díez, 2022; Liu et al., 2021).

Por ejemplo, Limma (Tabla 1), realiza una regresión lineal ponderada, también conocida como modelo lineal general, pero basándose en la idea que los residuos siguen la distribución normal. No obstante, los pesos asignados, al ser ponderada, se extraen de la relación media varianza. Mientras que en EdgeR y DESeq2 (Tabla 1), toman la aproximación de la dispersión de una forma distinta. Es decir, se sabe que matemáticamente, la suma de varias

distribuciones de Poisson con un parámetro agregado crea la sobredispersión. Matemáticamente la suma de dichas distribuciones conlleva a una nueva distribución llamada binomial negativa. La idea de la distribución binomial negativa es similar a la de Poisson, es decir, en Poisson, se cuenta el número de veces hasta que existe una observación. En la distribución binomial negativa, sin embargo, podemos pensar como cuando se tira un dado. Si se apuesta, por ejemplo, al número tres, tendremos una probabilidad de un sexto de que salga lo que queremos (Díez, 2022; Liu et al., 2021). No obstante, al tratar de la expresión génica, en una primera instancia, no sabemos qué genes están expresados, por lo tanto se apuesta a todos menos al que estamos interesados, y de hecho tiene un símil con Poisson, porque se hacen cuentas hasta que se observa lo que se desea (Stupnikov et al., 2021).

Tabla 1. *Cuadro comparativo que emplea las metodologías de RNA-seq en base a los paquetes estadísticos de expresión diferencial de datos*

	Limma-voom	**EdgeR**	**DESeq2**
Modelo probabilístico	Empírico lineal bayesiano	Binomial negativo	Binomial negativo

Fuente: Castañeda (2021).

Ya sea el modelo lineal ponderado, o un modelo de regresión binomial negativo, ambos modelos se les introduce dentro del marco bayesiano como la función verosimilitud, o, en otras palabras, la creencia de que los datos nos expliquen la expresión diferencial (Chamorro, 2019). Entonces ya sea una regresión lineal ponderada, o una regresión binomial negativa, al final se obtienen unos coeficientes, los cuales darán como resultado una función matemática que dirá cómo los datos explican la expresión diferencial de los genes. Sin embargo, suposición de que exista en realidad una expresión DE, difiere en los métodos, como se ha mencionado antes. Limma ya se ha discutido, por otro lado, en EdgeR y DESeq2, se observa que la relación de la media con la dispersión es inversamente proporcional (Chamorro, 2019; Díez, 2022; Sánchez, 2015).

En los tres métodos comentados, el prior o la creencia que va a actualizar el conocimiento de la función matemática se estima de forma empírica. Es decir, es un marco bayesiano empírico (Jiménez-Jiménez et al., 2021, Lee et al., 2015). En otras palabras, si bien la función matemática a la cual se le va a actualizar el conocimiento puede diferir, no obstante, la información con la que se actualizará sigue siendo la misma, empírica. Esto supone un balance entre la cantidad de datos, y el adecuado preprocesamiento de los datos, para que los resultados sean más fiables.

Un modelo bayesiano jerárquico, bien puede ser la solución para el problema de la sobredispersión de los datos, así como también una alternativa al marco empírico de bayes. En este marco, la suposición es que los datos siguen una distribución binomial negativa, no obstante, matemáticamente, se puede reescribir marginalmente respecto a la tasa de descubrimientos de la distribución de Poisson. Ahora bien, al computar la distribución marginal de la distribución binomial negativa, que cómo se ha comentado anteriormente se transforma de nuevo en una de Poisson, la tasa de descubrimiento de dicha distribución ya no es una constante, sino es otra distribución distinta (Lee et al., 2015).

Entonces, teniendo en cuenta que la tasa de descubrimientos de eventos de la distribución de Poisson es ahora una distribución log-normal, que simplemente es una distribución gaussiana donde la media y la varianza están en forma logarítmica. Donde la media de la tasa de descubrimientos de Poisson es la esperanza de dicha distribución logarítmica normal, la varianza toma otro parámetro, igualmente aleatorio, con distribución exponencial, el cual se le refiere como la sobredispersión de los datos (Neudecker y Katzfuss, 2023; Lee et al., 2015). En palabras más simples, se sabe que existe una sobredispersión, es decir, un parámetro a modelar, que produce el fenómeno en cuestión y dicho parámetro es naturalmente aleatorio. Donde dicha aleatoriedad va a estar definida por otros parámetros aleatorios, donde uno será la media de la sobredispersión y el otro la varianza de la

sobredispersión, siento otro parámetro aleatorio. Es decir, existe una jerarquía en cuanto a las creencias de cómo se comporta la biotecnología subyacente para la secuenciación masiva en paralelo de RNA (Lee et al., 2015).

Cómo se ha comentado, la metodología de BADER consta de una serie de creencias jerárquicas, mas no empíricas, las cuales permiten realizar el análisis DEG de una forma distinta a la de la metodología más popular. Donde en la última se realiza un análisis por gen, mientras que en BADER, se realiza dicho análisis en un conjunto, evitando el problema de la corrección de múltiples hipótesis, que puede llevar a cabo al descubrimiento de falsos biomarcadores (Alarcón, 2019).

En un marco completamente bayesiano, no empírico, se necesitan realizar, como se ha comentado, una actualización de conocimientos. Donde se puede pensar que dicha actualización de conocimientos como una cadena, o los elementos que formarán una cadena, paso por paso siguiendo una probabilidad matemática que modela el fenómeno detrás. El modelo del fenómeno de la secuenciación es la probabilidad de esta jerarquía de secuencias, se le conoce como Monte Carlo, y el proceso de crear la cadena, llevada por las probabilidades del modelo, es lo que se le denomina una cadena de Markov, en palabras sencillas, es cómo un punto en específico puede llevar a otro punto, siguiendo cierta probabilidad. Donde la probabilidad es justamente el modelo construido. En conjunto esta metodología se le conoce como remuestreo de Monte Carlo por cadenas de Markov (MCMC) (Lee et al., 2015; Vélez y Correa, 2013). En otras palabras, se quiere ir de un punto a otro, siguiendo un modelo probabilístico, el cual en teoría siguen los datos de RNA-seq.

1.3. Planteamiento del problema

La reducida aplicación de la Bioinformática en el procesamiento de datos complejos es un problema que se acentúa en países de Latinoamérica como Argentina, Bolivia, Venezuela y Ecuador (Castellanos y Melo, 2021). La Bioinformática es una ciencia multidisciplinar que

permite desarrollar, investigar y aplicar herramientas informáticas en el manejo de datos biológicos, las cuales son accesibles y se mantienen en constante evolución (Instituto de Salud Carlos III, 2020). Sin embargo, en el Ecuador la Bioinformática no es utilizada en las investigaciones en el campo de la salud y existe carencia de profesionales en el manejo de técnicas avanzadas de análisis de datos (Fernández, 2022).

Por consiguiente, se destaca la importancia de las herramientas de los NGS como el RNA-seq junto con herramientas bioinformáticas, para analizar de manera más precisa y detallada la expresión génica, lo que a su vez puede proporcionar una comprensión más completa entre las modificaciones de genes de fusión. Sin embargo, la gran cantidad de datos producidos por el RNA-seq requiere de técnicas avanzadas de análisis y modelización estadística para su interpretación adecuada (Chamorro, 2019; Nonell, 2019).

1.4. Justificación de la investigación

La tecnología de secuenciación se ha convertido en una herramienta fundamental en Biología molecular y Genética, con ello la secuenciación de alto rendimiento ha permitido que los datos genómicos están ampliamente disponibles. Una de estas tecnologías es el RNA-seq, que permite medir la expresión diferencial de genes en una serie de muestras, en determinados estudios. La información obtenida se puede utilizar para identificar genes entre diferentes condiciones, posibles mecanismos moleculares, comprender vías biológicas involucradas en una respuesta específica, medir el nivel de expresión de un gen, su relación con la predisposición de la enfermedad y reconocer el empalme aberrante (Hong et al., 2020; Marco-Puche et al., 2019).

Sin embargo, el análisis de los datos de RNA-seq presentan varios desafíos metodológicos, como la profundidad de la secuenciación, longitud del gen, variabilidad técnica y biológica, el ruido de fondo, degradación de transcripción y la baja frecuencia de expresión de muchos genes (Kellman et al., 2021; Vestal et al., 2020). Como se ha mencionado

anteriormente, la ingente cantidad de datos es necesario analizarlos con diversas técnicas bioestadísticas, por ejemplo, la metodología subyacente a DESeq2. Cabe recalcar que todos los métodos realizados hasta la fecha se basan en el marco bayesiano, por el hecho de la sobredispersión de los datos, el cual es unos de los problemas (Chamorro, 2019; Vardhanabhuti et al., 2013). Por lo tanto, el objetivo de este estudio es comparar si el método utilizado en BADER, el cual es un modelo meramente bayesiano, no empírico, donde se tienen múltiples creencias acerca de la distribución de los datos, por lo que se conoce como jerárquico, puede llegar a las mismas conclusiones que el artículo en cuestión.

1.5. Limitaciones

El presente proyecto es meramente comparativo. Es decir, a pesar de ser un estudio contundente, al comparar dos métodos dado un artículo de referencia ya publicado, no se tiene en consideración una simulación de datos con otros métodos. No obstante, a pesar de que BADER, es un método innovador, y el artículo correspondiente al método lo recalca con simulaciones matemáticas, son mejores que DESeq2, no está validado con otras simulaciones matemáticas, como en la que se basa DESeq2 (Neudecker y Katzfuss, 2023). Es decir, el método está validado para las creencias del método mas no para otros tipos de suposiciones sobre las distribuciones subyacentes en las que probablemente las lecturas de experimentos de RNA-seq en realidad provengan (Schurch et al., 2016).

1.6. Objetivos de la investigación

1.6.1. Objetivo general

Analizar la expresión diferencial en conjunto a la clasificación del fenotipo mediante modelos bayesianos jerárquicos para la obtención de información más fiable tanto de la variabilidad genética y la condición de estudio

1.6.2. Objetivos específicos

- Realizar la lectura, procesado y normalización de datos obtenidos de la expresión diferencial mediante paquetes de R para la organización de los datos del gen de fusión RUNX1-JAK2.
- Aplicar la expresión diferencial de los datos con el paquete DESeq2 utilizando como variable dependiente a los genes e independiente a la condición fenotípica para su interpretación mediante *Genome Set Enrichment Analysis* (GSEA).
- Introducir una modificación donde la variable dependiente representa la condición fenotípica e independiente a los genes, con el uso de un análisis de supervisión para el correcto empleo del modelo bayesiano jerárquico en las dos situaciones.

1.7. Hipótesis

La información conjunta de dos métodos, DESeq2 y modelos bayesianos jerárquicos, produce más información fiable que los métodos por separado.

CAPÍTULO II: MARCO TEÓRICO

2.1. Leucemia linfoblástica aguda precursora de células B (LLA-B)

La leucemia linfoblástica aguda precursora de células B (LLA-B), es una neoplasia pediátrica que afecta a la médula ósea y la sangre, generada por un desorden de las células inmaduras de la línea linfoide B (ausencia de blastos en la sangre periférica). En adición a ello, también puede provocar complicaciones secundarias como patologías cardiovasculares, valvulares y endocárdicas (Garaventa, 2018; Instituto Nacional del Cáncer, 2022; Tello y Novoa, 2020). La incidencia de esta enfermedad a nivel global tiene una frecuencia absoluta en la etapa infantil, sobre todo en niños de 3 a 5 años, con un valor de 5.3:100000, y aumenta una segunda alta frecuencia a partir de los 80 años 2.3:100000 (Onkopedia, 2022).

2.1.1. Diagnóstico de LLA-B

Previo al diagnóstico de la LLA-B, se deben considerar una serie de problemas que generalmente se suscitan, como: la debilidad, pérdida de peso, sudoración nocturna, cansancio, entre otros. Entre los principales hallazgos se encuentran los blastos en estirpe linfoide en sangre periférica (leucocitos) y también infiltrados en la médula ósea, así como, el aumento de LDH, trombopenia, entre otros (Lemes et al., 2022).

Los exámenes más empleados suelen ser de médula ósea en el cual los blastos se encuentran entre el 25 y 95% de los pacientes con LLA-B, hemograma y frotis periférico donde los blastocitos pueden estar al 90% de su recuento y es indispensable desempeñar un diagnóstico diferencial para observar la existencia de mononucleosis infecciosa provocado por el virus de Epstein-Barr (VEB) (Emadi y York, 2022; Lemes et al., 2022).

2.1.2. Patogenia de la LLA-B: gen de fusión RUNX1-JAK2

La LLA-B se produce por un compendio de aberraciones genéticas, donde la transformación maligna se genera en una célula madre pluripotente, o puede darse en una célula madre especializada con capacidad de autorrenovación limitada (Emadi y York, 2022).

Una vez dada esa afección, las células que adquirieron la mutación proliferan anormalmente y con el tiempo desplazan las células normales en la médula ósea, lo que genera un bloqueo de la diferenciación en cierto nivel de maduración. La LLA se asocia a mutaciones en el cromosoma 11 o 12, como en el caso de la translocación t (12;21) (p13; q22) que causa el gen de fusión ETV6-RUNX1 o conocido como TEL-AML1, que representan el 25% de los casos de LLA pediátrica. Suceso que no se da de igual forma con RUNX1-JAK2, del cual su incidencia no es muy conocida. A pesar de ello, las alteraciones en RUNX1 se producen particularmente en el exón 21 y de JAK2 en el exón 12 o 14 (Navarrete y Pérez, 2017; Onkopedia, 2022).

Es de esta forma que, los reordenamientos afectan los genes implicados en el proceso de señalización con el receptor de citoquinas o quinasas como son ABL1, ABL2, EPOR y JAK2. En este encuentro, normalmente, la vía de señalización JAK-STAT (J*anus kinases - Signal Transducers and Activators of Transcription*), transmite información desde las moléculas químicas que están fuera de la célula hasta la parte interna donde está el núcleo, para activar genes específicos. Entonces, la tirosina quinasa C-terminal JAK2 no receptora, desempeña funciones cruciales en la hematopoyesis, como son la diferenciación, proliferación y supervivencia. Sin embargo, las mutaciones que lo involucran activan de forma constitutiva las quinasas, que posteriormente permitirán la señalización proliferativa y/o antiapoptótica. Dicho en otras palabras, que las células se reproduzcan de forma incontrolada y sin un mecanismo de muerte celular dirigida a aquellas que lo ameriten. Dado esto, la fusión del gen JAK2 se ve impulsada por una proteína N-terminal que incide en la correcta hematopoyesis (Fortschegger et al., 2021; Layton, 2015).

Por otra parte, el compañero de fusión N-terminal RUNX1, es un factor de transcripción indispensable en la hematopoyesis temprana, y en etapas avanzadas de desarrollo interviene en la diferenciación y supervivencia del linaje megacariocito. Por ello, este gen cumple con la

actividad que produce la vía MYC, misma que une su factor de transcripción al DNA y regula las actividades como división celular, crecimiento, apoptosis y transformación celular (Fortschegger et al., 2021)

No es coincidencia que los cambios transcripcionales observados están gobernados principalmente por la señalización JAK-STAT mediada por RUNX1-JAK2 y la posterior activación de la vía MYC. Para que den como resultado el desarrollo de la LLA-B. Ya que de acuerdo con Fortschegger et al. (2021), la haploinsuficiencia, que es la limitada cantidad de un gen para que funcione adecuadamente, RUNX1 reduce la producción de células hematopoyéticas; por ende, al no contar con la cantidad requerida de células, el gen JAK2 se encarga de la transcripción aumentada, produciendo en conjunto un desorden neoplásico.

2.2. Secuenciación del transcriptoma: RNA-seq

La tecnología de RNA-seq es una técnica que permite un completo acercamiento al perfil de la expresión genética, en base a los niveles medidos de los transcritos y sus isoformas (Castañeda, 2021). De acuerdo con Marco-Puche et al. (2019), el flujo de trabajo para un conjunto de datos de RNA-seq es el siguiente:

1. Extracción y purificación de RNA: muestra de interés (tejido o células).
2. Preparación de la biblioteca: donde se convierte el RNA a cDNA, agregando adaptadores a la secuencia.
3. Secuenciación: NGS con plataformas Ilumina.
4. Alineación con el genoma de referencia: Se encajan los fragmentos que son productos de la amplificación por puente al genoma por defecto.
5. Cuantificación de expresión génica: recuento del número de lecturas que fueron alineadas.
6. Post-procesado de la expresión génica: normalización de datos, en el caso que fuese necesario, y empleo de controles de calidad.

7. Análisis de expresión diferencial: permite la comparación del número de veces que un transcrito se sobreexpresa o infraexpresa respecto a un control.

2.3. Expresión diferencial de los genes (DEG)

La expresión diferencial de genes (DEG) permite solamente observar, como su nombre bien indica, la diferencia entre condiciones. Es decir, entre contrastes que vienen dictados por el diseño del experimento. Cabe recalcar, que la expresión diferencial, no solamente aplica a datos de RNA-seq, sino a cualquier metodología que involucre una ómica. Dicho en otras, incorpora a cualquier capa biológica de interés, y no se limita a bloques generados por NGS. Otros métodos pueden ser Sanger, tecnología por nanoporos, tanto en transcritos, SNPs, proteoma, metaboloma, epigenoma, etc. (Hernández, 2021).

La herramienta diferencial, proporciona el cálculo del cambio relativo de una condición a otra. No obstante, a pesar de que se puede pensar a priori que se observa, o se contrasta una diferencia, es decir, una resta entre concentraciones, lecturas o expresión, al estar tratando con logaritmos, en realidad se mira el cambio de proporciones (Yu et al., 2021).

2.4. Lenguaje de programación R

R es un software ideado para la programación estadística, que nació a principios de la década de los 90s. A pesar de ser popularmente conocido por ser versátil en el ámbito de la estadística, y el manejo de grandes bases de datos, sobre todo con paquetes como *datatable,* es una herramienta flexible para cualquier problema Bioinformático, que además se entiende con otros lenguajes de programación, como Python (Rodríguez y Shishkova, 2019).

Al igual que Python como Biopython, como todo lenguaje de programación contiene un repositorio de librerías, o paquetes, que poseen funciones específicas para tareas en concreto, que dan especial énfasis a las temáticas dedicadas a la Bioinformática, este repositorio se lo conoce como *Bioconductor*. De hecho, muchos de los paquetes que se

encuentran dentro de este, se basan en la programación orientada a objetos, haciendo más fácil cualquier tarea relacionada con la bioestadística y la Bioinformática (Jiménez, 2019).

2.4.1. Definición de paquete (package)

Al ser un software de libre uso, las librerías o paquetes, son desarrollados por la comunidad. Además, cada paquete incluye código, manual de uso y conjuntos de datos específicos para entender mejor las funciones de cada librería y sus funciones (Sancho, s.f.).

2.4.2. Definición de función

Una función es cualquier bloque de código que desempeña una operación y que puede encapsularse para volver a ser utilizado. Se caracterizan por tener parámetros de entrada, *inputs*, que permiten pasar los argumentos de llamado y devolverlos como valor de salida, *outputs* (Whitney et al., 2023).

2.5. Flujo de trabajo

El análisis de los datos de RNA-Seq para la expresión diferencial comprende un procedimiento específico donde los genes deben ser tratados de manera cuidadosa, con el objeto de evitar resultados erróneos. Dado esto se explica la sistemática del análisis, así como la matriz teórica de cada etapa.

2.5.1. Post-procesado de datos

Una vez los fragmentos han sido alineados y las lecturas se presentan en una matriz de contajes, estas se procesan a cargar en el ambiente de trabajo. Posteriormente, si el análisis lo requiere son normalizados, o si no son normalizados con tal de realizar un control de calidad, con técnicas multivariantes como ahora el análisis de componentes principales. Una vez dichos controles son aprobados, se procede con el análisis DE (Burgos, 2021).

2.5.2. Lectura de datos

En este punto, se leen los datos crudos. Dichos datos pueden venir en distintos formatos. El más común es tsv (*tab delimited values*). No obstante, en el presente proyecto se obtuvieron

de la base de datos *GEO* (Gene Expression Omnibus), un repositorio de la página NCBI. Dichos datos en este caso se encuentran en formato delimitado por comas "csv" (Owens et al., 2019).

2.5.3. Filtrado de genes

El filtrado de genes es un proceso en el cual se eliminan aquellas lecturas nulas o con bajos recuentos. Proceso que mejora la robustez y eficiencia del proceso, reduciendo la complejidad computacional al momento en que se borran los genes que no competen un interés y mejoran los niveles de dispersión. Normalmente se filtran aquellas lecturas menores a 10 conteos. Es decir, todo gen o transcrito, a través de las muestras que tenga un conteo menor a 10 es eliminado, estos son los denominados "*outliers*" (Hernández, 2021).

Dentro del lenguaje de programación y en el ámbito matemático, la estrategia que se emplea es calculando el número total de lecturas para cada gen, se suman estas lecturas en todas las muestras y se otorga un rango de valores como ">" y la cantidad de lecturas mínimas que debe tener (Arias y Muñoz, 2019).

2.5.4. Normalización de datos

De acuerdo con Chamorro (2019) y Owens et al. (2019), la normalización de datos es un proceso mediante el cual se busca minimizar la cantidad de ruido técnico que se produce a razón de la secuenciación de datos en RNA-seq, para que puedan ser equiparables entre ellos, al momento del empleo del análisis de expresión diferencial. Es así como, normalizando se convierten los datos a una misma escala de todas las muestras. Existen algunas formas de la normalización de datos, sin embargo, la más atractiva y que emplea DESeq2 se conoce como "factor del tamaño o normalización", misma que asume que la mayoría de los genes no están diferencialmente expresados dentro de las diferentes condiciones experimentales, lo que dice que la variación en los contajes de lecturas para casi todos los genes se deba a razones técnicas en contraste con temas biológicos.

2.5.5. Control de calidad previo al análisis de expresión diferencial

El control de calidad previo al análisis de expresión diferencial, generalmente se realiza mediante análisis de componentes principales (PCA) o escalados multidimensionales. A pesar de que ambas técnicas difieren entre sí, disponen de formas de trabajo similares que, pretenden visualizar todo el transcriptoma, es decir más de 100000 dimensiones, en un solo gráfico de dos dimensiones, el cual conserva la máxima información posible del conjunto de todas las variables. Esto es posible, maximizando la varianza y computando unas nuevas variables, que se disponen en componentes principales (Gil, 2020).

Lo que permiten dichas técnicas, es poder visualizar ciertos errores sistemáticos en la distribución de las muestras. Pero también permiten ver "*a priori*", si el experimento está bien fundamentado. Además, de la posibilidad de observar si los grupos de interés se separan entre sí de acuerdo con la hipótesis planteada. Es por todo eso, que este proceso es un claro ejemplo de estudio de control, si ambos grupos se separan entre sí antes de realizar el análisis DE (Ferrer, 2018; Gil, 2018).

2.6. Paquete DESeq2

DESeq2 es un paquete empleado en el lenguaje de programación de R, el cual se encuentra dentro de *Bioconductor*, para el análisis de datos RNA-seq. Cumple con el trabajo de estimar la variabilidad de los recuentos de genes y ajustarlos al modelo de regresión binomial negativa (Chamorro, 2019).

La metodología que emplea se basa en el cálculo de la función de verosimilitud, es decir la "probabilidad", de que los datos expliquen cierta realidad. Este, realiza dos pruebas estadísticas, la primera se le conoce como "*Likelihood Ratio Test*" (*LRT*). Esta prueba, compara la regresión binomial negativa, que la podemos pensar como un "*ANOVA*" con una distribución diferente, y se va a comparar con un modelo nulo, es decir, un modelo donde no se tiene en cuenta ningún parámetro. Posteriormente, una vez la creencia o la realidad es computada a

partir de los datos, con la metodología empírica, se realiza la prueba de Wald, el cual básicamente es un "*t.test*" o comparación de "medias" de expresión sobre los coeficientes del modelo, que serían las condiciones experimentales sobre cada uno de los genes. Dicha prueba arroja un estadístico con su correspondiente al valor de "p" o "*p-value*". Donde la hipótesis nula es aquella en que las condiciones no tienen relación alguna con la expresión diferencial. Normalmente en estadística, se toma un valor de o menor a 0.05, para considerar el rechazo de la hipótesis nula, aunque dicho valor lo define el investigador. Dicho de otra forma, argumenta que los cambios de expresión genética dadas ciertas condiciones no se deben al azar y, también emplea a "*LogFoldChange*" (*LFC*), el cual mide la modificación de expresión de los genes en las condiciones del experimento. Este parámetro, representa la significancia biológica o que tanto un gen se va a expresar más o menos respecto a las condiciones del estudio (Love et al, 2023).

Finalmente, emplea la corrección de pruebas múltiples como mecanismo de control ante posibles falsos positivos y la visualización de los datos mediante el denominado "*Volcano Plot*". Dicho gráfico nos dice si los genes son estadísticamente significativos, además de que el cambio biológico lo sea. Cabe mencionar que la ventaja de DESeq2, es su marco de trabajo, donde solamente se requiere de una línea de código para realizar el análisis DEG (Arias y Muñoz, 2019; Piñero, 2017).

2.7. Paquete BADER

Hasta ahora DESeq2, como se ha mencionado, igualmente es un método bayesiano, donde la función de verosimilitud o el modelo matemático es una regresión binomial negativa. Es decir, es una función lineal, la cual nos dice el cómo dado un conjunto de datos de RNA-seq se van a expresar. No obstante, para actualizar el conocimiento del modelo se requiere de una aproximación empírica de la realidad (Yang et al., 2020).

Por otro lado, la librería BADER, emplea un modelo jerárquico. Se lo denomina de esa forma, porque hay creencias a partir de suposiciones. La distribución subyacente sigue siendo binomial negativa, pero se puede modificar a una de Poisson nuevamente. Donde el parámetro de la distribución de Poisson, es decir la tasa hasta que descubre un evento ya no es un número conocido, sino es una probabilidad. Donde dicha probabilidad está definida por una distribución log-normal, la cual esta parametrizada por la media de la tasa, y la varianza la cual a su vez está modelada por el parámetro de la sobredispersión en forma de una distribución exponencial. Y La estructura de un modelo jerárquico con el enfoque bayesiano, describe que dispone de un primer nivel que son las combinaciones de los hiperparámetros ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$), éstos a su vez, generan el parámetro "θ" responsable de la aparición de las condiciones de interés en "y" genes de un grupo de "n" genes de estudio (i = 1, 2, 3, ..., m). Es por este motivo que, lo que se busca inferir, es la distribución posterior o, en otras palabras, como la realidad viene explicada por los datos y no por especulaciones (Vélez y Correa, 2013).

De esta forma, Bayes (Ecuación 1), nos dice que la realidad explicada por los datos no es la misma que los datos que explican la realidad. Se necesita a parte tener en cuenta por sí misma dicha creencia. Donde la creencia es una serie de postulados jerárquicos. Es decir, es conjunto la creencia con la probabilidad de la creencia que los datos nos la van a explicar. Para lograr explicar la expresión diferencial verdadera, se requieren de múltiples pasos que la recreen, siguiendo el supuesto de que los datos en efecto nos van a dar la respuesta. A este proceso se le conoce como cadena de Markov guiada por un proceso de Monte Carlo. Donde la cadena son los pasos, y el proceso de Monte Carlo, no es más que la probabilidad que los datos en realidad nos puedan explicar la DEG (Alarcón, 2019; Risso et al., 2012; Vélez et al., 2013).

Ecuación 1. *Teorema de Bayes*

$$\mathrm{P}(\mathrm{y}|\text{⍉}) = \frac{P(\text{⍉}|y)\,P(y)}{\int[P(y|\text{⍉})][P(\text{⍉})]}$$

Donde,

P(y| θ)=Posterior

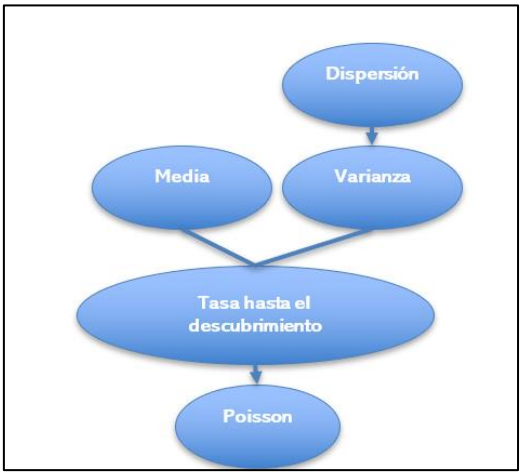
P(θ|y)=Función de verosimilitud

P(y)=Prior

$\int [P(y|\text{ᔕ})][P(\text{ᔕ})]$=Factor de normalización

En este pasaje, la estadística bayesiana, nos dice que la realidad explicada por los datos no es la misma que los datos que explican la realidad o estadística frecuentemente. En la perspectiva bayesiana, no se confía únicamente en los datos para formular interpretaciones sobre la realidad. Se propone que se debe considerar una "creencia" preexistente. Esta creencia previa se compone de una serie de postulados dispuestos de manera jerárquica. En este marco conceptual, se utilizan tanto la probabilidad de la creencia o "*prior*", como los datos para interpretar la realidad. En la situación particular del análisis de DEG, se requiere un procedimiento con múltiples etapas para reproducir de manera efectiva la expresión diferencial auténtica, fundamentándose, en la premisa de que los datos tienen la capacidad para brindarnos respuestas. Este método de análisis se lo conoce como cadena de Markov Monte Carlo (MCMC), donde el término "cadena" se refiere a las distintas etapas del procedimiento, y el "proceso de Monte Carlo" hace alusión a cómo se obtienen la probabilidad que los datos explican de DEG para extraer conclusiones posteriores. Por lo tanto, la clave de este método radica en que fusiona el conocimiento anterior "*prior*" y los propios datos para brindar una representación más integral y posiblemente más precisa de la realidad de DEG -Figura 1- (Neudecker y Katzfuss, 2023; Vélez y Correa, 2013).

Figura 1. *Estructura del modelo bayesiano jerárquico de la actividad funcional en la mutación.*



Fuente: Autores.

CAPÍTULO III: MARCO METODOLÓGICO

En el presente capítulo se expondrá el diseño de investigación, el nivel de investigación, la población y obtención de muestras, las variables utilizadas y el flujo de trabajo del análisis de datos.

3.1. Descripción del diseño general

3.1.1. Diseño de investigación

El diseño de la investigación fue descriptivo, debido al hecho que se compararon dos metodologías para la expresión diferencial, ambas bayesianas, DESeq2 y BADER. En lo que difieren ambas metodologías es que la primera es empírica respecto a la creencia o prior, y la segunda es puramente bayesiana con múltiples creencias acerca de la sobredispersión de los datos, es decir es jerárquica en el sentido de los múltiples priors que se utilizan. La comparativa se llevó a cabo en un experimento sobre células humanas sobre el gen de fusión RUNX1-JAK2. Donde cada variable dependiente son los genes y las independientes son las condiciones de fusión y control.

3.1.2. Nivel de investigación

El objetivo de la investigación fue analizar la expresión diferencial con las dos metodologías anteriormente mencionadas, comparando los resultados del análisis diferencial de expresión respecto a células humanas que contienen la fusión de RUNX1-JAK2, de resultados ya publicados. Es decir, es un caso de estudio comparativo y correlacional.

3.2. Población y obtención de muestras

La población de estudio fueron las células de la línea hematopoyética derivadas de iPSC y la muestra se obtuvo de los datos proporcionados “Expression of RUNX1-JAK2 in Human Induced Pluripotent Stem Cell-Derived Hematopoietic Cells Activates the JAK-STAT and MYC Pathways” de Fortschegger et al. (2021) de https://doi.org/10.3390/ijms22147576.

3.3. Variables

Las variables de interés en este estudio fueron las expresiones génicas de los genes, WT y RUNX1-JAK2, por otro lado, las variables dependientes al utilizar DESeq2 fue la expresión de los genes normalizada y transformada y al utilizar BADER la inferencia bayesiana de la expresión génica.

3.4. Recogida de datos

Los datos fueron recolectados de la base de datos GEO, utilizando el archivo GSE159261_raw_counts.txt.gz. (Fortschegger et al., 2021), los cuales se obtuvieron mediante la técnica de RNA-seq. Los datos se descargaron en formato *fastq*.

3.5. Softwares y paquetes de R usados para el análisis

Para la aplicación del análisis de expresión diferencial y modelos bayesianos jerárquicos al estudio del gen de fusión RUNX1-JAK2, se emplearon dos enfoques de análisis mediante los paquetes DESeq2 y BADER, ambos implementados en el lenguaje de programación R.

3.5.1. Programa de R

El lenguaje de programación R (Figura 2), que se utiliza en este estudio, se encuentra disponible de manera gratuita bajo la licencia GNU. El archivo de instalación puede ser descargado desde la página oficial de R, https://www.r-project.org/. En cuanto al ambiente de trabajo, se accedió a R a través de *RStudio* (Figura 3), que es un entorno de desarrollo integrado (IDE) de código abierto específico para R. La instalación de *RStudio* también es gratuita y puede ser realizada a través de su sitio web oficial, https://cran.rstudio.com/ (Chamorro, 2019). En este trabajo, se utilizó la versión 4.3.0 de *RStudio*.

Figura 2. *Logo de R*

Nota: Tomado de *The Comprehensive R Archive Network*, s.f., https://cran.r-project.org/

Figura 3. *Logo de* RStudio

Fuente: Autores.

3.5.2. Bioconductor

En el contexto de esta investigación, se optó por utilizar la plataforma *Bioconductor* (Figura 4) debido a su reputación en la comunidad científica como un recurso robusto y confiable para el análisis de datos biológicos. Su amplia variedad de paquetes, cada uno diseñado para un propósito específico, ofrece a los investigadores las herramientas necesarias para abordar diversos desafíos analíticos en la genómica y la biomedicina (Bioconductor, 2023).

Los paquetes de *Bioconductor* son conocidos por su acceso intuitivo, la confiabilidad de sus datos y su formato de código abierto, lo que facilita su uso en análisis estadísticos y exploratorios. Además, proporcionan la posibilidad de enriquecer los datos con metadatos adicionales procedentes de fuentes bibliográficas confiables, como *PubMed*, y de sistemas de anotación funcional de genes, como *Entrez Gene* (Gallego, 2021).

En este contexto, se seleccionaron específicamente los paquetes DESeq2 y BADER de *Bioconductor* versión 3.17.

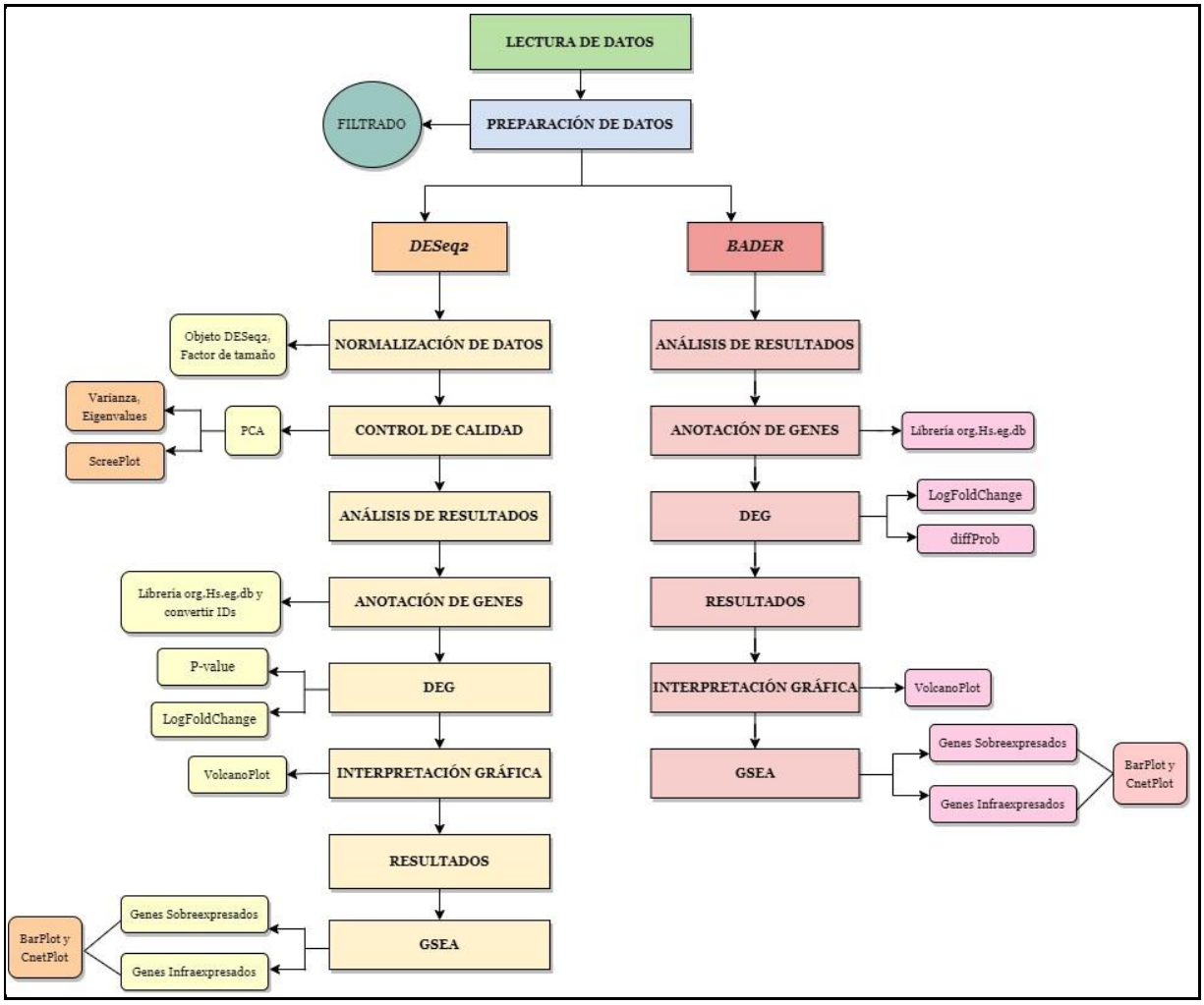
Figura 4. *Logo de Bioconductor*

Nota: Tomado de *Bioconductor*, Bioconductor, 2023, https://www.bioconductor.org/

3.6. Flujo de trabajo de análisis de datos

El flujo de trabajo realizado en este estudio se centra en analizar archivos *fastq.* Por lo que, se desarrolla en 5 fases: lectura de datos, preparación de los datos, análisis de expresión diferencial con DESeq2 y BADER y análisis de resultados -Figura 5-.

Figura 5. *Flujo de trabajo con el cual se trabajó en el análisis DEG con los paquetes DESeq2 y BADER*



Fuente: Autores.

3.6.1. Lectura de los datos

El análisis de expresión diferencial parte de las matrices de conteo que se obtuvieron tras la alineación y de las secuencias en relación con el genoma de referencia. Está matriz de conteos dispone a los genes en filas y las muestras en las columnas (Anexo 2).

3.6.2. Preparación de los datos

El archivo *fastq* que contiene la matriz de contajes se cargó en R, luego, se determinó la condición de análisis, *wyld type* y gen de fusión RUNX1-JAK2 (WR y RJ), para realizar el diseño de análisis en relación con la condición. Para ello se realizó, el filtrado de genes con baja expresión, ya que, en la matriz de conteo se obtiene un cierto número de genes que presentan pocas o ninguna lectura en la mayoría de las muestras. Ya que, los genes que nos permiten realizar la expresión génica diferencial son aquellos cuyo nivel de expresión es significativamente diferente a los otros genes (Chamorro, 2019). Por este motivo, se filtraron los genes que tienen recuentos menores a 10 lecturas en total con tal de evitar un recuento alto de genes no diferencialmente expresados, que podrían causar problemas en el análisis aguas abajo (Anexo 2) (Love et al., 2023).

3.6.3. Análisis de expresión diferencial con DESeq2

Después, de haber realizado el filtrado, reduciendo considerablemente el conjunto de genes a analizar tras la eliminación de aquellos genes con nula o poca expresión, se procede a normalizar los datos, realizar el control de calidad mediante PCA y la expresión diferencial.

3.6.3.1. Normalización de los datos

Según Chamorro (2019) y Owens et al. (2019), el término "normalización de los datos", se refiere a reducir el "ruido" o las inconsistencias técnicas que se pueden generar durante la secuenciación de datos en RNA-seq, para convertir los datos a una misma escala para que sean comparables. Por lo que, mediante DESeq2, la forma de normalizar los datos se le conoce como "normalización del factor de tamaño", la cual asume que la mayoría de los genes no cambian

mucho en su expresión en diferentes condiciones experimentales, por lo que, cualquier variación en la cantidad de lecturas de genes se debe probablemente a factores técnicos más que cuestiones biológicas. Por lo tanto, los recuentos de genes se normalizaron para eliminar el sesgo introductorio por las diferencias en la profundidad de secuenciación en las muestras (Anexo 2).

3.6.3.2. Análisis de componentes principales (PCA)

Las técnicas de análisis de calidad, como el análisis de componentes principales (PCA), se utiliza antes de realizar la expresión diferencial. La cual permite simplificar y visualizar un conjunto grande y complejo de datos (más de 100000 dimensiones en el caso del transcriptoma) en un gráfico simple de dos dimensiones. Esto es posible mediante la maximización de la variación entre los datos y la creación de nuevas variables no correlacionadas entre sí, llamadas componentes principales, que resumen la información más importante de los datos. Permitiendo así detectar errores sistemáticos en la distribución de las muestras, verificar si la base del experimento es sólida y nos muestra si los grupos que estamos estudiando se separan entre sí de acuerdo con la hipótesis que estamos probando, en este caso, WT y RJ (Ferrer, 2018; Gil, 2020).

Por lo tanto, se realizó un PCA para explorar la variación en los datos y las relaciones entre las muestras de los datos de expresión basada en el análisis realizado por Kassambara y Mundt (2020) (Anexo 2).

3.6.3.3. Análisis de expresión diferencial

El análisis de expresión diferencial mediante DESeq2 permite estimar la variabilidad del recuento de los genes y las ajusta utilizando un modelo matemático llamado regresión binomial negativa. En este contexto, utiliza una función llamado verosimilitud, medida de la probabilidad que los datos expliquen un determinado resultado o realidad; por lo que, se realizan dos pruebas estadísticas, la prueba de razón de verisimilitud y la prueba de Wald. La

prueba de razón de verisimilitud (LRT) imagina que un evento tiene dos explicaciones y las compara para determinar cuál es más probable que sea correcta. En cambio, la prueba de Wald analiza la importancia de cada condición experimental (WT y RJ) en la expresión del gen. Por consiguiente, después de estas pruebas se obtiene un valor p que es una forma de medir cuán confiables son los resultados, por lo que, un valor menor a 0.05 sugiere que nuestros hallazgos son significativos y no al azar. Además, obtiene un *LogFoldChange* que mide cuánto cambia la expresión diferencial de un gen bajo las condiciones WT y RJ (Chamorro, 2019; Love et al., 2019). Por lo tanto, se realiza la expresión diferencial mediante DESeq2 para determinar la variación en la expresión de los genes entre diferentes condiciones experimentales, WT y RJ, basada en el análisis realizada por Love et al. (2023) (Anexo 2).

3.6.4. Análisis de expresión diferencial con BADER

Después de haber realizado la preparación de los datos, se utilizan los datos filtrados para realizar el análisis de expresión diferencial mediante BADER. Este análisis se basa únicamente en los resultados de los datos teniendo en cuenta tus propias creencias y suposiciones previas sobre las condiciones. En este contexto, se desarrolla mediante un modelo llamado jerárquico, que implica niveles de probabilidad de un evento (genes de interés) para llegar a un resultado final, lo cual sigue una distribución “Log normal” que considera otros factores que influyen en la dispersión de los datos. Por consiguiente, este modelo parte de ciertos parámetros conocidos por hiperparámetros (hiperpriors), que generan otro parámetro que determina la condición que nos interesa de los genes, es decir, se trata de deducir cómo los datos explican la realidad en lugar de simplemente hacer suposiciones. Este proceso se basa en la inferencia de Bayes, que afirma que los datos que explican la realidad no son exactamente la misma cosa que la realidad explicada por los datos, es decir, hay que tener en cuenta las creencias de los datos, ya que, se convierten en una serie de suposiciones de manera jerárquica (Diaz, 2019; Neudecker y Katzfuss, 2023).

Además, para obtener una representación precisa de la expresión diferencial (cómo los genes cambian su actividad en respuesta de ciertas condiciones), BADER sigue un proceso llamado cadena de Markov Monte Carlo (MCMC), el cual consta de varios pasos que se basan en la suposición de que los datos nos darán la respuesta correcta, donde la cadena son los pasos que se siguen y el proceso de Monte Carlo se refiere a cómo usamos los datos para estimar la probabilidad de que nos puedan explicar la expresión diferencial (Alarcón, 2019; Risso et al., 2012; Vélez et al., 2013).

Por ello, se aplica la metodología de BADER, para realizar el análisis DE, el cual se basa en un modelo completamente bayesiano jerárquico con múltiples priors (Anexo 2).

3.6.5. Análisis de resultados

3.6.5.1. Volcano plot

Un *volcano plot* o gráfico de volcán nos ayuda a entender de un vistazo si los cambios que estamos viendo en los genes son importantes desde un punto de vista estadístico, es decir, si podemos confiar en que son reales y no son casualidades. Además, nos muestra si esos cambios son biológicamente significativos. En este contexto, en el eje X, muestra el cambio de expresión de cada gen entre dos condiciones, WT y RJ, mientras que el eje Y, muestra cuan seguros estamos de ese cambio (significancia estadística), por lo tanto, se determina según límites los genes que son estadísticamente significativos como biológicamente importantes (Arias y Muñoz, 2019; Piñero, 2017). Por lo que, con tal de analizar los datos, se computó un *volcano plot* tanto para los resultados de análisis de expresión diferencial de DESeq2 como BADER, el cual se realiza con la significancia estadística (p-valores) y la biológica (fold change). Donde en el eje de las ordenadas se colocan los p-valores y en las abscisas el fold change (Anexo 2).

3.6.5.2. GSEA

El análisis de enriquecimiento de conjunto de genes (GSEA), es una técnica que nos ayuda a entender como un grupo de genes puede influir en ciertos procesos biológicos o rutas en nuestro cuerpo, todo esto es a partir de la información obtenida de las expresión de genes. Por lo que, toma una lista de genes ya ordenados a ciertos criterios que nos indican que tan diferentes se están expresando los genes (menor a 0.05 mayor significancia) en comparación de una condición de referencia. Además, permite etiquetar grupos de genes que comparten características similares o funciones en común (Khalfan, 2021). Por lo tanto, se computó un GSEA tanto para los datos de análisis de expresión diferencial de DESeq2 como de BADER, donde se determina los genes individuales que muestran los mayores cambios de la expresión (genes DE) y las vías biológicas que pueden estar implicadas en LLA (Anexo 2).

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1. Análisis de datos

Los datos utilizados en el estudio fueron recolectados de la base de datos GEO del documento llamado GSE159261_raw_counts.txt.gz., https://bit.ly/45beXV3. Donde se determinó 60683 contajes y 18 condiciones -Tabla 2-.

Tabla 2. *Número de muestras utilizadas*

Nº de muestra	Código	Nombre de Referencia
1	GSM4824469	wild-type_1_DMSO
2	GSM4824470	wild-type_1_dTAG
3	GSM4824471	RUNX1-JAK2_cloneA6_DMSO
4	GSM4824472	RUNX1-JAK2_cloneA6_dTAG
5	GSM4824473	RUNX1-JAK2_cloneC6_DMSO
6	GSM4824474	RUNX1-JAK2_cloneC6_dTAG
7	GSM4824475	RUNX1-JAK2_cloneE5_DMSO
8	GSM4824476	RUNX1-JAK2_cloneE5_dTAG
9	GSM4824477	wild-type_2_DMSO
10	GSM4824478	wild-type_2_dTAG
11	GSM4824479	RUNX1-JAK2_cloneG10_DMSO
12	GSM4824480	RUNX1-JAK2_cloneG10_dTAG
13	GSM4824481	RUNX1-JAK2_cloneH1_DMSO
14	GSM4824482	RUNX1-JAK2_cloneH1_dTAG
15	GSM4824483	wild-type_3_DMSO
16	GSM4824484	wild-type_3_dTAG
17	GSM4824485	RUNX1-JAK2_cloneF2_DMSO
18	GSM4824486	RUNX1-JAK2_cloneF2_dTAG

Nota: Los nombres de referencia son las condiciones de las muestras, donde se determina wild type y RUNX1-JAK2.

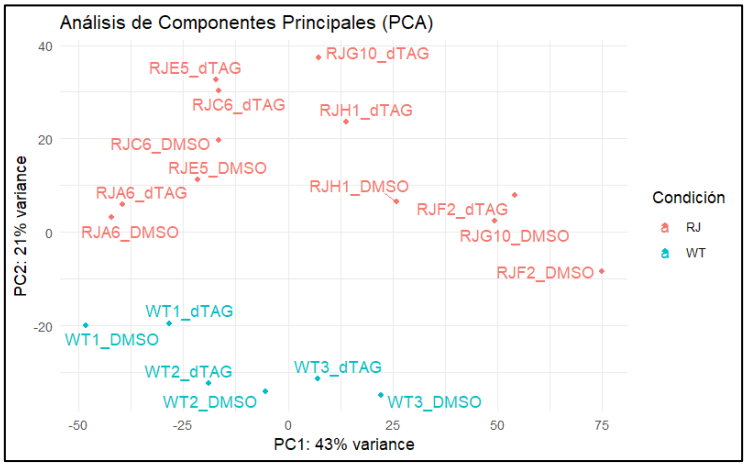
Fuente: Fortschegger et al. (2021).

4.2. Presentación de los datos

4.2.1. Resultados del análisis de componentes principales

Para obtener una visualización general de los datos del RNA-seq, se realizó el PCA, lo cual permitió visualizar la variación total de la expresión génica de las muestras.

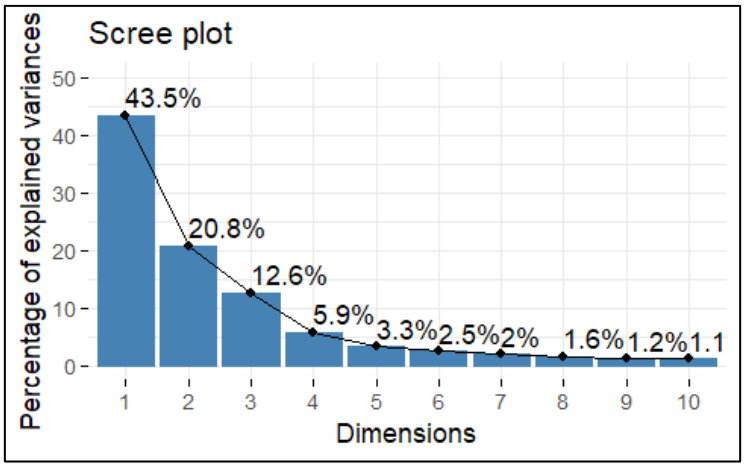
Figura 6. *Varianza de los componentes principales (PCA)*



Fuente: Autores.

De la Figura 6 los componentes principales PC1 y PC2 explican que la varianza total de la expresión génica es 64%, la PC1 explica el 43% y el PC2 el 21%, lo cual muestra una clara separación de las condiciones WT y RJ debido a la segunda componente. Por otro lado, la varianza explicada por cada componente principal Scree Plot (Figura 7) determina cada punto un componente principal y su respectiva proporción de varianza en relación al conjunto de datos.

Figura 7. *Gráfico de Scree Plot con los porcentajes de la proporción de las varianzas*



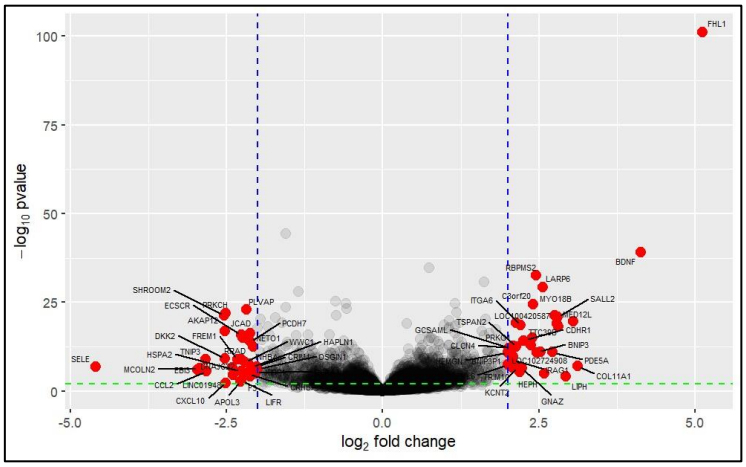
Fuente: Autores

4.2.2. Resultados del análisis de expresión diferencial con DESeq2

El análisis de expresión diferencial utilizando DESeq2, identificó 60 genes diferencialmente expresados entre las condiciones p ajustado < 0.01 y log2(fold change) >2,

por lo que, el *Volcano Plot* determina la distribución de los genes en función del cambio de la expresión y una alta significancia estadística -Figura 8-.

Figura 8. *Volcano Plot del análisis de expresión diferencial realizada con DESeq2*



Nota: Los límites utilizados fueron p value igual a 0.01 y log2 (fold change) de -2 y 2. Fuente: Autores.

Los genes diferencialmente expresados significativamente incluyen genes sobreexpresados que son 29 -Tabla 3-: COL11A1, TSPAN2, KCNT2, GCSAML, EML6, ITGA6, C3orf20, MED12L, LIPH, PDE5A, TRIM10, CLCN4, HEPH, FHL1, TTC39B, HEMGN, IRAG1, BDNF, PRKG1, CDHR1, BNIP3, SALL2, BNIP3P1, RBPMS2, LARP6, LOC100420587, LOC1027224908, GNAZ y MYO18B.

Tabla 3. *DEGs sobreexpresados determinados con DESeq2*

Nº DE GENES	NOMBRE DE LOS GENES
1	COL11A1
2	TSPAN2
3	KCNT2
4	GCSAML
5	EML6
6	ITGA6
7	C3orf20
8	MED12L
9	LIPH
10	PDE5A
11	TRIM10
12	CLCN4
13	HEPH
14	FHL1
15	TTC39B
16	HEMGN
17	IRAG1
18	BDNF
19	PRKG1
20	CDHR1
21	BNIP3
22	SALL2
23	BNIP3P1
24	RBPMS2
25	LARP6
26	LOC100420587
27	LOC102724908
28	GNAZ
29	MYO18B

Nota: DEGs=Genes diferencialmente expresados. Fuente: Autores.

Los genes diferencialmente expresados significativamente incluyen genes sobreexpresados que son 29 -Tabla 4-: COL11A1, TSPAN2, KCNT2, GCSAML, EML6, ITGA6, C3orf20, MED12L, LIPH, PDE5A, TRIM10, CLCN4, HEPH, FHL1, TTC39B, HEMGN, IRAG1, BDNF, PRKG1, CDHR1, BNIP3, SALL2, BNIP3P1, RBPMS2, LARP6, LOC100420587, LOC1027224908, GNAZ y MYO18B.

Tabla 4. *DEGs infraexpresados determinados con DESeq2*

Nº DE GENES	NOMBRE DE GENES
1	MCOLN2
2	F3
3	SELE
4	CRIM1
5	XIRP1
6	PCDH7
7	CXCL10
8	DKK2
9	TNIP3
10	LIFR
11	LINC01948
12	CRHBP
13	HAPLN1
14	ECSCR
15	WWC1
16	AKAP12
17	INHBA
18	SHROOM2
19	FREM1
20	JCAD
21	GFRA1
22	DNAJC22
23	PRKCH
24	HSPA2
25	RRAD
26	OSGIN1
27	CCL2
28	NETO1
29	EBI3
30	PLVAP
31	APOL3

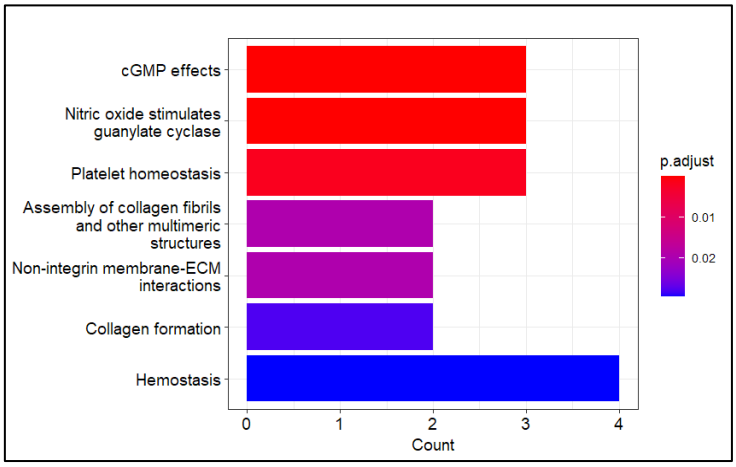
Nota: DEGs=Genes diferencialmente expresados. Fuente: Autores.

4.2.2.1. Resultado de genes sobreexpresados de DESeq2 con GSEA

De acuerdo con el análisis de enriquecimiento para genes sobreexpresados después de emplear el DESeq2, se evidenciaron 8 categorías a nivel de proceso biológico. La hemostasia

contiene la mayor cantidad de genes, sin embargo, mantiene un p valor de casi 0.03. Por otro lado, para los efectos de cGMP, el óxido nítrico estimulante de guanilato ciclasa y la homeostasis plaquetaria mantiene un p valor menor a 0.01, por lo que, tiene una cantidad moderada de genes. Además, tanto el ensamblaje de fibrillas de colágeno y otras estructuras multiméricas e interacciones membrana-ECM no mediadas por integrinas mantienen un p valor de más o menos 0.02 y una cantidad moderada de genes. En cambio, la vía de formación de colágeno mantiene un p valor de 0.03 más o menos y una cantidad moderada de genes -Figura 9-.

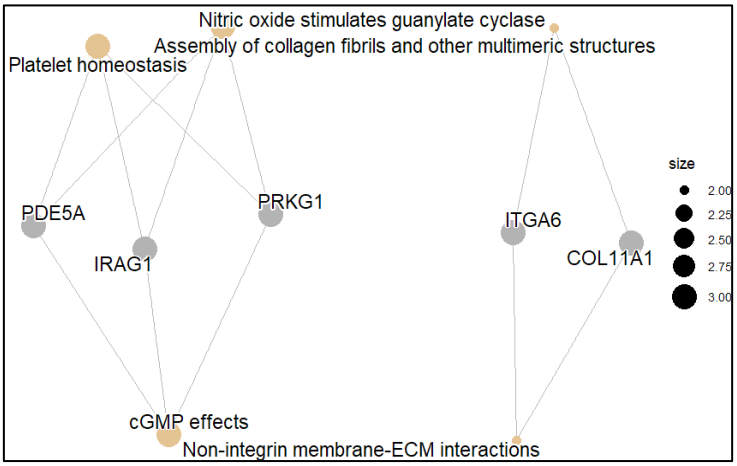
Figura 9. *BarPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con DESeq2*



Nota: El número de genes se representan en el eje X con un significado ajustado (p<0.03) indicado por orden y tendencia de color. Fuente: Autores.

Según la Figura 10, se determina que tres genes (PRKG1, PDE5A y IRAG1) se asocian con las vías de ensamblaje de fibrillas de colágeno y otras estructuras multiméricas, la homeostasis plaquetaria y los efectos de cGMP. Por otro lado, la vía del óxido nítrico que estimula el guanilato ciclasa y las interacciones membrana-ECM no mediadas por integrinas se relacionan con los genes: ITGA6 y COL11A1.

Figura 10. *CnetPlot de los genes sobreexpresados con sus vías de señalización determinado con DESeq2*



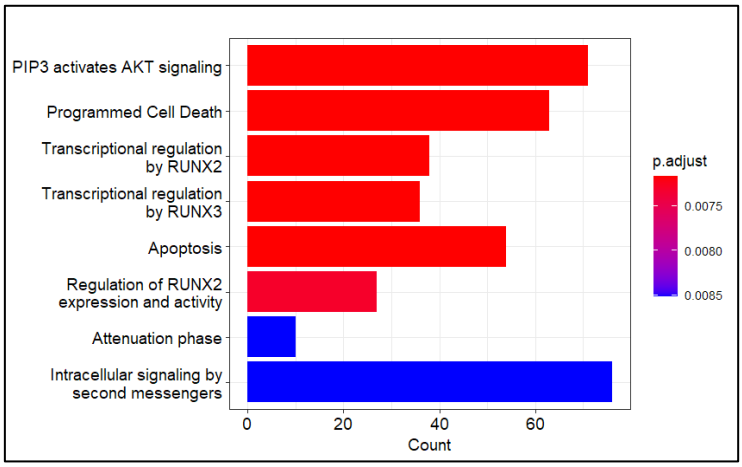
Nota: Los nodos grises representan los genes sobreexpresados y los nodos dorados representan las vías señalización biológica, donde el tamaño del nodo es proporcional al número de genes asociados en cada camino.

Fuente: Autores.

4.2.2.2. Resultado de genes infraexpresados de DESeq2 con GSEA

En base al análisis de enriquecimiento para genes infraexpresados después de emplear el DESeq2, se evidenciaron 8 categorías a nivel de proceso biológico. PIP3 activa la señalización de AKT, muerte celular programada o apoptosis, la cantidad de genes infraexpresados no se deben al azar y mantiene un p valor menor a 0.0075. Sin embargo, en las vías de regulación transcripcional por RUNX2, regulación transcripcional por RUNX3 y regulación de la expresión y actividad de RUNX2, se establecen con un p valor menor a 0.0075. Por otro lado, tanto para la fase de atenuación y señalización intracelular por segundos mensajeros se mantiene un p valor menor a 0.0085 teniendo en cuenta que para la primera vía hay menor cantidad de genes en comparación a la segunda vía que tiene una gran cantidad de genes -Figura 11-.

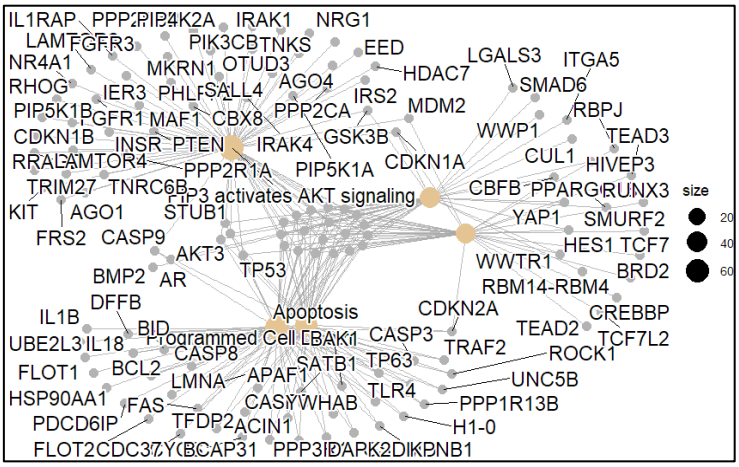
Figura 11. *BarPlot de los genes infraexpresados según p value y sus vías de señalización determinado con DESeq2*



Nota: El número de genes se representan en el eje X con un significado ajustado (p<0.0085) indicado por orden y tendencia de color. Fuente: Autores.

Según la Figura 12, se reveló que los genes diferencialmente expresados están asociados con 5 vías de señalización (PIP3 activa la señalización de AKT, apoptosis y muerte celular programada) y varios genes.

Figura 12. *CnetPlot de los genes infraexpresados con sus vías de señalización determinado con DESeq2*

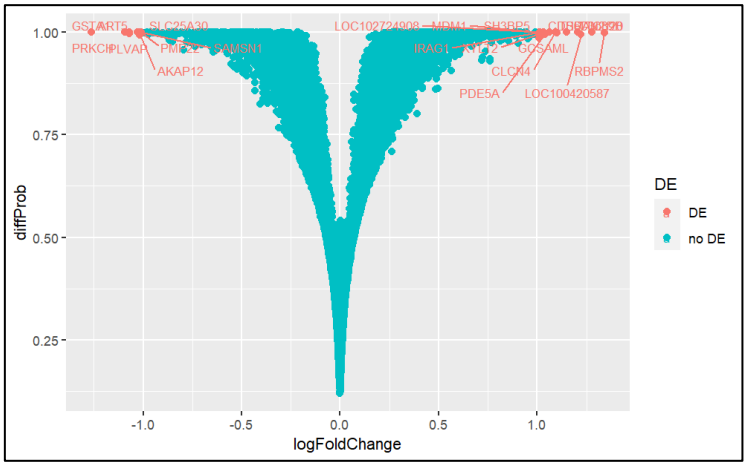


Nota: Los nodos grises representan los genes infraexpresados y los nodos dorados representan las vías señalización biológica, donde el tamaño del nodo es proporcional al número de genes asociados en cada camino. Fuente: Autores.

4.2.3. Resultados del análisis de expresión diferencial con BADER

El análisis de expresión diferencial utilizando BADER reveló un total de 23 genes diferencialmente expresados entre las dos condiciones estudiadas, WT y RUNX1-JAK2. Para visualizar los resultados se generó un gráfico de *Volcano Plot* -Figura 13-, en el que cada punto representa un gen, de los cuales los genes diferencialmente expresados se representaron de color rojo y los genes con menor significancia estadística de color azul.

Figura 13. *Volcano Plot del análisis de expresión diferencial realizada con BADER*



Nota: DE=Genes diferencialmente expresados, NO DE=Genes que no son diferencialmente expresados. Fuente: Autores.

Por otro lado, de la Figura 13, el eje x (*LogFoldChange*) muestra el cambio de expresión del gen entre las condiciones, >1, y el eje y (diffProb o -Log10 (padj)) demuestra la significancia estadística de diferencia observada > 0.95. Demostrando que los genes con sobreexpresión (>1) son 14 -Tabla 5-: TSPAN2, IRAG1, GCSAML, RBPMS2, LOC100420587, PDE5A, XYLT2, SH3BP5, LOC102724908, MDMD, C3orf20, TTC39B, CDH6 y CLCN4. Y los genes con infraexpresión son 8 -Tabla 5-: AKAP12, PRKCH, GSTA1, ART5, PLVAP, SLC25A30, PMP22 y SAMSN1.

Tabla 5. *Nombre de los genes diferencialmente expresados obtenidos por BADER*

Nº DE GENES	DE	NOMBRE DE LOS GENES
1	DE SOBREEXPRESADOS	TSPAN2
2		GCSAML
3		NA
4		C3orf20
5		SH3BP5
6		PDE5A
7		CDH6
8		CLCN4
9		TTC39B
10		IRAG1
11		MDM1
12		RBPMS2
13		XYLT2
14		LOC100420587
15		LOC102724908
1	DE INFRAEXPRESADOS	GSTA1
2		AKAP12
3		ART5
4		SLC25A30
5		PRKCH
6		PMP22
7		PLVAP
8		SAMSN1

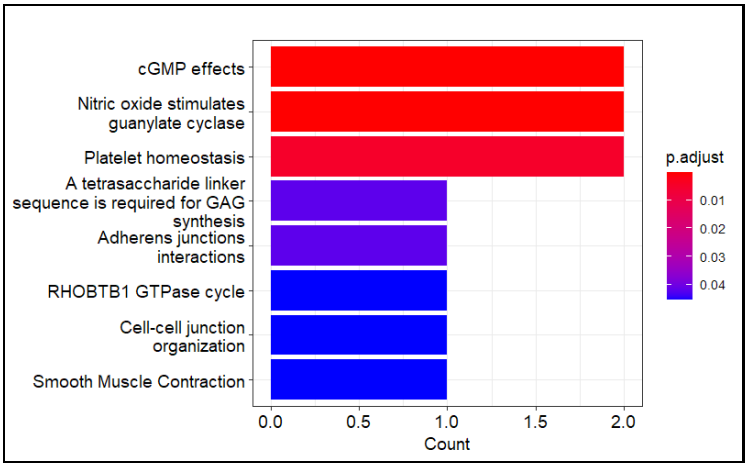
Fuente: Autores.

4.2.3.1. Resultados de genes sobreexpresados de BADER con GSEA

De acuerdo con el análisis de enriquecimiento para genes sobreexpresados después de emplear BADER, se evidenciaron 9 categorías a nivel de proceso biológico. Los efectos del cGMP, óxido nítrico que estimula el guanilato ciclasa y la homeostasis plaquetaria contienen el rango con mayor cantidad de genes sobreexpresados con un p por debajo de 0.01. Otros procesos como la secuencia de análisis de enlace tetrasacárido para la síntesis de GAC y las interacciones de las uniones inherentes, contienen una cantidad moderada de genes con un p valor de más o menos 0.04. Sin embargo, en el caso del ciclo de GTPasa RHOBTB1, la

organización de las células intercelulares y la contracción del músculo liso se visualizan con una cantidad moderada de genes, pero con un valor de p de más o menos 0.05 -Figura 14-.

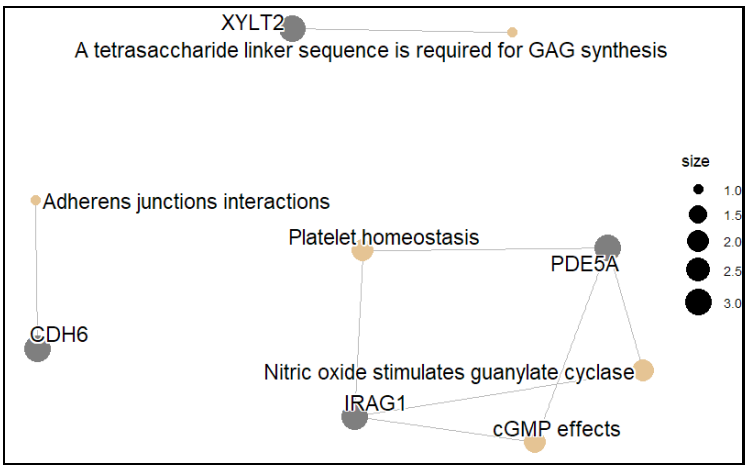
Figura 14. *BarPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con BADER*



Nota: El número de genes se representan en el eje X con un significado ajustado ($p<0.05$) indicado por orden y tendencia de color. Fuente: Autores.

Según la Figura 8, se reveló que los genes diferencialmente expresados están significativamente enriquecidos con complejidades potencialmente biológicas, está asociación determina que un gen (CDH6) asociado con la vía de interacciones de las uniones adherentes, un gen (XYLT2) asociado con la vía de secuenciación del enlace tetrasacárido para la síntesis de GAG. Por otro lado, tanto el gen PDE5A e IRAG1 se asocian con 3 vías biológicas como la homeostasis plaquetaria, el óxido nítrico que estimula el guanilato ciclasa y los efectos del cGMP.

Figura 15. *CnetPlot de los genes sobreexpresados según p value y sus vías de señalización determinado con BADER*



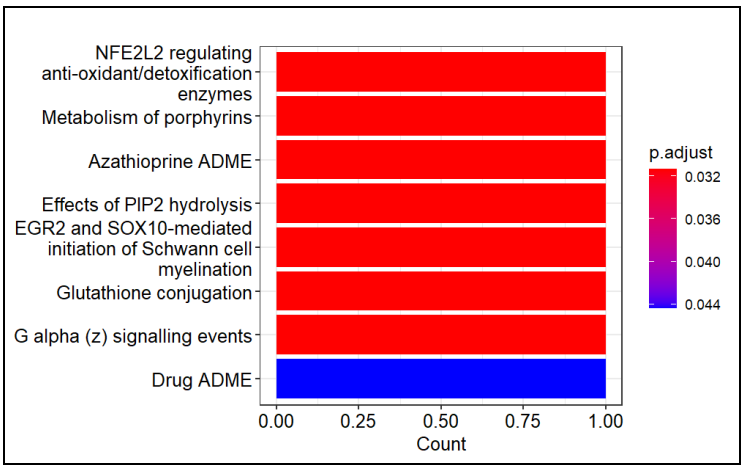
Nota: Los nodos grises representan los genes sobreexpresados y los nodos dorados representan las vías señalización biológica, donde el tamaño del nodo es proporcional al número de genes asociados en cada camino.

Fuente: Autores.

4.2.3.2. Resultados de genes infraexpresados de BADER con GSEA

Se determinan diferentes vías de señalización, siendo un total de 8 vías, donde 7 se determinan con un p valor por debajo del 0.036, que tienen una cantidad alta de genes, los cuales son: NFE2L2 regulando enzimas antioxidantes y de detoxificación, metabolismo de la porfiria, ADME de azatioprina, efectos de la hidrólisis de P1P2, iniciación mediada por EGR2 y SOX10 de la mielinización de las células de Schwann, conjugación del glutatión y eventos de señalización de G alfa "z". Por otro lado, se determina una vía de señalización "*ADME* de medicamentos" con un valor de p de más o menos 0.04, lo que indica que tiene una menor cantidad de genes infraexpresados. -Figura 16-.

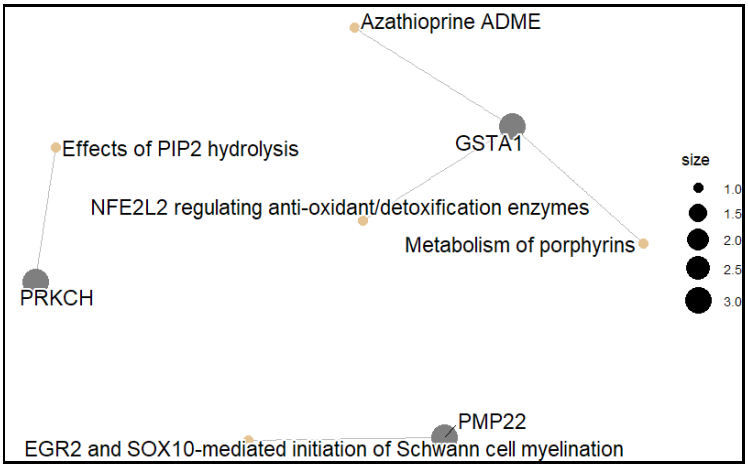
Figura 16. *BarPlot de los genes infraexpresados según p value y sus vías de señalización determinado con BADER*



Nota: El número de genes se representan en el eje X con un significado ajustado ($p<0.05$) indicado por orden y tendencia de color. Fuente: Autores.

Según la Figura 17, se reveló que los genes diferencialmente expresados están significativamente enriquecidos con complejidades potencialmente biológicas, está asociación determina que un gen (PRKCH) se relaciona con los efectos de la hidrólisis P1P2, un gen (GSTA1) está relacionado con *ADME* "Absorción, Distribución, Metabolismo y Excreción" de azatioprina, NFE2L2 regulando enzimas antioxidantes y el metabolismo de la porfiria, por último, un gen (PMP22) se relaciona con la iniciación mediada por EGR2 y SOX10 de la mielinización de las células de Schwann.

Figura 17. *CnetPlot de los genes infraexpresados con sus vías de señalización determinado con BADER*



Nota: Los nodos grises representan los genes infraexpresados y los nodos dorados representan las vías señalización biológica, donde el tamaño del nodo es proporcional al número de genes asociados en cada camino.

Fuente: Autores.

4.3. Discusión

En primera instancia, se determinó que al utilizar tanto DESeq2 como BADER, se obtienen diferentes cantidades de genes diferencialmente expresados, debido a las distintas aproximaciones, ya que, DESeq2 utiliza métodos empíricos, mientras que BADER utiliza un método jerárquico (Sánchez, 2015). Por lo que, BADER, dispone, de ventajas potenciales como son que la precisión de las estimaciones es mayormente aceptable debido a que emplea información previa o "*a priori*". Además de que estos datos pueden compartir particularidades entre los diversos niveles de jerarquía lo cual es beneficioso para la correlación de los genes (Alcalde, 2022). Caso que no sucede con DESeq2, ya que este al manejarse con una distribución binomial negativa, modela los conteos de los genes de acuerdo con la estimación de la sobredispersión. Por ende, si se observa una gráfica de estos resultados, los genes se verifican de forma desorganizada y son expuestos en base a la cantidad de su expresión. Es decir, si un gen en primera posición tiene una menor cantidad que otro gen en segunda posición con mayor cantidad, la gráfica de interpretación de los resultados se vería similar a la curva de una campana. Todo lo contrario, a BADER, en donde su caracterización se maneja de acuerdo

con los niveles de mayor expresión se guían de un orden hasta llegar a uno de menor manifestación.

Por otro lado, dado que, con el análisis de expresión diferencial mediante DESeq2 se determinaron 60 genes diferencialmente expresados, en cambio con el modelo de BADER se determinaron 22 genes. De los cuales, los infraexpresados identificados tanto con DESeq2 como BADER fueron: AKAP12, PRKCH y PLVAP. Y los genes sobreexpresados fueron: TSPAN2, GCSAML, RBPMS2, LOC100420587, PDE5A, LOC102724908, C3orf20, TTC39B y CLCN4. A pesar de que se emplearon métodos diferentes, la cantidad total de genes que se expresaron diferencialmente con cada paquete, le dan más relevancia a estos genes, que no fueron comentados en el contexto de estudio. Es por ello, la razón de que BADER ofrece una mayor probabilidad de los resultados obtenidos, debido a que considera la variación biológica y tecnológica, mientras que DESeq2 conserva la ideología de un panorama con mayor cantidad de genes. Pues no es lo mismo cantidad que calidad (Neudecker y Katzfuss, 2023).

Desde otro punto de vista, según la variación génica y las vías de señalización determinados con GSEA para los genes sobreexpresados, se tiene que con DESeq2 se determinaron las ***vías de efectos de cGMP***, el ***óxido nítrico estimula la guanilato ciclasa***, ***homeostasis plaquetaria***, ensamblaje de fibrillas de colágeno y otras estructuras multiméricas, interacciones membrana-ECM no mediadas por integrinas, formación del colágeno y la hemostasia para genes sobreexpresados, sin embargo, con BADER se determinaron efectos de ***cGMP, el óxido nítrico estimula la guanilato ciclasa***, ***homeostasis plaquetaria***, secuencia de enlace de tetrasacárido para la síntesis de *GAG,* interacciones de las uniones adherentes, ciclo de GTPasa RHOBTB1, organización de las uniones intercelulares y contracción del músculo liso. De los procesos anteriormente mencionados y a pesar de que no se conoce del todo la lógica que envuelve al desarrollo de la *LLA-B*, en ambas metodologías, procesos que se repiten

como la ***homeostasis plaquetaria, el óxido nítrico que estimula el guanilato ciclasa y los efectos de cGMP***, son factores que pueden ***afectar la estabilidad de las células en la línea hematopoyética*** y producir o empeorar la LLA-B. Además de que, se identifica inmiscuido al gen *PRKG1*, como un principal responsable de estas fallas. Por lo tanto, estás vías revelan que podrían estar relacionadas con alteraciones en la estructura y la función de las células leucémicas, en cuanto a la proliferación celular, diferenciación y apoptosis (Tello y Novoa, 2020).

En cambio, en el análisis de los genes infraexpresados con DESeq2, se encontraron genes menos activos que están vinculados con varias funciones celulares, como la señalización celular, el crecimiento y desarrollo de ciertas células, y la muerte celular programada. Entre estos genes, se destacan *RUNX2* y *RUNX3*, que son importantes para el desarrollo de ciertas células, incluyendo las células B, que son un tipo de células del sistema inmunológico (Sánchez, 2018). Por otro lado, utilizando *BADER*, se encontraron genes relacionados con la ***eliminación de sustancias dañinas del cuerpo***, ***el metabolismo de ciertos compuestos, y la formación de la mielina en células específicas del sistema nervioso.*** Entre estos, se destacan genes como ***NFE2L2***, que están involucrados en la ***protección celular contra el estrés oxidativo*** y pueden ***influir en la resistencia a la quimioterapia***, un tratamiento común para el cáncer (Carballar del Valle, 2022). Por consiguiente, con DESeq2 se determina una variabilidad de genes relacionados con las vías de señalización y con BADER solo se determinaron los genes ***PRKCH, GSTA1 y PMP22***.

En este contexto, se observa que DESeq2 identificó una mayor variedad de genes relacionados con las vías de señalización en comparación con BADER. Esto puede sugerir que DESeq2 es más sensible en la detección de cambios en la expresión génica. Sin embargo, es importante destacar que algunos genes específicos fueron identificados sólo por BADER (***PRKG1***, ***PRKCH***, ***GSTA1***, ***PMP22***, etc.), lo que indica que este método también puede aportar

información valiosa, ya que estos genes están involucrados en procesos biológicos, tales como son: la regulación de la función plaquetaria por ***PRKG1***, codificación a proteínas quinasa por ***PRKCH*** y procesos de desintoxicación celular con ***GSTA1***, mismos que no fueron mencionados en el estudio base.

Por consiguiente, según Fortschegger et al. (2021) la variabilidad de la LLA-B según RUNX1-JAK2 está guiado por otros genes que regulan las vías de ***JAK-STAT*** y ***MYC***, lo que determina que el análisis de expresión diferencial de DESeq2 y BADER determinan una alterabilidad de la que relacionan con otros genes según la significancia de los mismos, lo que puede sugerir que se debe al enfoque dado en el análisis de la expresión génica y la variabilidad de la cantidad de genes normalizados y filtrados.

Finalmente y a modo de exploración, a pesar de que los dos softwares disponibles tanto DESeq2 y BADER son adecuados para el análisis de la expresión diferencial de los datos de RNA-seq y que generan los resultados de las vías de enriquecimiento para genes, resultados del estudio, sin embargo, con este trabajo, se puede dejar la tentativa presunción de que BADER, no solo analiza a los genes inmiscuidos en la patología, sino que también otro tipo de procesos que podrían estar alimentando a la mutación, sin imaginar que existe la posibilidad del caso. En este contexto, podemos mencionar a uno de los mecanismos, el de contracción del músculo liso, que no se identificó con ningún gen establecido. Desde esta perspectiva, podemos dejar como un dato de estudio para próximas investigaciones, que se pueda dar la posibilidad de que las neoplasias, no solo tengan que relacionarse con parámetros propios del caso, sino quizás con otros que no parezcan tener vínculo alguno.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En conclusión, las implicaciones más importantes de los resultados fueron la revelación de genes de interés por ambos métodos, DESeq2 y BADER. Estos genes, como PRKG1, PRKCH, GSTA1, PMP22, entre otros, podrían estar implicados en la patogénesis de la LLA-B y podrían ser objetivos potenciales para la investigación futura y el desarrollo de nuevas terapias. Además, se obtuvo información sobre vías de señalización alteradas como la homeostasis plaquetaria, el ensamblaje de fibrillas de colágeno y la muerte celular programada se relacionan con la LLA-B. Lo cual, esta información puede ser útil para comprender la biología de la enfermedad y orientar la búsqueda de nuevas terapias.

Además, los resultados sugieren que puede ser necesario replantearse cómo se lleva a cabo el análisis de la expresión génica. Ambos métodos de análisis (DESeq2 y BADER) aportaron información valiosa, pero también mostraron diferencias notables en los genes y las vías de señalización que identificaron. Esto sugiere que puede ser beneficioso utilizar múltiples métodos de análisis para obtener una visión más completa de la expresión génica.

Estos resultados también tienen implicaciones prácticas. La identificación de genes y vías de señalización alteradas puede ayudar a guiar la búsqueda de nuevas terapias para la LLA-B. También podría ser útil para identificar biomarcadores para el diagnóstico y la predicción del pronóstico.

En cuanto a la justificación de los motivos, se ha explicado claramente la necesidad de utilizar múltiples métodos de análisis para obtener una visión más completa de la expresión génica. En términos de qué se debe hacer, quién debe hacerlo y cómo, estos resultados sugieren que los investigadores en el campo de la genómica y la biología del cáncer deberían considerar la utilización de múltiples métodos de análisis en sus investigaciones.

5.2. Recomendaciones

- Para el análisis de expresión diferencial se debe tener en cuenta el número de muestras para el análisis de las condiciones, además la probabilidad y la distribución que maneja cada método.
- Se debe considerar que para emplear un paquete en el análisis de expresión diferencial las probabilidades de los datos y cuales tienen menor cantidad de falsos positivos, denominado método *False Discovery Rate* (FDR).
- Se debe considerar la simulación de los datos para evaluar el método adecuado.

REFERENCIAS BIBLIOGRÁFICAS

Alarcón, M. (2019). *Análisis conjunto mediante modelos lineales jerárquicos y modelos jerárquicos Bayesianos. Una aproximación mediante análisis multivariado* [Tesis de Grado, Universidad Santo Tomas]. Primer Claustro Universitario de Colombia.

Alcalde, M. (2022). *Modelos Jerárquicos Bayesianos* [Tesis de grado, Universidad de Zaragoza].

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., y Love, M. I. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature methods, 17*(3), 137-145. https://doi.org/10.1038/s41592-019-0654-x

Arias, J., y Muñoz, J. (2019). *Métodos alternativos para evaluar expresión diferencial sin réplicas de los tratamientos de materiales Rubus glaucus Benth tolerantes al ataque de Colletrotrichim gloesporiodes con el fin de identificar genes de importancia asociados a tolerancia* [Tesis de Maestría, Universidad Tecnológica de Pereira].

Baptiste, A., y Anton, A. (2022). *Miscellaneous Functions for "Grid" Graphics.* CRAN. https://bit.ly/3rms4nz

Bioconductor. (2023). *Bioconductor open source software for bioinformatics.* https://bit.ly/3OjwEMy

Burgos, A. (2021). *Análisis de datos NGS para la determinación de nuevos factores moleculares implicados en la adrenoleucodistrofia* [Tesis de maestría, Universitat Oberta de Catalunya].

Canzoneri, R., Lacunza, E., y Abba, M. (2019). Genómica y bioinformática como pilares de la medicina de precisión en oncología. *Medicina (Buenos Aires), 79* (6), 587-592.

Carlson, M. (2019). *org.Hs.eg.db: Genome wide annotation for Human.* R package version 3.8.2. https://doi.org/doi:10.18129/B9.bioc.org.Hs.eg.db

Carballar del Valle, R. (2022). *Estrategia terapéutica para el carcinoma de células escamosas de pulmón basada en la generación de ROS por anisomicina* [Tesis de pregrado, Universidad Politécnica de Madrid].

Castellanos, O. A., y Melo, M. X. (2021). *Estudio de la prefactibilidad para la creación de una maestría profesional en bioinformática en la Universidad Central del Ecuador, con el desarrollo de un portal web multiplataforma, durante el periodo 2020-2021* [Tesis de pregrado, Universidad Central del Ecuador]. https://bit.ly/3KRN9O0

Castañeda, P. (2021). *Análisis de metodologías estadísticas en RNA-seq, con aplicación a cáncer de pulmón* [Tesis de grado, Universidad Nacional de Colombia].

Chamorro, C. (2019). *Análisis de datos de RNA-seq empleando diferentes paquetes de Bioconductor para estudios de expresión* [Tesis de Maestría, Universidad Oberta De Catalunya].

Diaz, R. (2019). Métodos Bayesianos para modelos ocultos de Markov en series de tiempo con conteo [Tesis de maestría, Universidad Nacional de Colombia].

Díez, J. (2022). *Determinación del perfil de expresión génica inducido por la dexametasona mediante una revisión sistemática con metaanálisis* [Tesis de pregrado, Universidad de Salamanca].

Emadi, A., y York, J. (2022). Leucemia Linfoblástica Aguda. Manual MSD. https://msdmnls.co/3q5Fioq

Fernández, J. (2022). *Desarrollo del prototipo de sistema web para la gestión de datos de información genética (bioinformática)* [Tesis de pregrado, Universidad Católica del Ecuador]. https://bit.ly/40gy8dj

Ferrer, A. (2018). *Visualización de la calidad de los datos de RNA-Seq relacionados con el sistema inmune y el cáncer visualizados mediante shiny* [Tesis de maestría, Universidad Oberta de Catalunya].

Fortschegger, K., Husa, A. M., Schinnerl, D., Nebral, K., y Strehl, S. (2021). Expression of RUNX1-JAK2 in Human Induced Pluripotent Stem Cell-Derived Hematopoietic Cells Activates the JAK-STAT and MYC Pathways. *International Journal of Molecular Sciences, 22* (14), 7576. https://doi.org/10.3390/ijms22147576

Gallego, A. (2021). *Uso de paquetes de R/Bioconductor para análisis funcional de datos ChIP-Seq* [Tesis de Maestría, Universitat Oberta de Catalunya].

Garaventa, M. (2018). *Leucemia Mieloide Crónica. IBC Laboratorios.* https://bit.ly/3Oxq5WE

Gil, C. (2018). *Análisis de Componentes Principales (PCA).* RStudio Pubs. https://bit.ly/3O6hLMh

Gil, C. (2020). *Análisis de Datos scRNA-Seq con Bioconductor*. RStudio Pubs. https://bit.ly/3rDB5bY

Guo, W., Wang, D., Wang, S., Shan, Y., Liu, C., y Gu, J. (2021). Cancer: a package for automated processing of single-cell RNA-seq data in cancer. *Brief Bioinform, 22(*3). https://doi.org/10.1093/bib/bbaa127

Hernández, M. (2021). *Análisis de expresión diferencial en células PBMC de pacientes sanos vs. COVID-19* [Tesis de Grado, Universitat Rovira i Virgili].

Hong, M., Tao, S., Zhang, L., Diao, L., Huang, X., Huang, S., Xie, S., Xiao, Z., y Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology*, *13*(1). https://doi.org/10.1186/s13045-020-01005-x

Husson, F., Josse, J., Le, S., y Mazet, J. (2023). *FactoMineR: Multivariate exploratory data analysis and data mining.* CRAN. https://bit.ly/3pEWgdl

Instituto de Salud Carlos III. (2020). *¿Qué es la bioinformática y qué aplicaciones tiene en biomedicina?*. Divulgación ISCIII. http://bit.ly/3KRQZXq

Instituto Nacional del Cáncer. (2022). *Tratamiento de la leucemia linfoblástica aguda infantil.* https://bit.ly/3KGZsve

Jiménez, J. (2019). *Introducción a R y RStudio* [Universidad Tecnológica de Panamá].

Jiménez-Jiménez, V., Martí-Gómez, C., Del Pozo, M. A., Lara-Pezzi, E., y Sánchez-Cabo, F. (2021). Bayesian Inference of Gene Expression. *Exon Publications eBooks*. https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch5

Khalfan, M. (2021). *Gene Set Enrichment Analysis with ClusterProfiler.* NGS Analysis. https://bit.ly/44FRKdC

Kassambara, A., y Mundt, F. (2020). *factoextra: extract and visualize the results of multivariate Data analyses*. CRAN. https://bit.ly/3pO7lc4

Kellman, B. P., Baghdassarian, H. M., Pramparo, T., Shamie, I., Gazestani, V. H., Begzati, A., Li, S., Nalabolu, S., Murray, S. S., Lopez, L. C., Pierce, K., Courchesne, E., y Lewis, N. S. (2021). Multiple freeze-thaw cycles lead to a loss of consistency in poly(A)-enriched RNA sequencing. *BMC Genomics*, *22*(1). https://doi.org/10.1186/s12864-021-07381-z

Layton, C. (2015). Factores de pronóstico en leucemia linfoblástica aguda pediátrica: posibles marcadores moleculares. *Revista de medicina e investigación*, *3*(1), 85-91. https://doi.org/10.1016/j.mei.2015.02.008

Lee, J., Ji, Y., Liang, S., Cai, G., y Müller, P. E. (2015). Bayesian Hierarchical Model for Differential Gene Expression Using RNA-Seq Data. *Statistics in Biosciences*, *7*(1), 48-67. https://doi.org/10.1007/s12561-013-9096-7

Lemes, V., Gutiérrez, M., Fernández, F., y Riesco, S. (2022). Leucemia linfoblástica aguda tipo B. La importancia de un diagnóstico precoz. *Nuevo Hosp,* (*18*)3, 40-44. https://bit.ly/3O6OH8v

Limma Law, C. W., Chen, YShi, W., y Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology, 15*(2). https://doi.org/10.1186/gb-2014-15-2-r29

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., y Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: LIMMA, EDGER, DESEQ2. *Journal of Visualized Experiments*, *175*. https://doi.org/10.3791/62528

Love, M. I., Anders, S., y Huber, W. (2019). RNA-Seq workflow: gene-level exploratory analysis and differential expression. 8, 1-10. https://doi.org/10.12688/f1000research.16684.2

Love, M., Huber, W., y Anders, S. (2023). *Analyzing RNA-seq data with DESeq2.* https://bit.ly/3K0kINe

Marco-Puche, G., Lois, S., Benitez, J., y Triviño, J. C. (2019). RNA-Seq Perspectives to Improve Clinical Diagnosis. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01152

Masip, D. (2019). *Desarrollo de un pipeline Bioinformático mediante R: Análisis basados en panel de cáncer de pulmón* [Tesis de maestría, Universitat Oberta de Catalunya].

Miao, Z., Moreno-Romero, J., y Eswaran, G. (2021). TissueClear: a next-generation tissue-clearing protocol for plant biological research. *Plant Physiol, 186*(1), 403-416. https://doi.org/10.1093/plphys/kiaa106

Morgan, M., Twisk, D., y Cheng, Y. (2023). *dplyr-base Access to Bioconductor Annotation Resources.* Bioconductor. https://bit.ly/3Db2rJ7

Navarrete, M., y Pérez, P. (2017). Alteraciones epigenéticas en leucemia linfoblástica aguda. *Boletín médico del Hospital Infantil de México*, *74*(4), 243-264. https://doi.org/10.1016/j.bmhimx.2017.02.005

Neudecker, A., y Katzfuss, M. (2023). *Bayesian Analysis of Differential Expression in RNA Sequencing Data.* Bioconductor. https://bit.ly/3rxaZY1

Nonell, L. (2019). *New approaches in omics data modelling* [Tesis doctoral, Universitat Pompeu Fabra]. https://bit.ly/41g0lSJ

Onkopedia. (2022). *Leucemia Linfoblástica aguda (LLA).* Sociedad Alemana de Hematología y Oncología. https://bit.ly/3Ye2nSD

Ortiz, Í. (2018). *Inferencia Bayesiana* [Tesis de pregrado, Universidad de Sevilla]. https://bit.ly/43Mnlcn

Owens, N. D. L., De Domenico, E., y Gilchrist, M. J. (2019). An RNA-Seq Protocol for Differential Expression Analysis. *Cold Spring Harbor protocols*, *2019*(6), 10.1101/pdb. prot098368. https://doi.org/10.1101/pdb.prot098368

Pagés, H., Carlson, M., Falcon, S., y Li, N. (2023). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. https://bit.ly/457chaV

Piñero, M. (2017). *Differential expresión analysis of TYLCV infected tomato genes* [Tesis de Grado, Universidad de Málaga].

Risso, D., Sales, G., Romualdi, C., y Chiogna, M. (2012). A Hierarchical Bayesian Model for RNA-Seq Data. *Springer eBooks,* 215-227. https://doi.org/10.1007/978-88-470-2871-5_17

Rodríguez, G., y Shishkova, S. (2019). Estudio del transcriptoma mediante RNA-seq con énfasis en las especies vegetales no modelo. *Revista de Educación Bioquímica*, *37*(3), 75-88. https://bit.ly/43NObkz

Robinson, M. D., McCarthy, D. J., y Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Sánchez, S. *Análisis de datos de RNA-seq: comparación de métodos para el estudio de expresión génica diferencial* [Tesis de pregrado, Universidad de Sevilla].

Sancho, R. S. (s.f.). *Paquetes-Ciencia de datos con R.* https://bit.ly/45alwr1

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., y Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA (New York, N.Y.)*, *22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Slowikowski, K. (2023). *Getting started with ggrepel.* CRAN. https://bit.ly/43NJXcn

Soriano, B. (2023). *Soluciones bioinformáticas para el análisis de datos ómicos, descubrimiento de conocimiento y diagnóstico genético en Sparus aurata y otros organismos biológicos* [Tesis doctoral, Universitat de Valencia].

Stupnikov, A., McInerney, C. E., Savage, K. I., McIntosh, S., Emmert-Streib, F., Kennedy, R. D., Salto-Tellez, M., Prise, K. M., y McArt, D. G. (2021). Robustness of differential gene expression analysis of RNA-SEq. *Computational and structural biotechnology journal*, *19*, 3470-3481. https://doi.org/10.1016/j.csbj.2021.05.040

Sundaram, A., Tengs, T., y Grimholt, U. (2017). Issues with RNA-seq analysis in non-model organisms: A salmonid example. *Developmental & Comparative Immunology, 75*, 38–47. https://doi.org/10.1016/j.dci.2017.02.006

Owens, N. D. L., De Domenico, E., y Gilchrist, M. J. (2019). An RNA-Seq Protocol for Differential Expression Analysis. *Cold Spring Harbor protocols*, *2019*(6), 10.1101/pdb. prot098368. https://doi.org/10.1101/pdb.prot098368

Piñero, M. (2017). *Differential expresión análisis of TYLCV infected tomato genes* [Tesis de Grado, Universidad de Málaga].

Tello, S., y Novoa, K. (2020). Leucemia-linfoma linfoblástico de células precursoras B con eosinofilia severa, en una paciente adulta joven. *Revista del cuerpo médico del HNAAA, 12*(4), 337-339. https://doi.org/10.35434/rcmhnaaa.2019.124.568

Universidad de las Palmas de Gran Canaria. (s.f.). *Análisis Bayesiano para datos de Poisson. https://bit.ly/3YeqaBU*

Valdespino-Gómez, V., Valdespino-Castillo, P., y Valdespino, C. (2013). Organización estructural y funcional del genoma humano: variación en el número de copias predisponentes de enfermedades degenerativas. *Gaceta Mexicana de Oncología, 12* (6). https://bit.ly/3DCWWmS

Vardhanabhuti, S., Li, M., y Li, H. (2013). A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-sample RNA-Seq Data. *Statistics in Biosciences*, *5*(1), 119-137. https://doi.org/10.1007/s12561-011-9052-3

Vélez, J. I., y Correa, J. A. (2013). Cuantificación de variantes genéticas utilizando modelos jerárquicos bayesianos. *Comunicaciones en Estadística, 6*(1), 59-73.

Vestal, B., Moore, C. M., Wynn, E. H., Saba, L., Fingerlin, T. E., Y Kechris, K. (2020). MCMSeq: Bayesian hierarchical modeling of clustered and repeated measures RNA sequencing experiments. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-03715-y

Wickham, H., Chang, W., Henry, L., Lin, T., Takashi, K., Wilke, C., Woo, K., Yutani, H. y Dunnington, D. (2023). ggplot2: create elegant data visualisations using the Grammar of Graphics. CRAN. https://bit.ly/3Ob6MCp

Whitney, T., Robertson, C., Sharkey, K., Torble, T., Lagwankar, S., Toliver, K., Blome, M., Jones, M., Hohenson, G., y Cai, S. (2023). *Documentación del Lenguaje C++. Microsoft.* https://bit.ly/3DUuop7

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu x, Liu S, Bo X., y Yu G (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2 (3), 1-10. https://doi.org/10.106/j.xinn.2021.100141

Yang, X., Zhang, S., He, S., Xiong, Q., Wang, Y., y Sun, Z. (2020). Comparative evaluation of RNA-seq analysis methods for degraded or low-input samples. *Bioinformatics, 36(*7), 2084-2090. https://doi.org/10.1093/bioinformatics/btz862

Yu, X., Li, M., Guo, C., Wu, Y., Zhao, L., Shi, Q., Song, J., y Song, B. (2021). Therapeutic targeting of cancer: epigenetic homeostasis. *Frontiers in Oncology, 11*. https://doi.org/10.3389/fonc.2021.747022

Yu, G., y Petyuk, V. (2023). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, *12*(12), 477-479. https://doi.org/10.1039/C5MB00663E

APÉNDICES Y ANEXOS

Anexo 1. *Librerías utilizadas en R*

LIBRERÍA	VERSIÓN	DESCRIPCIÓN	REFERENCIA
AnnotationDbi	1.62.2	Manipulación de anotaciones basadas en SQLite, que posee el nivel de: organismo, plataforma y sistema biológico, en Bioconductor.	Pagés et al., 2023
BADER	1.38.0	Análisis de expresión diferencial para datos de secuenciación de RNA.	Neudecker y Katzfuss, 2023
clusterProfiler	4.8.2	Herramienta de enriquecimiento universal para interpretar datos ómicos.	Wu et al., 2021
DESeq2	1.40.2	Expresión diferencial de los genes basado en al análisis de la distribución binomial negativa,	Love et al, 2023
ggplot2	3.4.2	Gráficos de datos, detalles de forma estética (aesthetic mappings) como color exterior, relleno, formas de puntos, tipo de línea y tamaño.	Wickham et al., 2023
dplyr	1.1.2	Maneja la manipulación de datos que facilita la transformación, filtrado, ordenamiento y el resumen de estos.	Morgan et al., 2023
factoextra	1.0.7	Funciones para extraer de forma fácil y visualizar la salida de datos multivariados incluidos de PCA, MCA y HMFA.	Kassambara y Mundt, 2020
FactoMineR	2.8	Permite análisis factorial que reduce la dimensionalidad de variables, puede efectuar PCA, relación entre variables categóricas y análisis de componentes múltiples.	Husson et al., 2023
ggrepel	0.9.3	Etiquetas en la posición de texto no superpuesto, de tipo ggplot2.	Slowikowski, 2023
gridExtra	2.3	Organización flexible y combinación múltiple de gráficos. Favorece la exportación de figuras en diversos formatos.	Baptiste y Anton, 2022
org.Hs.eg.db	3.17.0	Anota el genoma humano, basado en Gene IDs de Entrez Gene	Carlson, 2019
ReactomePA	1.44.0	Analiza rustas de datos mediante el análisis de enriquecimiento de conjunto de genes, análisis de enriquecimiento. Además, permite la visualización de gráficos.	Yu y Petyuk, 2023

Fuente: Autores.

Anexo 2. *Código en R*

Preparación de los Datos

Chunk 1. Instalación de las librerías en el programa *R*

```r
{r}
#Graficos estadisticos
library(ggplot2)
#Manipular y transformas datos
library(dplyr)
#Analisis de datos de expresion genica
library(DESeq2)
#Combinacion de graficos
library(gridExtra)
#Herramientas para PCA, MCA, HCA
library(FactoMineR)
#Paquetes de Bioconductor
library(BiocManager)
#Visualizacion de PCA, MCA, HCA
library(factoextra)
#Anotaciones genes tipo SQLite
library(AnnotationDbi)
#Análisis de modelo Jerárquico bayesiano
library(BADER)
#Etiquetas de datos en gráficos
library(ggrepel)
#Interpretación de los datos ómicos
library(clusterProfiler)
```

Chunk 2. Creación del Objeto *DESeq2DataSet*

```r
{r}
#Lectura de datos
contajes.raw <- read.table("./datos/GSE159261_raw_counts.txt",header = TRUE, row
.names = 1)
#Renombrar el nombre de las columnas
colnames(contajes.raw) <- c("WT1_DMSO","WT1_dTAG","RJA6_DMSO","RJA6_dTAG"
,"RJC6_DMSO","RJC6_dTAG","RJE5_DMSO","RJE5_dTAG","WT2_DMSO","WT2_dTAG"
,"RJG10_DMSO","RJG10_dTAG","RJH1_DMSO","RJH1_dTAG","WT3_DMSO","WT3_dTAG"
,"RJF2_DMSO","RJF2_dTAG")
#Crear marco de datos de la condición
colData <- data.frame(condicion=factor(rep(c("WT","WT","RJ","RJ","RJ","RJ","RJ"
,"RJ","WT","WT","RJ","RJ","RJ","RJ","WT","WT","RJ","RJ"))))
#Asignar nombres de las filas al marco de datos de condición
rownames(colData) <- colnames(contajes.raw)
#Crear diseño de análisis
design <- formula(~ condicion)
#Crear objeto DESeqDataSet (dds)
dds <- DESeqDataSetFromMatrix(countData = contajes.raw, colData = colData, design
=design)
dds
```

Chunk 3. Preparación de los datos: Filtrado de datos

```r
#Conteo de las columnas
X <- ncol(counts(dds))
#Filtrado de genes
keep <- rowSums(counts(dds) >= 10) >= X
#Actualizacion del objeto dds
dds <- dds[keep,]
```

Chunk 4. Obtención de los recuentos en base a las condiciones

```r
# Obtener los recuentos de expresión génica con DESeq2
contajes <- counts(dds)
contajes
colnames(contajes)<-c("WT","WT","RJ","RJ","RJ","RJ","RJ","RJ","WT","WT","RJ","RJ"
,"RJ","RJ","WT","WT","RJ","RJ")
```

Análisis de expresión diferencial con DESeq2

Chunk 5. Normalización de datos, basándose en la metodología de *DESeq2* con el "*factor de tamaño*"

```r
#Datos normalizados, parametros dispersión y prueba hipótesis nula
dds <- DESeq(dds)
```

```
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

```r
normalized_counts<-counts(dds, normalized=TRUE)
```

Chunk 6. Transformación de datos y desarrollo de *PCA*

```r
# Transformación de tamaño de muestra de varianza estabilizante (vst)
vst_data <- vst(dds, blind = FALSE)

# Ahora podemos realizar un PCA en los datos transformados
PCA_res<- prcomp(t(assay(vst_data)))

# Ver los resultados
print(PCA_res)
```

Chunk 7. Gráfico del *PCA* en base a la varianza de los componentes

```r
{r}
# Extraer los resultados PCA y convertirlos a un data frame
PCA_data <- as.data.frame(PCA_res$x)

# Añadir la condición experimental de cada muestra al data frame
PCA_data$condicion <- vst_data$condicion

#Calculo de la varianza de cada componente
percentVar<-(PCA_res$sdev^2) / sum(PCA_res$sdev^2)

# Crear el gráfico PCA usando ggplot2 y ggrepel
ggplot(PCA_data, aes(PC1, PC2, color = condicion)) +
  geom_point() +
  geom_text_repel(aes(label = rownames(PCA_data)))+
  theme_minimal() +
  xlab(paste0("PC1: ", round(percentVar[1] * 100), "% variance")) +
  ylab(paste0("PC2: ", round(percentVar[2] * 100), "% variance")) +
  labs(color = "Condición")+
  ggtitle("Análisis de Componentes Principales (PCA)")
```

Chunk 8. Resultados del *PCA* y exportación de la lista de estos componentes

```r
{r}
#Resultados de PCA
PCA<-results(dds)

# Obtener los datos del PCA
PCA1<- as.data.frame(PCA@listData)
```

Chunk 9. Cálculo de la varianza de *PCA*, impresión de los resultados en base a sus dimensiones

```r
{r}
# Cálculo de la varianza en los componentes principales
eigenvalues<-get_eigenvalue(PCA_res)
eigenvalues
```

Chunk 10. Gráfico de un fviz_screeplot de las dimensiones de *PCA*

```r
{r}
# Crear un Scree plot
fviz_screeplot(PCA_res, addlabels = TRUE, ylim = c(0, 50))
```

Chunk 11. Anotación de los genes mediante columnas por lectura "*org.Hs.eg.db*"

```r
{r}
#Obtener la lista de columnas disponibles
columns(org.Hs.eg.db)
```

Chunk 12. Conversión de los identificadores

```r
{r}
#Conversion de identificadores
convertIDs <- function( ids, from, to, db, ifMultiple=c("putNA", "useFirst")) {
  stopifnot( inherits( db, "AnnotationDb" ) )
  ifMultiple <- match.arg( ifMultiple )
  suppressWarnings( selRes <- AnnotationDbi::select(
    db, keys=ids, keytype=from, columns=c(from,to) ) )
  if ( ifMultiple == "putNA" ) {
    duplicatedIds <- selRes[ duplicated( selRes[,1] ), 1 ]
    selRes <- selRes[ ! selRes[,1] %in% duplicatedIds, ]
  }
  return( selRes[ match( ids, selRes[,1] ), 2 ] )
}
```

Chunk 13. Conversión de los identificadores de los genes, con el propósito del empleo de varios recursos para datos genómicos.

```r
{r}
PCA$hgnc_symbol <- convertIDs(row.names(PCA), "ENSEMBL", "SYMBOL", org.Hs.eg.db)
PCA$entrezgene <- convertIDs(row.names(PCA), "ENSEMBL", "ENTREZID", org.Hs.eg.db)
PCA2<-PCA@listData
```

Chunk 14. Gráfico de Volcán o "*Volcano Plot*" que evalúa el nivel de expresión de los genes tratados con DESeq2

```r
{r}
tab = data.frame(logFC = PCA$log2FoldChange, negLogPval = -log10(PCA$pvalue),
hgnc_symbol = PCA$hgnc_symbol)

ggplot(tab, aes(x = logFC, y = negLogPval)) +
  geom_point(alpha = 0.1, size = 3) +
  geom_point(data = subset(tab, abs(logFC) > 2 & negLogPval > -log10(0.01)),
             aes(x = logFC, y = negLogPval), color = "red", size = 3) +
  geom_vline(xintercept = c(-2, 2), color = "blue", linetype = "dashed") +
  geom_hline(yintercept = -log10(0.01), color = "green", linetype = "dashed") +
  geom_text_repel(data = subset(tab, abs(logFC) > 2 & negLogPval > -log10(0.01)),

                  aes(x = logFC, y = negLogPval, label = hgnc_symbol),
                  box.padding = unit(0.2, "lines"),
                  point.padding = unit(0.2, "lines"),
                  size = 2,
                  max.overlaps = Inf) +
  xlab(expression(log[2]~fold~change)) +
  ylab(expression(-log[10]~pvalue))
```

Chunk 15. Obtención de los Genes de interés en base al parámetro de "*LogFoldChange*" y "*P-value*"

```r
{r}
# Crear subconjunto de datos para genes de interés
genes_de_interes = subset(tab, abs(logFC) > 2 & negLogPval > -log10(0.01))

# Ver la tabla de genes de interés
View(genes_de_interes)
```

Chunk 16. Establecimiento de los valores para "*LogFoldChange*" en caso de los genes sobreexpresión e infraexpresados para identificar cuáles son.

```r
{r}
#Establecer umbrales de sobreexpresión e infraexpresión
pvalor_umbral <- 0.05
foldchange_umbral <- 2
foldchange_umbral2 <- -2

#Identificar genes sobreexpresados e infraexpresados
genes_sobreexpresados <- subset(PCA, padj < pvalor_umbral& log2FoldChange >
foldchange_umbral)

genes_infraexpresados <- subset(PCA, padj < pvalor_umbral & log2FoldChange
>foldchange_umbral2)
```

Chunk 17. Análisis de Enriquecimiento de rutas o conjunto de genes (*GSEA*)

Creación de datos-universo para identificar los genes sobreexpresados que se encuentran enriquecidos por procesos biológicos relacionados con la patología.

```r
{r}
#Conjunto universo de Genes sobreexpresados para GSEA
universe <- AnnotationDbi::select(x = org.Hs.eg.db,keys = rownames
(contajes),columns = c("ENSEMBL","ENTREZID"),keytype = "ENSEMBL")

#Crear el componente GSEA
enrich.result <- ReactomePA::enrichPathway(gene =
genes_sobreexpresados$entrezgene,
                                    pvalueCutoff = 0.05,
                                    readable = T,
                                    pAdjustMethod = "BH",
                                    organism = "human",
                                    universe =universe$ENTREZID)
```

Chunk 18. Gráfico "*BarPlot*" que indica el grado de enriquecimiento de genes sobreexpresados en los datos de expresión diferencial

```r
{r}
barplot(enrich.result)
```

Chunk 19. Gráfico "*CnetPlot*" para observar las conexiones y nodos de genes relacionados con los sobreexpresados, con la patología

```r
{r}
cnetplot(enrich.result)
```

Chunk 20. Creación de datos-universo para identificar los genes infraexpresados que se encuentran enriquecidos por procesos biológicos relacionados con la patología.

```r
{r}
#Universo de datos infraexpresados
universe <- AnnotationDbi::select(x = org.Hs.eg.db,keys = rownames
(contajes),columns = c("ENSEMBL","ENTREZID"),keytype = "ENSEMBL")

#Crear el componente GSEA para genes infraexpresados
enrich.result1 <- ReactomePA::enrichPathway(gene =
genes_infraexpresados$entrezgene,
                                    pvalueCutoff = 0.05,
                                    readable = T,
                                    pAdjustMethod = "BH",
                                    organism = "human",
                                    universe =universe$ENTREZID)
```

Chunk 21. Gráfico "*Barplot*" que indica el grado de enriquecimiento de genes infraexpresados en los datos de expresión diferencial

```r
{r}
barplot(enrich.result1)
```

Chunk 22. Gráfico "*CnetPlot*" para observar las conexiones y nodos de genes relacionados con los infraexpresados, con la patología

```r
{r}
cnetplot(enrich.result1)
```

Análisis de expresión diferencial con BADER

Chunk 23. Empleo del paquete *BADER* (*Bayesian Analysis of differential Expression in RNA-seq data)* que contempla el Modelo bayesiano jerárquico sobre los contajes.

```r
{r}
results1 <- BADER(contajes,colData$condicion)
saveRDS(results1,"./resultados_bayesianos.rds")
resultados1 <- readRDS("./resultados_bayesianos.rds")
```

Chunk 24. Obtención de los parámetros de "*LogFoldChange*" y "*diffProb*" específicamente en este enfoque bayesiano, para obtener la probabilidad posterior de la expresión diferencial de los genes "*DEG*".

```r
{r}
# Extraer los valores de logFoldChange y diffProb
logFC <- resultados1$logFoldChange

diffProb <- resultados1$diffProb

# Crear un data frame con los valores extraídos
genes_significativos <- data.frame(logFoldChange = logFC, diffProb = diffProb)

# Filtrar por genes diferencialmente expresados mayores a 0,95
genes_significativos <- genes_significativos[genes_significativos$diffProb > 0.95
, ]

# Imprimir los genes diferencialmente expresados
print(genes_significativos)
```

Chunk 25. Gráfico de Volcán o "*Volcano Plot*" que imprimen los resultados de los genes significativos una vez aplicado BADER

```r
{r}
ggplot(genes_significativos,aes(x = logFoldChange, y = 1-(diffProb))) +

  geom_point(size = 1, color = "black") + geom_point( size = 2) + geom_hline
(yintercept = -log10(0.05), linetype = "dashed", color = "red") +

  labs(x = "logFoldChange", y = "-log10(diffProb)", title = "Volcano Plot") +
theme_minimal()
```

Chunk 26. Desarrollo del gráfico "*Volcano Plot*" para la identificación de los resultados obtenidos de *BADER* empleando las etiquetas de la librería "*ggrepel*" para la visualización de los genes implicados positivamente *DEG* o no *DEG*.

```
{r}
resultados1$hgnc_symbol <- convertIDs(row.names(resultados1), "ENSEMBL", "SYMBOL"
, org.Hs.eg.db)
resultados1$entrezgene <- convertIDs(row.names(resultados1), "ENSEMBL",
"ENTREZID", org.Hs.eg.db)

# Crear una submuestra de los datos que contengan solo los genes DE
genes_DE_subset <- subset(resultados1, DE == "DE")

# Crear el gráfico de volcan, etiquetando solo los genes DE
ggplot(resultados1, aes(x = logFoldChange, y = diffProb, color = DE)) +
 geom_point(size = 2) +
 geom_text_repel(data = genes_DE_subset,
                 aes(label = hgnc_symbol),
                 size = 3,
                 box.padding = unit(0.3, "lines"),
                 point.padding = unit(0.1, "lines"),
                 max.overlaps =Inf)
```

Chunk 27. Almacenamiento de los genes diferencialmente expresados del análisis con BADER

```
{r}
resultados1[resultados1$DE=="DE",]
genesDE <- rownames(resultados1[resultados1$DE=="DE",])
genesDE
```

Chunk 28. Creación de un marco de datos para su posterior lectura con *GSEA*.

```
{r}
resultados1 <- readRDS("./resultados_bayesianos.rds")
resultados1 <- as.data.frame(resultados1)
rownames(resultados1) <- rownames(contajes)
```

Chunk 29. Creación del entorno de análisis para GSEA en el caso de los genes sobreexpresados

```
{r}
up <- resultados1[resultados1$logFoldChange> 1 & resultados1$diffProb > 0.95,]
up.names <- rownames(up)
up.names <- AnnotationDbi::select(org.Hs.eg.db,
                                  keys = up.names,
                                  columns = c("ENSEMBL", "ENTREZID"),
                                  keytype = "ENSEMBL")
up.names <- up.names$ENTREZID
universe <- AnnotationDbi::select(org.Hs.eg.db,
                                  keys = rownames(contajes),
                                  columns = c("ENSEMBL",
                                              "ENTREZID"),
                                  keytype = "ENSEMBL")
universe <- universe$ENTREZID
```

Chunk 30. Obtención de los resultados empleado en los genes de sobreexpresión para conocer las vías de enriquecimiento de los genes de estudio.

```{r}
genes <- genes_significativos
results.gsea <- enrichPathway(gene = up.names,organism = "human",
                              pvalueCutoff = 0.05,
                              universe =universe,
                              readable = T )
```

Chunk 31. Gráfico de "BarPlot" de los genes sobreexpresados ante el análisis de GSEA con BADER

Se presentan las barras que distribuyen la cantidad de procesos involucrados

```{r}
barplot(results.gsea)
```

Chunk 32. Gráfico de "CnetPlot" de los genes sobreexpresados una vez conocido la cantidad de procesos biológicos y genes con "BarPlot"

Se presentan redes que conectan los genes del estudio con los procesos involucrados

```{r}
cnetplot(results.gsea,foldChange = 1)
```

Chunk 33. Creación del entorno de análisis para GSEA en el caso de los genes infraexpresados

```{r}
up <- resultados1[(resultados1$logFoldChange) < -1 & (resultados1$diffProb) > 0
.95,]
up.names <- rownames(up)
up.names <- AnnotationDbi::select( org.Hs.eg.db,
                                   keys = up.names,
                                   columns = c("ENSEMBL","ENTREZID"),
                                   keytype = "ENSEMBL")
up.names <- up.names$ENTREZID
universe <- AnnotationDbi::select(org.Hs.eg.db,
                                  keys = rownames(contajes),
                                  columns = c("ENSEMBL",
                                              "ENTREZID"),
                                  keytype = "ENSEMBL")
universe <- universe$ENTREZID
```

Chunk 34. Obtención de los resultados empleado en los genes de infraexpresión para conocer las vías de enriquecimiento de los genes de estudio.

```r
genes <- genes_significativos
results.gsea1 <- enrichPathway(gene = up.names,organism = "human",
                               pvalueCutoff = 0.06,
                               universe =universe,
                               readable = T )
```

Chunk 35. Gráfico de “BarPlot” de los genes infraexpresados ante el análisis de GSEA con BADER

Se presentan las barras que distribuyen la cantidad de procesos involucrados

```r
barplot(results.gsea1)
```

Chunk 36. Gráfico de “CnetPlot” de los genes infraexpresados una vez conocido la cantidad de procesos biológicos y genes con “BarPlot”

Se presentan redes que conectan los genes del estudio con los procesos involucrados

```r
cnetplot(results.gsea1,foldChange= -1)
```