UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE BIOTECNOLOGÍA

EVALUACIÓN Y COMPARACIÓN DE TÉCNICAS BIOINFORMÁTICAS PARA EL ANÁLISIS DE EXPRESIÓN DIFERENCIAL EN NGS

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

AUTORAS: JACQUELINE MICHELLE JARA MOSCOSO

KERLY SAMANTHA SAQUIPAY NIEVES

TUTORA: DRA. INÉS PATRICIA MALO CEVALLOS, PhD.

Cuenca - Ecuador

2023

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotras, Jacqueline Michelle Jara Moscoso con documento de identificación N° 0106079890 y Kerly Samantha Saquipay Nieves con documento de identificación N° 0105826408; 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, 17 de agosto del 2023

Atentamente,

Jacqueline Michelle Jara Moscoso
0106079890

Kerly Samantha Saquipay Nieves
0105826408

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

Nosotras, Jacqueline Michelle Jara Moscoso con documento de identificación No. 0106079890 y Kerly Samantha Saquipay Nieves con documento de identificación No. 0105826408, 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 autoras del Trabajo experimental: "Evaluación y comparación de técnicas bioinformáticas para el análisis de expresión diferencial en NGS", el cual ha sido desarrollado para optar por 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 hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 17 de agosto del 2023

Atentamente,

Jacqueline Michelle Jara Moscoso
0106079890

Kerly Samantha Saquipay Nieves
0105826408

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: EVALUACIÓN Y COMPARACIÓN DE TÉCNICAS BIOINFORMÁTICAS PARA EL ANÁLISIS DE EXPRESIÓN DIFERENCIAL EN NGS, realizado por Jacqueline Michelle Jara Moscoso con documento de identificación N° 0106079890 y por Kerly Samantha Saquipay Nieves con documento de identificación N° 0105826408, obteniendo como resultado final el trabajo de titulación bajo la opción Trabajo experimental que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 17 de agosto del 2023

Atentamente,

Dra. Inés Patricia Malo Cevallos, PhD.

0102291044

DEDICATORIA

A mis padres por su resiliencia, constancia y apoyo incondicional en cada una de las etapas de mi vida, por invertir en la familia confiando siempre en el proceso.

A mis abuelos por enseñarme que el niño interior nunca descansa, siempre experimenta el mundo con curiosidad y asombro.

Michelle Jara

DEDICATORIA

Quiero dedicar esta tesis a mis papás, Flavio y Norma; pues sin su amor y apoyo no lo habría logrado. Puedo decir que ellos han sido el pilar fundamental en mi vida y en sí todo lo que tengo se lo debo a ellos. Nada hubiera sido posible sin el amor que me han brindado y esta es una manera de retribuirles todo lo que han hecho por mí. Aún faltan más logros, pero cada logro será dedicado especialmente a mis papás.

Al Ing. Edmond Géraud, por su paciencia, apoyo y compromiso para realizar esta tesis. Todos los conocimientos impartidos por su parte fueron de gran ayuda, y así se ha convertido en un gran ejemplo a seguir.

También a mi hermana, que a pesar de ser pequeña me ha sabido comprender, animar y ayudar en lo que ha podido.

Samantha Saquipay

AGRADECIMIENTOS

A los docentes de la carrera por el esfuerzo que realizan para formar buenas personas, por compartir su conocimiento y enseñanzas de vida.

Al Ing. Edmond Géraud por su compromiso y los conocimientos impartidos durante todo el proceso de investigación, por su profesionalismo y por ser buen ejemplo de superación personal.

A mis amigas por el apoyo incondicional en este proceso de crecimiento personal, por las buenas experiencias y enseñanzas que me han permitido vivir.

A mis hermanos por las vivencias compartidas en momentos de ocio y por influir siempre de manera positiva en mi vida.

Michelle Jara

AGRADECIMIENTOS

Quiero agradecer a mis papás por todo el amor y la compresión que me han brindado. Por confiar en mí y enseñarme que las cosas no son fáciles, pero con esfuerzo y siendo honestos se pueden lograr muchas cosas.

A mi hermana, pues con sus ocurrencias y su forma tan divertida de ser me ha levantado los ánimos, se ha quedado conmigo los días y noches que tenía que hacer trabajos o tenía que estudiar, y de cierta manera me ha ayudado a ser feliz. Siempre diré que es la persona que más quiero en esta vida.

Agradezco de una manera muy especial al Ing. Edmond Géraud, por su paciencia y compresión; puesto que no habría sido posible realizar la tesis sin su ayuda. En realidad las palabras no me alcanzan para agradecerle por todo el apoyo y sobre todo los ánimos que nos brindó. Y en sí, agradezco por todas las enseñanzas impartidas por los docentes que tuve durante la carrera.

A las amigas que hice durante la carrera, Oda, Michis y Gaby; pues han sido las mejores personas que he conocido. Y me han ayudado a crecer en muchos aspectos. Jamás olvidaré todo lo que han hecho por mí, por escucharme y apoyarme en todo. Me han demostrado que si existe una verdadera amistad y compañerismo.

También agradezco a mis amigos David, Wendy, Alejandra y Erick, que igualmente siempre me han apoyado, me han demostrado amor sincero y se han preocupado por mí. Les agradezco por haberme escuchado y apoyado en los momentos más difíciles, cuando sentía que no podía siempre estuvieron ahí.

Samantha Saquipay

ÍNDICE DE CONTENIDO

ABREVIATURAS

DE: diferencialmente expresados

DGE: análisis génico de expresión diferencial

diffexp: número de genes DE

FPR: tasa de falsos positivos

m: número de transcritos totales

n: número de muestras

NB: binomial negativa

NGS: secuenciación de próxima generación

RNA-Seq: secuenciación de ácido ribonucleico (ARN)

TPR: tasa de verdaderos positivos

RESUMEN

Existen varias herramientas bioinformáticas que se pueden implementar para el análisis génico de expresión diferencial (DGE), sin embargo, no existe un consenso claro sobre las mejores prácticas, lo que hace que la elección de un método apropiado de análisis sea difícil. Por lo tanto, mediante simulaciones con datos de *RNA-Seq* se llevaron a cabo análisis DGE con paquetes de *Bioconductor* (*DESeq2, EdgeR y Limma*) empleando flujos de trabajo estándar para determinar parámetros óptimos en relación al número de muestras (n), transcritos totales (m) y genes DE (diffexp). Mediante un análisis posterior con un conjunto de datos reales se comprobó que *DESeq2* identifica más genes diferencialmente expresados (DE), así como también junto a *Limma* son paquetes óptimos cuando se obtiene un número de transcritos alto. Por último, *Limma* ofrece un enfoque más conservador en comparación con *DESeq2*, pero es computacionalmente más rápida de ejecutar.

Palabras clave: RNA-Seq, DGE, DE, *DESeq2*, *Limma*, *EdgeR*.

ABSTRACT

There are several bioinformatics tools that can be implemented for differential gene expression (DGE) analysis, however there is no clear consensus on best practice, making the choice of an appropriate method of analysis difficult. Therefore, through simulations with RNA-Seq data, DGE analyzes were carried out with *Bioconductor* packages *(DESeq2, EdgeR and Limma)* using standard workflows to determine optimal parameters in relation to the number of samples (n), total transcripts (m) and DE (diffexp) genes. Through a subsequent analysis with a real data set, it was verified that *DESeq2* identifies more differentially expressed (DE) genes, as well as together with *Limma* they are optimal packages when a high number of transcripts is obtained. Lastly, *Limma* offers a more conservative approach compared to *DESeq2*, but is computationally faster to run.

Keywords: RNA-Seq, DGE, DE, *DESeq2*, *Limma*, *EdgeR*.

CAPÍTULO 1

INTRODUCCIÓN

La secuenciación de ARN de alto rendimiento (*RNA-Seq*) ha revolucionado la forma en que se mide la expresión génica. Actualmente es posible analizar el transcriptoma en profundidad utilizando *RNA-Seq*, ya que la tecnología aprovecha las capacidades de secuenciación de próxima generación (NGS). Muchos estudios han demostrado que los métodos de *RNA-Seq* proporcionan mediciones más sensibles, precisas y completas de la expresión génica (Mortazavi et al., 2008, Nagalakshmi et al., 2008, Wang et al., 2009).

En un experimento de *RNA-Seq*, el ARNm (transcrito) primero se aísla y se fragmenta aleatoriamente, luego el ARNm se convierte en ADN complementario (ADNc) y se prepara para la secuenciación. Los fragmentos de ADNc se secuencian simultáneamente, generando cientos de millones de lecturas cortas de *RNA-Seq*. Luego, las lecturas de *RNA-Seq* se alinean con una base de datos de referencia, y la frecuencia relativa de las lecturas de *RNA-Seq* correspondientes a un gen sirve como medida de la expresión de ese transcrito. Por lo tanto, deben usarse métodos estadísticos para identificar genes DE. Los problemas estadísticos se complican por la necesidad de probar un gran número de genes, muestras biológicas pequeñas y una variabilidad que no se puede modelar mediante distribuciones de probabilidad de uso común. Se debe tener especial cuidado en combinar un gran número de genes con muestras pequeñas, ya que la potencia de la prueba depende de esta combinación. Por estas razones se emplean herramientas bioinformáticas, capaces de realizar un análisis exhaustivo de la expresión diferencial, basándose en diferentes metodologías y distribuciones, tal es el caso de la distribución de Poisson y la distribución binomial negativa; que permiten modelar los datos de conteo de *RNA-Seq*.

Robinson y Smyth (2007, 2008), mencionaron que la distribución binomial negativa ofrece un modelo más realista para la variabilidad del conteo de *RNA-Seq* que el modelo de Poisson, por la propia sobredispersión de los datos. De hecho, la sobredispersión de los datos permite parametrizar la binomial negativa. Es más, el conjunto de distribuciones de Poisson forma una distribución binomial negativa (Di, Schafer, Cumbie & Chang, 2011). La distribución propuesta por Robinson (2007), es una forma inversa a la distribución binomial. Por ejemplo, la transcripción y el error de la medición permiten ver la probabilidad de que se observe un gen. No obstante, al cuantificar todos los transcritos, en realidad, no se sabe cuál se observa; es decir no se puede mirar la probabilidad de un ARNm en específico. En cambio, se puede observar la probabilidad de que salgan todos los transcritos menos el de interés. Esta es la idea principal detrás de modelar el transcriptoma con la distribución binomial negativa.

Una vez modelados los datos de *RNA-Seq* se debe realizar propiamente el análisis de expresión génica diferencial (DGE). Para ello, se emplean los paquetes de datos más conocidos en el lenguaje de programación R, como son *DESeq2*, *EdgeR* y *Limma*; estos paquetes permiten realizar un procesamiento de los datos, la normalización y la aplicación de pruebas estadísticas. Cada paquete tiene sus propias características y especificaciones para diferentes tipos de datos y diseños experimentales, en consecuencia, se obtienen resultados diferentes.

1.1 PLANTEAMIENTO DEL PROBLEMA

La Bioinformática es una disciplina con muchos enfoques en las ciencias de la vida. De esta forma, se pueden utilizar diversas herramientas bioinformáticas para analizar datos biológicos, biomédicos, químicos, biofísicos, etc., facilitando su recogida, interpretación, gestión y almacenamiento. A pesar de que la Bioinformática es una de las disciplinas más destacadas en los últimos años y con muchas perspectivas de futuro, en el Ecuador no es un área de investigación

consolidada. Ayala (2020), menciona que según *"Bioinformatics in Latin America Review"*, que recopiló datos sobre el número de publicaciones en 20 países entre 1991 y 2016, Ecuador escribió 15 artículos (que representan el 0,71% de todas las investigaciones publicadas en América Latina). Además, otro estudio global de publicaciones de Bioinformática hasta 2019 señaló que en Ecuador se publicaron 40 estudios. A pesar de un aumento significativo en el número de publicaciones en esta área de investigación, Ecuador aún se encuentra lejos del promedio mundial de logros científicos.

Por otra parte, en las últimas décadas, la secuenciación del ARN *(RNA-Seq)* mediante la aplicación de tecnologías de secuenciación de nueva generación (NGS) se ha convertido en un enfoque esencial en la elaboración de perfiles de transcriptomas. Un problema de investigación fundamental en muchos estudios de *RNA-Seq* es la identificación de métodos confiables de análisis de expresión diferencial (DGE); para lo cual, no existe un consenso claro sobre las mejores prácticas, lo que hace que la elección de un método apropiado sea difícil, especialmente para un usuario básico sin una sólida formación estadística o computacional. No obstante, al hacer uso de cada paquete, se presenta variabilidad de resultados puesto que existen diferencias en los métodos de normalización de datos, así como también en los modelos probabilísticos. Por lo tanto, se plantea el análisis de la expresión génica diferencial comparando tres paquetes *(DESeq2, EdgeR, Limma*) del repositorio *Bioconductor*.

1.2 JUSTIFICACIÓN

Las técnicas de expresión génica se utilizan a menudo en estudios de Biología Molecular para obtener actividad transcripcional inmediata en diferentes tejidos o poblaciones celulares. Estos perfiles luego se comparan para identificar cambios en la expresión génica asociados con la condición tratada o el fenotipo de interés. El análisis de la expresión génica se basa en

experimentos diseñados en los que se perturban los sistemas biológicos, como la inactivación de genes o la aplicación de factores estresantes específicos, o simplemente para observar comparaciones de casos y controles. Dichos experimentos son muy importantes porque brindan información sobre los procesos celulares normales, así como sobre la patogénesis de la enfermedad. Incluso es posible realizar estudios observacionales que comparen diferentes fenotipos, tejidos enfermos y normales, o incluso células de diferentes poblaciones. Dichos estudios son comunes en la investigación del cáncer y la investigación del desarrollo celular. La tecnología *RNA-Seq* es muy beneficiosa para el análisis de expresión diferencial, que implica algunas preguntas específicas que se pueden resumir en cinco pasos. Primero, las muestras de ARN se fragmentan en secuencias de ADNc, que luego se secuencian desde una plataforma de alto rendimiento. Posteriormente, las secuencias resultantes se asignan a genomas o transcriptomas. En tercer lugar, se evalúa el nivel de expresión de cada gen o subtipo. Los datos mapeados después se normalizan usando métodos bioestadísticos y se identifican los genes expresados diferencialmente. Finalmente, se evalúa la relevancia de los datos obtenidos en relación con el contexto biológico (Costa-Silva, Domingues, & Lopes, 2017).

Vale la pena mencionar que los estudios de expresión génica generalmente involucran sólo una pequeña cantidad de réplicas biológicas debido a su complejidad. Al hacer preguntas estadísticas, el objetivo es aprovechar al máximo cada conjunto de datos, por lo que se desarrolló un *software* para realizar un análisis diferencial de la expresión génica a partir de estos datos. Se incluyen los paquetes *DESeq2*, *EdgeR* y *Limma* del repositorio de *Bioconductor*, que utilizan el lenguaje de programación estadístico R. No obstante, cada paquete tiene su propia especificación, modelo probabilístico, normalización y métodos de filtrado. En consecuencia, esto conduce a resultados diferentes, lo que a su vez genera problemas de reproducción de los estudios realizados

(Ritchie et al., 2015). Por lo tanto, el presente trabajo abordará las limitaciones y ventajas de cada paquete, así como también una comparación entre los resultados para brindar más información que ayude a la comunidad científica a elegir el método apropiado.

Por lo tanto, se plantea la siguiente pregunta de investigación:

¿Son el número de muestras, el número de transcritos totales, los genes diferencialmente expresados, los métodos de normalización, el modelo subyacente en los que se fundamentan los distintos paquetes (*DESeq2, EdgeR, Limma*), los que hacen que se obtengan distintos resultados en el análisis de datos de *RNA-Seq* y juegan un papel importante en el análisis diferencial de expresión?

Se llevarán a cabo análisis comparativos utilizando muestras simuladas de *RNA-Seq* para investigar la distribución natural, sobre la cual se supone que los modelos estadísticos están construidos. A través de estas simulaciones, se computarán métricas correspondientes a un método de clasificación de tipo caso-control para evaluar la sensibilidad y especificidad de cada modelo. Además, se realizarán pruebas de permutación con diferentes tamaños de muestra para evaluar las diferencias entre los tres métodos. Una vez que se haya seleccionado el método adecuado basado en las simulaciones, se compararán los resultados de los tres métodos utilizando el conjunto de datos reales obtenidos por Tao y colaboradores (2022), en el artículo "*ISG15 is associated with cervical cancer development",* para una vez más determinar el mejor método empleado para el análisis DGE.

1.3 ANTECEDENTES

En las últimas décadas, *RNA-Seq* se ha convertido en un método fundamental en la investigación médica para el perfilado de transcritos, que juega un papel importante en el diagnóstico, pronóstico y tratamiento de enfermedades, una de las más importantes es el cáncer.

Las herramientas de análisis computacional para *RNA-Seq* han aumentado significativamente durante la última década, y la elección de herramientas específicas debe basarse en el propósito y la precisión de la aplicación (Hong et al., 2020).

Según un estudio de Chamorro (2019) utilizando diferentes paquetes de software desarrollados por el proyecto *Bioconductor*, mencionó que *Limma*, *EdgeR* y *DESeq2* permiten el análisis de datos de *RNA-Seq* para estudios de análisis DGE con resultados similares excepto cuando la cantidad de muestras por grupo es muy bajo. Sin embargo, se ha demostrado que *Limma* tiene un bajo poder estadístico para muestras muy pequeñas. *EdgeR* y *DESeq2*, basados en la distribución binomial negativa, son los dos paquetes con el enfoque más sencillo. Por el contrario, *Limma* ofrece un enfoque más conservador, puesto que obtiene un número menor de genes DE, pero es computacionalmente más rápido independientemente del tamaño de la muestra. Por otro lado, Seyednasrollah, Laiho y Elo (2015) compararon ocho paquetes de software para la detección de disfunción eréctil en estudios de *RNA-Seq* en un estudio y concluyeron que se debe considerar el número de réplicas y la heterogeneidad de la muestra al elegir un método apropiado. Además, también se puede demostrar que en la comparación actual, el paquete *Limma* generalmente funciona bien en muchos casos y también realiza cálculos más rápidos. Tong (2021) mencionó en un artículo que, a medida que la tecnología ha evolucionado, han surgido muchos métodos para analizar genes DE, como *Limma*, *DESeq2* y *EdgeR*. Sin embargo, no está claro si el uso de diferentes métodos producirá resultados diferentes. Una comparación de los resultados obtenidos con *DESeq2* y *Limma* durante un análisis posterior de los datos de *RNA-Seq* mostró que la cantidad de genes encontrados difería entre sí. Se evidenció que *DESeq2* detecta más genes que *Limma*. Sin embargo, más del 90% de los genes detectados por ambos métodos fueron consistentes, lo que implica que ambos métodos son confiables. Finalmente, se concluyó que si los resultados

aproximados son aceptables, ambos paquetes son suficientes, pero si se requieren resultados exactos, se recomienda *Limma*.

1.4 OBJETIVOS

1.4.1 General

- Realizar una simulación de datos transcriptómicos para la selección de la mejor herramienta bioinformática de análisis de expresión diferencial (DGE) comparando tres paquetes de *Bioconductor (EdgeR y DESeq2, Limma)* según las condiciones de estudio y su posterior validación con un conjunto de datos reales.

1.4.2 Específicos

- Realizar una simulación matemática de datos transcriptómicos siguiendo la distribución binomial negativa para los análisis posteriores.
- Realizar un análisis DE para cada gen simulado mediante la aplicación de las condiciones por defecto de cada paquete para la determinación de la especificidad y sensibilidad.
- Evaluar el rendimiento de cada procedimiento mediante el uso de curvas ROC para la identificación del método más apropiado.

CAPÍTULO 2

MARCO TEÓRICO REFERENCIAL

2.1 Secuenciación de Nueva Generación (NGS) y Secuenciación del ARN (*RNA-Seq*)

La secuenciación de próxima generación (NGS) es una tecnología de secuenciación paralela masiva, también conocida como secuenciación de alto rendimiento, que permite el análisis de grandes fragmentos de genomas de ADN y ARN con alta sensibilidad. La tubería de NGS consta de dos partes principales: la parte de laboratorio húmedo, que incluye la preparación, amplificación y secuenciación de muestras; y la segunda parte es un flujo de trabajo de bioinformática que utiliza datos obtenidos en un laboratorio húmedo para obtener secuencias (Saeed & Usman, 2019). Inicialmente, los estudios de expresión génica basados en técnicas de bajo rendimiento, como Northern blot y qPCR, se limitaban a la medición de transcritos individuales. Los primeros estudios transcriptómicos se realizaron utilizando tecnología de *Microarrays,* basada en la hibridación de moléculas complementarias de ADN entre sí, los cuales proporcionaron un alto rendimiento a un costo relativamente bajo. Sin embargo, este método tenía varias limitaciones, como la necesidad de conocer la secuencia analizada, la alta calidad y pureza de las muestras o la capacidad para identificar con precisión genes de baja y muy alta expresión (Kukurba & Montgomery, 2015). El desarrollo de NGS de alto rendimiento para permitir el análisis de ARN a través de la secuenciación de ADNc ha revolucionado la transcriptómica. Un transcriptoma contiene el conjunto completo de transcritos para un tejido, organismo o célula en particular bajo ciertas condiciones fisiológicas. Las transcripciones incluyen ARN mensajero codificante de proteínas (ARNm) y ARN no codificante, como ARNr, ARNt y otros ARNnc. *RNA-Seq* proporciona una imagen más detallada y cuantitativa de la expresión génica, el empalme alternativo y la expresión específica de alelos. Los avances recientes en los flujos de trabajo de

RNA-Seq, desde la preparación de muestras hasta las plataformas de secuenciación y el análisis de datos bioinformáticos, han permitido la creación de perfiles transcriptómicos profundos con el potencial de dilucidar una variedad de condiciones fisiológicas y patológicas (Gaur & Chaturvedi, 2017). Los avances en curso en Bioinformática han facilitado el procesamiento de datos de *RNA-Seq*. Existen varias herramientas bioinformáticas, servidores web y canalizaciones completas para procesar y analizar datos de *RNA-Seq*. Además, algunas estrategias adecuadas para el análisis de datos de *RNA-Seq* se pueden implementar en *Bioconductor* utilizando el lenguaje estadístico "R" (https://www.r-project.org) (Huber et al., 2015).

2.2 Análisis de Expresión Génica Diferencial (DGE)

La expresión diferencial implica comparar los niveles de expresión génica entre dos condiciones. Se puede decir que los genes se expresan diferencialmente si existe una diferencia estadísticamente significativa o un cambio en el recuento de lecturas entre las dos condiciones. Preprocesamiento adicional y mapeo de expresiones diferenciales; es necesario analizar la distribución del número de lecturas, que suele expresarse en forma de matriz. *Bioconductor* tiene varios paquetes que admiten el análisis DGE de datos de *RNA-Seq*. La mayoría de los paquetes esperan que los datos de entrada sean una matriz de valores enteros. Dado que no existe un estándar universal para el análisis DGE, no se puede esperar que dos estrategias de análisis diferentes para los mismos datos conduzcan a los mismos resultados, aunque aún se espera similitud (Gaur & Chaturvedi, 2017).

2.3 Matrices de conteo

Los datos de conteo se refieren al número de lecturas alineadas con un transcrito específico. Estos datos se describen en forma de matriz, donde las columnas representan muestras y las filas

representan genes. Cabe mencionar que estos son datos discretos, es decir, no son continuos. Por lo tanto, no pueden ser modelados por una distribución continua.

2.4 Distribución de Poisson

Para modelar los datos de recuento se utiliza una distribución de probabilidad discreta conocida como distribución de Poisson. Esta distribución se basa en la tasa de descubrimientos, que podemos denominar “r”, donde la media de la distribución es lo mismo que el mencionado parámetro que es igual a la media. No obstante, los transcritos no siguen una distribución de Poisson como tal, debido a lo que se conoce como sobredispersión de los datos. Por ende, a dicha probabilidad se le puede añadir otro parámetro el cual podrá modelar la dispersión (Pan et al., 2023).

2.5 Distribución Binomial Negativa (NB)

Este es otro modelo basado en datos discretos. Los datos que se obtienen son lecturas, es decir el número de veces que se ha leído un gen debido al alineamiento y el ensamblaje. A diferencia del modelo de Poisson clásico, la distribución binomial negativa maneja recuentos sobredispersos para adecuar la variación de los datos. Por lo tanto, esta distribución permite que la varianza sea mayor que la media, es decir modela la dispersión de los datos. Si solamente proviene de una mezcla de distribuciones de Poisson sobredispersas (Liu et al., 2022).

La distribución binomial negativa se puede formular de la siguiente manera, originalmente:

$$p(x) = \frac{\Gamma(x+n)}{\Gamma(n)!} p^n (1-p)^x \quad (1)$$

Donde el nominador de la fórmula de la probabilidad es la función gamma. Que no es más que la extensión de lo que conocemos como factorial. Es decir, la extensión a números reales de por ejemplo 5!.

La distribución NB surge como una distribución del número de fallas (x) antes del *n*-ésimo éxito en pruebas independientes, con probabilidad de éxito *p* en cada prueba (en consecuencia, para x= 0,1,2,... n > 0 y 0 < $p \leq 1$) (Lindén & Mäntyniemi, 2011).

Alternativamente, proviniendo de distribuciones de Poisson, se puede parametrizar de la siguiente manera:

$$\mu = n(1-p)/p \tag{2}$$

$$\sigma^2 = \frac{n(1-p)}{p^2} = \mu + \frac{\mu^2}{x} \tag{3}$$

Siendo *x* la dispersión que equivale el número de sucesos hasta encontrar el gen.

Donde la media,

$$E[Y|x_i] = 2^{X\beta} \tag{4}$$

Considerando que la media de varias observaciones de Poisson, va a seguir un modelo, cuya parte lineal se encuentra bajo el exponente 2, y va a coincidir con la media de la binomial negativa.

La varianza se puede reescribir de la siguiente manera:

$$Var(Y|x_i) = \mu_i(1 + \mu/\alpha)|Var\left(Y\middle|x^i\right) > E[Y|x_i] \tag{5}$$

Básicamente la fórmula anterior indica que la varianza es mayor a la media, y el segundo término de la expresión en la suma, la media dividida por el parámetro alfa, es la dispersión de los datos.

Sin embargo, la relación entre la media y la sobredispersión es inversamente proporcional, puesto que a medida que crece la media, la varianza disminuye. A su vez, la varianza depende del parámetro de dispersión.

$$NB(y|\mu,\alpha = \binom{y+\alpha-1}{y}\binom{\mu}{\mu+\alpha}^{y}\binom{\alpha}{\mu+\alpha}^{\alpha} \quad (6)$$

Siendo alfa el parámetro de la dispersión, y el número de veces que no se encuentra el transcrito en cuestión y mu es la media de la media de la distribución NB.

Varias distribuciones de Poisson pueden tener medias distintas, es decir, que siguen su propia distribución. Tal es el caso de los genes, debido a que es común si unos genes tienen pocas lecturas a través de las muestras y otros genes tienen más; en este caso las medias y las varianzas van a diferir. Al tomar en cuenta cada observación correspondiente a las muestras para un gen en específico, se considera que los datos siguen una distribución de Poisson con sobredispersión. En consecuencia, la varianza será mayor a la media. De esta manera, al unir varias distribuciones de Poisson se puede llegar a tener una sola distribución NB.

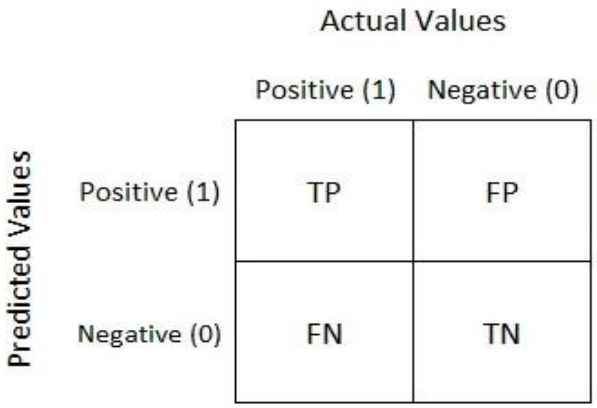
2.6 Matrices de confusión

La matriz de confusión es una matriz *N x N*, donde *N* es el número de niveles de la variable categórica a estudiar, las columnas son los recuentos reales o *"gold standard"*, mientras que las filas de la matriz son los contajes predecidos por la herramienta a evaluar. Dicha matriz compara los valores objetivo reales con los predichos por el modelo o herramienta a evaluar. A partir de dicha matriz, se consiguen diversas métricas útiles para computar el rendimiento de algoritmos, herramientas, modelos estadísticos, de aprendizaje automático, etc. En una matriz de confusión binaria, es decir, en la cual se evalúa una variable categórica de dos niveles, podemos obtener las observaciones clasificadas correctamente conocidos como verdaderos positivos (TP) y las observaciones clasificadas correctamente en la clase negativa se denominan verdaderos negativos (TN). Las instancias de la clase positiva clasificadas falsamente como negativas se denominan falsos negativos (FN) y las observaciones de la clase negativa clasificadas falsamente como

positivas se denominan falsos positivos (FP). Es decir, en esta tabla se muestra la frecuencia con la que el algoritmo predice correcta o incorrectamente el resultado (Ruuska et al., 2018; Düntsch & Gediga, 2020).

Figura 1.

Matriz de confusión



Nota. Matriz de confusión. Fuente: Narkhede (2021).

Según Ruuska y colaboradores (2018), a partir de las frecuencias computadas, se pueden calcular indicadores que reflejan cómo se desempeña el clasificador al detectar la clase dada. Se debe considerar que para todos los indicadores el mejor valor es 1, mientras que el peor valor es 0. Los indicadores más comunes y sus formulaciones son las siguientes:

$$Precisión = \frac{TP}{TP + FP} \tag{7}$$

La precisión se refiere al número real de valores pronosticados correctamente que resultaron ser positivos.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{8}$$

La tasa de sensibilidad hace referencia a las observaciones reales que se predicen correctamente. Para calcular esta métrica, se divide el número total de resultados positivos que se pronosticaron correctamente por el número total de resultados positivos reales.

$$Especificidad = \frac{TN}{TN + FP} \tag{9}$$

La especificidad se conoce también como tasa negativa verdadera.

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

La tasa de exactitud mide la frecuencia con la que el modelo realiza una predicción correcta. Se puede calcular como la relación entre el número de predicciones correctas y el número total de predicciones realizadas por los clasificadores.

Además, se debe considerar que un buen modelo debe tener altas tasas de TP y TN, mientras que las tasas de FP y FN deben ser bajas, no obstante, cabe decir, que ningún modelo de clasificación, obtendrá una perfección al 100%. En el caso que se fuera así, se obtendría un modelo sobreajustado, el cual no podría predecir bien futuras clases.

Cabe mencionar que, en el presente trabajo, solamente se tomarán en cuenta la especificidad, es decir la tasa de verdaderos negativos y la sensibilidad que es la tasa de verdaderos positivos.

2.7 Curvas ROC y el área bajo la curva (AUC)

La curva ROC (Receiver-Operating-Characteristic) es una herramienta que nos permite representar en un gráfico el rendimiento del modelo de clasificación. En la curva se representa la sensibilidad (= tasa de verdaderos positivos) en función de 1-especificidad (= la tasa de falsos positivos) para diferentes puntos de corte. A diferentes niveles, parámetros o umbrales, se calcula

ordenando, la sensibilidad y la tasa de falsos positivos. El mejor clasificador, será aquel en el que el parámetro óptimo se encuentre un balance para la sensibilidad y la tasa de falsos positivos (Martínez & Pérez, 2022). En otras palabras, dada cierta herramienta, en distintos puntos de corte se evalúan las métricas de la sensibilidad y la especificidad en un barrido de puntos, conformando la curva ROC.

El AUC o área bajo la curva simplemente es la integral de la función computada. La cual mide en un conjunto la sensibilidad y la especificidad del método. No obstante, no indica qué umbral es el mejor, pero es otra métrica muy útil para combinar ambas métricas. Por otro lado, cuanto más cerca esté la curva ROC de la esquina superior izquierda del gráfico, mayor será la precisión de la prueba debido a que se puede observar una sensibilidad del 100% y una tasa de falsos positivos de 0%. La curva ROC ideal tiene así un área bajo la curva del 100%. Mientras tanto, un AUC = 50% significa que el modelo no tiene capacidad de discriminación para distinguir entre clase positiva y clase negativa, es decir el modelo es igual de preciso que un evento aleatorio (Nahm, 2022).

Al hacer una comparación de dos métodos, uno será mejor que el otro cuando un método tenga un área bajo la curva mayor que el segundo.

2.8 Índice de Youden

Aunque el área bajo la curva ROC (AUC) es el índice global de precisión diagnóstica más utilizado, el índice de Youden permite ver cual es el umbral específico o parámetro específico que nos da el mejor modelo. A diferencia del área bajo la curva ROC (AUC), el índice de Youden determina en qué punto exacto el modelo tiene una mejor clasificación.

$$J(c) = \{Se(c) + Sp(c) - 1\} = \{Se(c) - (1 - Sp(c))\} \tag{11}$$

El valor máximo del *índice de Youden* es 1 que corresponde a una prueba perfecta, y el mínimo es 0 cuando el clasificador no tiene poder de predicción. El mínimo ocurre cuando $sensibilidad = 1 - especificidad$, es decir, la línea diagonal de la curva ROC. La distancia vertical entre la línea diagonal y la curva ROC es el índice *J* para ese corte en particular. El índice *J* está representado por la propia curva ROC.

2.9 Bioconductor

Es un conjunto de librerías, propio para la investigación en Bioinformática. Estas herramientas están escritas en el lenguaje de programación R, además de estar disponibles de forma gratuita, de tal manera que son fáciles de descargar, instalar y modificar a través de un modelo de código abierto. El proyecto *Bioconductor* tiene como misión desarrollar, respaldar y difundir un *software* gratuito de código abierto, capaz de facilitar el análisis riguroso y reproducible de experimentos en cualquier ámbito de las ciencias de la vida.

Originalmente, *Bioconductor* estaba centrado en el análisis y la anotación de *Microarrays*. No obstante, actualmente, el proyecto abarca librerías que ayudan al análisis de enfermedades, redes metabólicas, el estudio de las ómicas en cualquier tipo de instrumento entre muchas otras aplicaciones. Una característica de los datos ómicos en específico es la metadata, es decir la información acerca de los datos, por lo que muchas librerías han optado por un lenguaje de programación orientado a objetos, la cual permite guardar variables y los datos de dichas variables. Un ejemplo en particular es el objeto *ExperimentData*. Por otro lado, *Bioconductor*, mediante interfaces de programación para aplicaciones (API, por sus siglas en inglés), permite anotar datos biológicos con importantes bancos de datos públicos, como lo puede ser RefSeq (Sepulveda, 2020).

2.9.1 *DESeq2*

A diferencia de otros métodos, como *EdgeR* o *Limma*. Al igual que *EdgeR*, se asume que los datos siguen una distribución NB, con una media la cual consta del logaritmo de un modelo de regresión lineal. A partir de estos supuestos sigue un tratamiento empírico bayesiano. Cabe destacar que se realizan dos tipos de pruebas, la LRT (Likelihood Ratio Test) y la prueba de Wald. El primero se utiliza para corroborar el modelo de la función de verosimilitud, mientras que el segundo se utiliza para realizar los contrastes de interés. *DESeq2* realiza análisis DGE mediante modelos lineares generalizados (GLM) para lo cual parte de una matriz de conteos con una fila para cada gen y una columna para cada muestra.

Prueba de Wald

Una vez que los GLM se ajustan a cada gen, se puede probar si cada coeficiente del modelo difiere significativamente de cero. En concreto, se prueba la hipótesis en el que el coeficiente no es necesario. *DESeq2* computa el error estándar para cada estimación del cambio logarítmico (LFC), en otras palabras, de la significancia biológica.

2.9.2 *Limma*

Es otra librería del repositorio *Bioconductor*. Originalmente pensada para experimentos de *Microarrays*. Sin embargo, con el surgimiento de la secuenciación de nueva generación (NGS), tuvo que actualizarse principalmente con la función *Voom*, la cual se describirá más adelante.

La recomendación actual de acuerdo con la guía del usuario de *Limma* es usar la normalización TMM *(Media truncada de M valores)*, mencionada anteriormente, del paquete *EdgeR*, y la transformación de los datos a conteos por millón. La normalización establece la relación de la media con la varianza para computar pesos a nivel observacional. Dichos pesos serán

usados para el modelo estadístico posterior (Bottomly, y otros, 2011). Entonces, el flujo para realizar un análisis DGE según las mejores prácticas de *Limma* son: (1) a partir de la matriz de conteos, filtrar los datos como se ha mencionado para *EdgeR*, después (2) calcular los factores de normalización mediante TMM, (3) computar la transformación *logCPM* y la estimación de la media con la varianza, con tal de obtener los pesos para el modelo lineal general. (4) Con la función *lmFit*, se realiza dicho modelo estadístico.

Un modelo lineal general, en pocas palabras, realiza una regresión lineal simple para cada gen (ANOVA si las covariables son variables categóricas). Es decir, es un modelo en el que los residuos se suponen normales. Para lograr esto, a los coeficientes que se deben de estimar, se le asignan los pesos estimados mediante la función *Voom* (5). Tomando en cuenta el modelo lineal general, a partir de una realidad que se asume, es decir, que existen genes DE, dadas unas condiciones; dado el conjunto de datos, ¿esta realidad existe? De igual manera, se utiliza el teorema de Bayes para responder a dicha pregunta. Es decir, el teorema de Bayes, relaciona estas dos preguntas, mediante lo que se conoce como el *prior* y la función de verosimilitud. Donde la segunda, es el modelo lineal general, y el *prior* se establece a partir de los datos, respondiendo así a la pregunta en cuestión (Chamorro, 2019).

Normalización para *Limma*: media recortada de normalización de valores M (TMM)

El principal propósito de la normalización por la normalización recortada es la media después de quitar un cierto porcentaje a los datos por encima y por debajo. En realidad, el método TMM no normaliza las lecturas, como se ha comentado previamente, en realidad calcula factores de normalización. De manera que, cuando se calculan los conteos por millón, se escalan precisamente por el tamaño de los factores, que provienen de la librería total. La idea, es que los genes pueden tener valores más altos de lecturas, debido a razones técnicas, y dichos sesgos no

son idóneos para computar el tamaño de la librería. Se utiliza la suma total de las lecturas de los genes para obtener el tamaño de la librería, de manera que TMM recorta la mayoría de los genes más variables. Los valores M, del método de medias recortadas de los valores M, son los *log fold change*, es decir la significancia biológica entre cada muestra y una referencia. Después se calcular los factores de la normalización para ajustar el tamaño de la librería para computar las cuentas por millón en forma logarítmica (*logCPM*).

Test estadístico empírico de Bayes

Según Smyth (2004), y Phipson y colaboradores (2016), la idea detrás del método empírico de Bayes es utilizar diversas funciones para comprimir las varianzas residuales genéticas hacia un valor común. En este test en lugar de suponer alguna distribución para el *prior* en el teorema de Bayes, es la propia distribución de los datos que hacen que formen dicha distribución. En una primera instancia, se calcula la función de verosimilitud, la cual sigue unos residuos normales a partir de una regresión lineal ponderada. Y mediante los propios datos se calcula el *prior*. La multiplicación de ambos es proporcional al posterior, es decir, lo que supone que se observa dados los datos.

2.9.3 *EdgeR*

Inspirado en los análisis de *Microarrays*, *EdgeR*, toma como inicio la matriz de conteos o de lecturas, donde se asume que los datos son sobredispersos, tal como se ha comentado anteriormente. Dicha matriz se construye tal que las columnas son las muestras y las filas los transcritos. Donde un gen, en una muestra se modela tal que la lectura de ese dato en específico proviene de una distribución NB. En cambio, el parámetro de la media viene dado por el tamaño de la librería, es decir, el número total de lecturas de ese gen, junto a la abundancia relativa del gen

en el grupo experimental que se está estudiando; con una variable de dispersión del gen en cuestión.

Cabe mencionar que la NB (distribución binomial negativa), se reduce a una Poisson, cuando el parámetro de la dispersión es cero. Dicho parámetro representa el coeficiente de variación biológica, y *EdgeR*, es capaz de separar dicha variación de la variabilidad técnica. Es más, *EdgeR* estima la dispersión por gen mediante una función máxima de verosimilitud condicionada al número total de lecturas para el gen en cuestión.

De hecho, *EdgeR*, utiliza un método para conocer la abundancia de un gen, conocido como procedimiento empírico de Bayes. Es decir, a partir del teorema de Bayes, se puede conocer que existe una realidad dada por los datos, la cual es explicada por una función de verosimilitud. La realidad en términos Bayesianos es conocida como *prior*. Normalmente en la metodología de Bayes, el *prior*, se establece como una función de distribución conocida, cuyos parámetros hay que estimar; sin embargo, en los procedimientos empíricos, dicha distribución viene dada por los datos. De manera que dicho procedimiento, recorta la dispersión hacia un valor consenso, tomando la información de los genes.

Finalmente, el análisis DE, se realiza con un test análogo al test de Fisher exacto, pero adaptado para datos sobredispersos. Es decir, la hipótesis nula significa que las condiciones no son diferentes entre ellas.

Procedimiento para realizar análisis DGE en *EdgeR*

En una primera instancia, para que el resultado DE sea efectivo, se necesitan filtrar aquellos genes a través de las muestras con un determinado umbral de lecturas, normalmente 10 lecturas. Después, es necesario calcular los factores de normalización mediante el método TMM *(Media truncada de M valores)*, por consiguiente, los datos se normalizan en conteos por millón de forma

logarítmica. TMM calcula factores teniendo una muestra de referencia, calculando el *log fold change* entre cada muestra y la referencia. Posteriormente se calcula la dispersión, es decir la variabilidad técnica y biológica, mediante la maximización de la función de verosimilitud de la binomial negativa (NB) para así obtener parámetros necesarios. Finalmente, se realiza el análisis DE. Dicho análisis se computa respecto a las diferencias en las medias de los genes, con un análogo al test de Fisher; donde la hipótesis nula es aquella donde no existen diferencias entre las medias.

CAPÍTULO 3

MATERIALES Y MÉTODOS

3.1 Modelado de datos de *RNA-Seq*

Los datos de expresión en técnicas de secuenciación de la expresión masiva en paralelo, tratan de unas matrices de conteo. Dichas matrices provienen del alineamiento de millones de fragmentos, producto de la técnica subyacente. Estos fragmentos son cortos, que pueden variar de 70 a 150 pb, dependiendo del kit. Cada uno de estos millones de fragmentos se alinean con el genoma de referencia, por ejemplo, con el genoma hg38. Cabe mencionar que cada uno de estos fragmentos pueden alinearse más de una vez en un rango de posiciones, lo cual se le conoce como profundidad de cobertura en cada par de bases. Al número de fragmentos alineados a un transcrito en específico se le conoce como el número de lecturas. Y al número total de lecturas se le conoce como el tamaño de la librería.

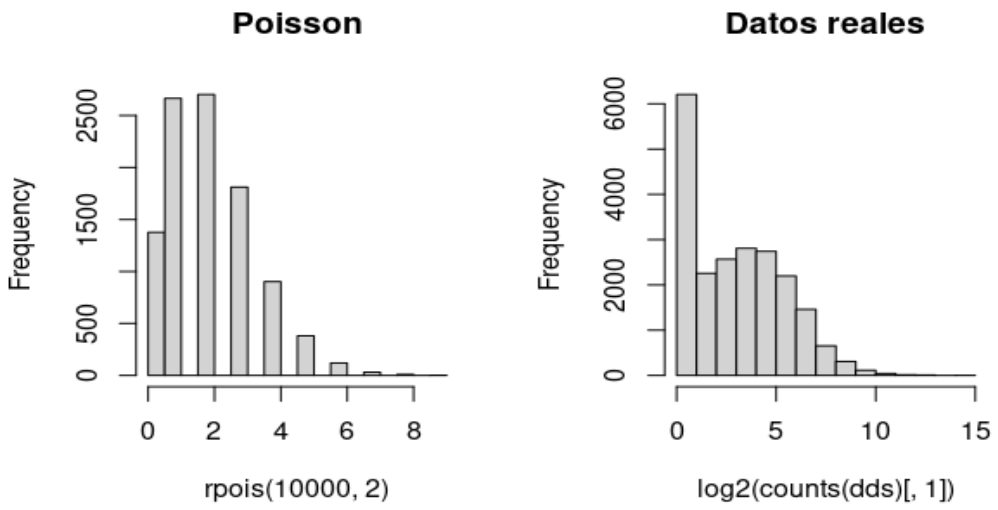
Dado que las lecturas se basan en el conteo, no pueden seguir una distribución normal. Estas lecturas o contajes son datos discretos, por lo que corresponden a valores enteros no negativos, que representan el número de veces que ocurre un evento. Además, estos datos pueden ser contados de una forma precisa y no tienen un rango infinito.

Los datos de *RNA-Seq* representan una gran cantidad de ARN y la probabilidad de extraer una transcripción en particular es muy pequeña; los genes poco expresados tienen una varianza mucho mayor que los genes más expresados. Para modelar estos datos de conteo se considera adecuado emplear la distribución NB, puesto que el modelo binomial negativo permite que la varianza sea mayor que la media, es decir, maneja recuentos sobredispersos para capturar la variación en el conjunto de datos. En cambio, en la distribución de Poisson, la media y la varianza

son iguales. Por lo cual, como alternativa a la distribución de Poisson, la distribución NB es más flexible al permitir que la media y la varianza sean diferentes (Liu et al., 2022).

Figura 2.

Histogramas de las distribuciones de Poisson con datos simulados y datos reales



a) b)

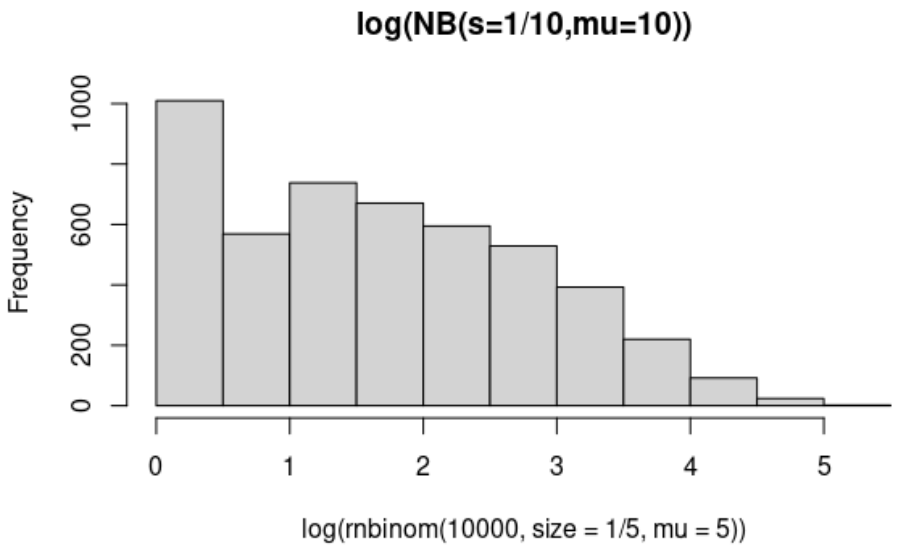
Nota. Los histogramas representan el comportamiento de los datos que siguen una distribución de Poisson. a) Histograma con datos simulados que siguen una distribución de Poisson. b) Histograma con datos reales que siguen una distribución de Poisson. Fuente. Elaboración propia.

Al analizar el comportamiento de los datos, se trata de buscar un método adecuado que se ajuste a la distribución que siguen los mismos. En la Figura 2, se puede observar que ambos gráficos tienen una figura similar. Sin embargo, en el histograma con los datos reales se observa un pico alrededor de 0. A pesar de seguir una distribución de Poisson, el bulto seguido después del 0 indica una sobredispersión de datos. Por lo que el objetivo de *EdgeR* y *DESeq2*, es modelar este bulto. Como se mencionó anteriormente, se pueden modelar estos datos considerando que los

genes provienen de la suma de varias distribuciones de Poisson, y a su vez la suma de estas distribuciones da como resultado una distribución NB.

Figura 3.

Histograma que sigue una distribución de Poisson



Nota. El histograma sigue una distribución de Poisson con un parámetro de sobredispersón inversamente proporcional a la media. Se puede observar cierta similitud con el histograma de datos reales. Esta es la base para conocer la metodología que se empleará. Fuente. Elaboración propia.

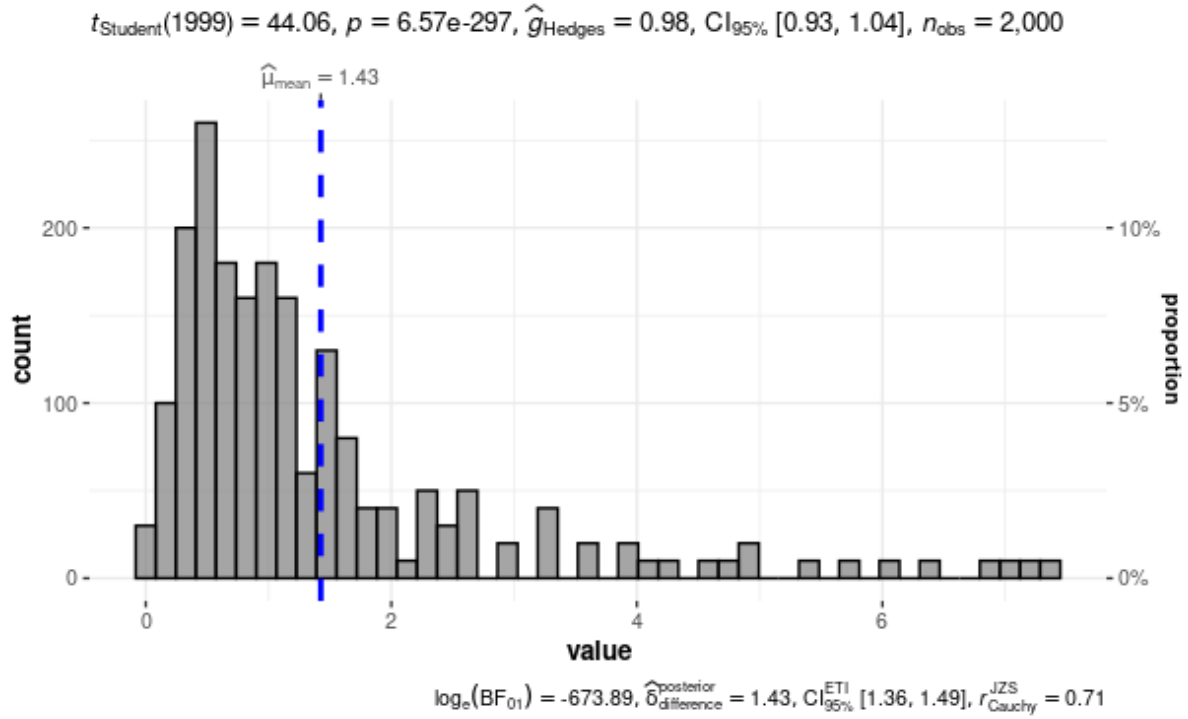
Anteriormente, se mencionó que el parámetro de dispersión es inversamente proporcional a la varianza. Además, considerando que cada gen sigue una distribución de Poisson, se puede modelar matemáticamente la media de dicha distribución con un modelo matemático lineal, el cual se encuentra como exponente de una base, en este caso 2. En el modelo de Poisson existe un número de eventos en un intervalo específico, hasta que ocurre lo que se desea observar. Por tal motivo, tiene coherencia con la binomial negativa, en la que se "apuesta" por los eventos que no son hasta que se puede observar el que realmente es.

Por lo tanto, se puede modelar la media de un caso-control, que sigue el modelo lineal, si se asume que $\beta \sim N(0,1)$.

```
beta <- matrix(rnorm(200),ncol=2,nrow=100) ## 200 genes

caso_control <- as.factor(rep(c("A","B"),each=10)) ## 20 muestras

diseno <- model.matrix(~caso_control)

media <- t(2^(diseno %*% t(beta)))

colnames(media) <- rep(c("A","B"),each=10)

rownames(media) <- paste0("gen_",rep(1:100))

media.melt <- reshape2::melt(media)

ggstatsplot::gghistostats(media.melt,value)
```

Figura 4.

Histograma de la media que sigue una distribución binomial negativa (NB)



Nota. El histograma representa una modelización de la media de un caso-control de la distribución NB.

Fuente. Elaboración propia.

Como la varianza es inversamente proporcional al parámetro de dispersión:

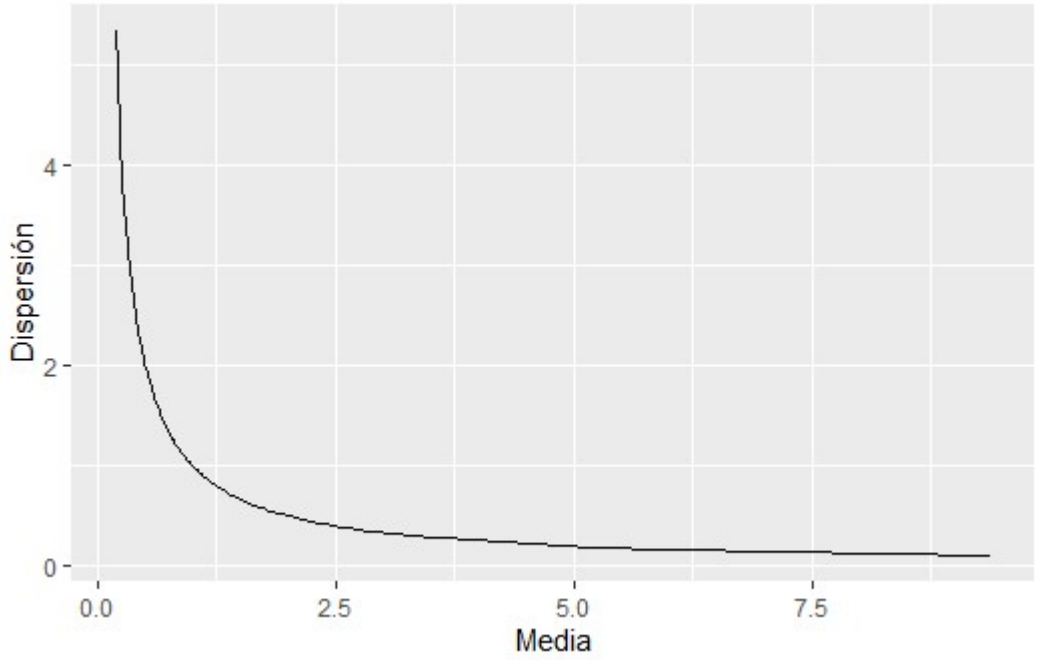
$$Var(Y|x_i) = \mu_i(1 + \mu/\alpha) \rightarrow \alpha \sim 1/\mu + C \quad (12)$$

Donde C es una constante, y para estimar la relación entre la dispersión y la media de los genes de una muestra, se creó la Figura 5.

```
media_gen1 <- media[,1] ## media para una observación
alfa <- 1/media_gen1
dispersion_media <- data.frame(media=media_gen1,dispersion=alfa)
ggplot(dispersion_media,aes(media,dispersion))+geom_line()+ggtitle("Dispersion
-Media a través de los genes en una muestra")
```

Figura 5.

Relación entre la dispersión y la media



Nota. Gráfica que representa la relación entre la dispersión y la media a través de los genes de una muestra.

Fuente. Elaboración propia.

3.2 Simulación de datos de *RNA-Seq*

La simulación de datos de *RNA-Seq* fue generada con el paquete *compcodeR*, el cual contiene funciones para simular datos de conteo e interfaces a varios paquetes para realizar el análisis DGE. Se utilizó la función `generateSyntheticData` cambiando parámetros en relación con el número de muestras *(n)*, el número de transcritos totales *(m)*, y los genes diferencialmente expresados *(diffexp)*.

```
funcion_sim <- function(n, m, diffexp) {
  ## funcion para determinar la relación media dispersión
  dispMeanRel <- function(x) 4/x +0.1
  ## Nuestras condiciones
  ## Estamos diciendo que nos tome un modelo balanceado
  ## es decir el mismo número de muestras para ambos grupos
  colData <- DataFrame(condition = factor(rep(c("A", "B"),
                                              times = c(
                                                ceiling(n / 2), floor(n /2)
                                              ))))

  set.seed(123456) ## fijamos semilla de aleatorización

  beta <- rnorm(m) ## tenemos en cuenta sólo un parámetro
  mu <- t(2 ^ ((beta))) ## computamos la media

  dispersion <- dispMeanRel(2 ^ (beta)) ## computamos la dispersión
  ## generamos los datos con compcodeR

 datos <-
    generateSyntheticData(
      "sim_counts",
      n.vars = m,
      samples.per.cond = ceiling(n / 2),
      n.diffexp = diffexp,
      relmeans = mu,
      dispersions = dispersion,
      filter.threshold.total = 0
```

```
    )

  return(datos)

}
```

El núcleo de la función es `compcodeR:generateSyntheticData`. Sin embargo, como se requiere observar diferentes barridos; entre ellos, la media, el número de variables y el de genes DE. Lo que se encuentra arriba de la función establece la relación de la media con la varianza, según coeficientes; los cuales se pretenden determinar con los métodos de *EdgeR* y *DESeq2*. Sin embargo, estos coeficientes se determinan a *priori* siguiendo una distribución normal con el número de transcritos totales DE. Esto se logró al modelar la media, de acuerdo a los paquetes anteriormente mencionados.

Luego, se aplicó dicha función variando el número de muestras de 6 a 56 en pasos de 6. Posteriormente, fijando los valores óptimos del número de muestras (n), para cada paquete, se varió el número de transcritos totales (m) y posteriormente fijando los valores óptimos de m y n, el número de genes DE (diffexp). En función de esta aproximación, se observa cuáles son los parámetros óptimos para cada paquete. Después, se realizó el análisis DGE con los tres paquetes para los datos reales escogidos. Para ilustrar los resultados, se realizó un diagrama de Venn, que indicaba las interacciones y los genes comunes entre los paquetes, para así determinar cuál es el paquete que más genes aporta.

Por otro lado, se realizaron simulaciones con los tres paquetes, considerando el número de observaciones de los datos originales, fijando el número de transcritos totales con esos mismos datos, y variando de valores DE del mínimo encontrado con el máximo encontrado.

3.3 Determinación de parámetros óptimos

Se determinarán los parámetros óptimos en función de las variables *n, m* y *diffexp,* realizando el análisis DGE para *DESeq2*, *Limma* y *EdgeR*. Posteriormente se generaron matrices de confusión para la obtención de curvas ROC y AUC. Cabe mencionar que un paso previo al análisis DGE es el filtrado de los datos; no obstante, por simplicidad solo se lo realizó en *DESeq2,* puesto que es un flujo de trabajo ya elaborado.

Para obtener el área bajo la curva AUC simplemente se integra. Por lo cual, al ser valores discretos se sumaron pequeños rectángulos de la siguiente manera:

$$AUC = Sens \sum \Delta Esp \tag{13}$$

Además, los parámetros óptimos se escogieron según índice de Youden con valores de *J* máximos.

$$J = sensibilidad + especificidad - 1 \tag{14}$$

$$J(c) = Se(c) + Sp(c) - 1 = Se(c) - \big(1 - Sp(c)\big) \tag{15}$$

Para establecer los parámetros óptimos de *n, m* y *diffexp,* se realizó un barrido con un amplio rango cambiando los valores de *n, m* y *diffexp.* Estos parámetros se establecieron para cada flujo de trabajo de los paquetes *DESeq2*, *Limma* y *EdgeR.*

```
## variando número de muestras(n)
n <- seq(6,56,6)
m <- 70000
diffexp <- 700
```

```
## variando número de transcritos totales(m)
n <- as.numeric(unique(names(op_n_dds))) #parámetro óptimo de n
m <- seq(60000,120000,4000)
diffexp <- 700

## variando número de genes (m)
n <- as.numeric(unique(names(op_n_dds)))#parámetro óptimo de n
m <- as.numeric(unique(names(op_m_dds)))#parámetro óptimo de m
diffexp <- seq(1000,10000,1000)
```

El análisis DGE con cada paquete se basó en los flujos de trabajo estándar detallados a continuación.

3.3.1 DESeq2: flujo de trabajo estándar para análisis DGE

En primer lugar, se creó el objeto `DESeq-DataSet`. Al tener una matriz de conteos se usó la función `DESeqDataSetFromMatrix`, además de añadir la información sobre las muestras (las columnas de la matriz de conteo) como `data.frame` y la fórmula de diseño. Es importante que las columnas de la matriz de conteo y las filas de la columna de datos estén en el mismo orden, debido a que *DESeq2* no hace conjeturas sobre qué columna de la matriz de conteo pertenece a qué fila de los datos de la columna, estos datos se proporcionaron a *DESeq2* ya en un orden coherente (Love, Anders, & Huber, 2023).

Si bien no es necesario filtrar previamente los genes de conteo bajo antes de ejecutar las funciones *DESeq2*, hay dos razones que hacen que el filtrado previo sea útil; al eliminar las filas

en las que hay muy pocas lecturas, se reduce el tamaño de la memoria del objeto de datos *dds*, y se aumenta la velocidad de las funciones de transformación y prueba dentro de *DESeq2*. También puede mejorar las visualizaciones, ya que las características sin información para la expresión diferencial no se trazan. Se realiza un filtrado previo mínimo para mantener solo las filas que tienen al menos 10 lecturas en total. Se debe tomar en cuenta que se aplica automáticamente un filtrado más estricto para aumentar la potencia a través de un filtrado independiente sobre la media de recuentos normalizados dentro de la función de resultados (Love, Anders, & Huber, 2023).

Para el análisis de expresión diferencial en *DESeq2* se dispuso de una función `DESeq` que realiza de forma predeterminada todos los pasos necesarios. En el estudio realizado por Chamorro (2019), se mencionó que estos pasos incluyen:

- La normalización de los datos mediante la estimación de los tamaños de muestra y los factores de normalización.
- La estimación de la dispersión.
- El ajuste de los datos a un modelo lineal generalizado (GLM) binomial negativo.
- La comprobación de la expresión diferencial de cada transcrito mediante una prueba paramétrica de Wald y la validación del modelo con LRT.

Las tablas de resultados se generaron utilizando la función `results,` que extrae una tabla de resultados con cambios de log2, valores de *p* y valores de *p ajustados.* Sin argumentos adicionales a los resultados, el cambio de pliegue log2 y el valor *p* de la prueba de Wald son para la última variable en la fórmula de diseño, y si esto es un factor, la comparación es el último nivel de esta variable sobre el nivel de referencia (Love, Anders, & Huber, 2023).

Dicho flujo se realizó dentro una función, debido a que se considera un código útil que se debía reutilizar varias veces:

```
funcion_dds <- function(n, m, diffexp) {
  colData <- DataFrame(condition = factor(rep(c("A", "B"),
                                              times = c(
                                                ceiling(n / 2), floor(n /2)
                                              ))))
  simulacion <- funcion_sim(n = n, m = m, diffexp = diffexp)
  genes.de.actual <-
    simulacion@variable.annotations$differential.expression
  design <-    as.formula("~ condition", env = .GlobalEnv)
  dds <-
    DESeqDataSetFromMatrix(simulacion@count.matrix,
                           colData = colData,
                           design = design)
  dds <- DESeq(dds)
  res <- results(dds)
  dgenes.de <- as.factor(ifelse(res$pvalue < 0.05, 1, 0))
  genes.de.act <-
    as.factor(ifelse(
      simulacion@variable.annotations$differential.expression == 1,
      1,0))
  miconf <- caret::confusionMatrix(dgenes.de, genes.de.act)
  miconf$overall[1]
```

```
  miconf$byClass[1:2]
  return(c(
      acc = miconf$overall[1],
      sens = miconf$byClass[1],
      esp = miconf$byClass[2]
    ))
}
```

Al obtener la matriz de confusión, no se sabe si es un gen DE por el método o en realidad está DE. Tan solo se comparó el número de genes DE. Donde el número de genes DE, se tomó con un valor absoluto mayor a 1 que el *log fold change*, y una significancia estadística con una confianza del 95%. Posteriormente, para cada barrido, se obtuvieron los parámetros de precisión, sensibilidad y especificidad para construir las curvas ROC y determinar el AUC.

3.3.2 EdgeR: flujo de trabajo estándar para análisis DGE

En base a los estudios de Chamorro (2019), el análisis en *EdgeR*, se realizó a partir de un objeto `DGEList`. Los métodos y funciones utilizadas tanto para el filtrado como para la normalización son comunes con los de *Limma*. Puesto que, en el filtrado se ocupa la función `cpm` y en la normalización se utiliza la función `calcNormFactors`, la cual se basa en el método TMM.

Para la estimación de las dispersiones de cada transcrito se debe considerar la variabilidad total de todos los genes, debido a que el paquete *EdgeR* emplea una distribución NB y proporciona una forma de estimar las dispersiones, conocida como qCML (verosimilitud máxima condicional ajustada por cuantiles). En general, este método se aplica a experimentos que poseen un solo factor,

además de ser más confiable y funcionar mejor en el caso de tener muchas muestras pequeñas que posean una dispersión común, por lo que es el óptimo para trabajar con datos de experimentos de *RNA-Seq*. La función `estimateDisp` permitió una visualización de la dispersión gen a gen y la dispersión común. Al conocer el valor de la dispersión común, se pudo obtener el coeficiente de la variación biológica (BCV). Este valor se relaciona con el número de genes DE. Luego, se realizó una gráfica de BCV por medio de la función `plotBCV`.

De forma similar a *DESeq2*, se estableció el flujo de trabajo para *EdgeR*.

```
funcion_dge <- function(n, m, diffexp) {
  simulacion <- funcion_sim(n = n, m = m, diffexp = diffexp)
  group <-factor(rep(c("A", "B"),
                       c(ceiling(n / 2), floor(n / 2)
                         )))
  dge <-DGEList(simulacion@count.matrix,group=group)
  dge <- calcNormFactors(dge)
  design<-model.matrix(~0+group)
  dgedis <- estimateDisp(dge)
  etest<-exactTest(dgedis,pair=c("B","A"))
  dgenes.de <- as.factor(ifelse(etest$table$PValue < 0.05, 1, 0))
  genes.de.act <-
    as.factor(ifelse(
      simulacion@variable.annotations$differential.expression == 1,
      1,
```

```
      0
    ))
  miconf <-confusionMatrix(dgenes.de, genes.de.act)


  miconf$overall[1]
  miconf$byClass[1:2]
  return(c(
      acc = miconf$overall[1],
      sens = miconf$byClass[1],
      esp = miconf$byClass[2]
    ))
}
```

3.3.3 Limma: flujo de trabajo estándar para análisis DGE

Una vez obtenida la matriz de conteos de lecturas, con filas para los genes y columnas para las muestras, en primer lugar, se creó un objeto `DGEList` utilizando el paquete *EdgeR*. El siguiente paso fue eliminar las filas que tenían sistemáticamente recuentos cero o muy bajos, donde `filterByExpr` es un código del paquete *EdgeR* que cumple esta función. Posteriormente, se aplicó una normalización de escala a los recuentos de lecturas *RNA-Seq*, utilizando el método de normalización TMM mediante `calcNormFactors`.

El método de la estimación media-varianza, consiguió establecer una serie de pesos, que se utilizan posteriormente para el análisis DGE. Mientras que transformó los datos, acorde a los factores de escala para obtener cuentas por millón logarítmicas. La transformación *Voom* utilizó

la matriz de diseño del experimento y produce un objeto `EList`. Después, se aplicaron los comandos habituales de *Limma* empleados en expresión diferencial mediante la función `lmFit`, para ajustar los modelos lineales generales por pesos, así se obtuvo la función de verosimilitud. Al multiplicar lo anterior por el *prior*, obtenido empíricamente con la función `eBayes`, se realizaron los contrastes de las condiciones del experimento (Law, Chen, Shi, & Smyth, 2014).

De forma similar, se realizó otra función, con el *pipeline* de *Limma*.

```
funcion_Limma<- function(n, m, diffexp) {
  simulacion<- funcion_sim(n = n, m = m, diffexp = diffexp)
  group <- factor(rep(c("A", "B"),
                      times = c(
                        ceiling(n / 2), floor(n / 2)
                        )))
  lim <-DGEList(counts= simulacion@count.matrix, group=group)
  lim <- calcNormFactors(object=lim)
  design<-model.matrix(~0+group)
  colnames(design)<-levels(lim$samples$group)
  cont.matrix<-makeContrasts(B-A, levels=design)
  v<-Voom(counts=lim,design=design)
  vfit <- lmFit(object=v, design=design)
  fit.cont<-contrasts.fit(fit=vfit, contrast=cont.matrix)
  efit <- eBayes(fit.cont)
```

```
 dgenes.de <- as.factor(ifelse(efit$p.value < 0.05, 1, 0))
  genes.de.act <-
    as.factor(ifelse(
      simulacion@variable.annotations$differential.expression == 1,
      1,
      0
    ))
  miconf <-confusionMatrix(dgenes.de, genes.de.act)
  miconf$overall[1]
  miconf$byClass[1:2]
  return(c(
      acc = miconf$overall[1],
      sens = miconf$byClass[1],
      esp = miconf$byClass[2]
    ))
}
```

3.4 Análisis DGE con datos reales

Se llevaron a cabo análisis DGE siguiendo los flujos de trabajo estándar para *DESeq2*, *Limma* y *EdgeR* a partir de la matriz de conteos obtenida de GEO (GSE192804: *ISG15 is associated with cervical cancer development*). De la matriz de conteos se consideraron 12 muestras en total, las cuales estaban divididas en muestras de tejido normal (NC) y tejido tumoral (TT). Se

omitieron los datos de tejido de paracarcinoma (TP), por motivos comparativos con las simulaciones.

3.5 Curvas ROC y el área bajo la curva (AUC)

Se realizó una simulación con un barrido en función de los genes DE obtenidos de los datos reales fijando valores de *n* y *m*, según los valores obtenidos del estudio. Se obtuvieron curvas ROC para *DESeq2*, *Limma* y *EdgeR* y se calcularon valores de AUC integrando la función computada.

```
#Curva ROC DESeq2

#lista obtenida del barrido con valores de precisión, sensibilidad y especificidad

dds.list_sim<-readRDS("dds.list_sim.rds")

csa_dds <- Reduce(rbind, dds.list_sim)

especificidad1 <- c(csa_dds[, 3])

sensibilidad1 <- c(csa_dds[, 2])


#Curva ROC Limma

lim.list_sim<-readRDS("lim.list_sim.rds")

csa_lim <- Reduce(rbind, lim.list_sim)

especificidad2 <- c(csa_lim[, 3])

sensibilidad2 <- c(csa_lim[, 2])


#Curva ROC EdgeR

dge.list_sim<-readRDS("dge.list_sim.rds")

csa_dge <- Reduce(rbind, dge.list_sim)

especificidad3 <- c(csa_dge[, 3])
```

```
sensibilidad3 <- c(csa_dge[, 2])


#Valores de AUC calculados para cada paquete
AUC_DESeq2 <- round(sum(sensibilidad1*diff(c(0, 1 - especificidad1))),2)
AUC_Limma <- round(sum(sensibilidad2*diff(c(0, 1 - especificidad2))),2)
AUC_EdgeR<-round(sum(sensibilidad3*diff(c(0, 1 - especificidad3))),2)
```

CAPÍTULO 4

RESULTADOS Y DISCUSIÓN

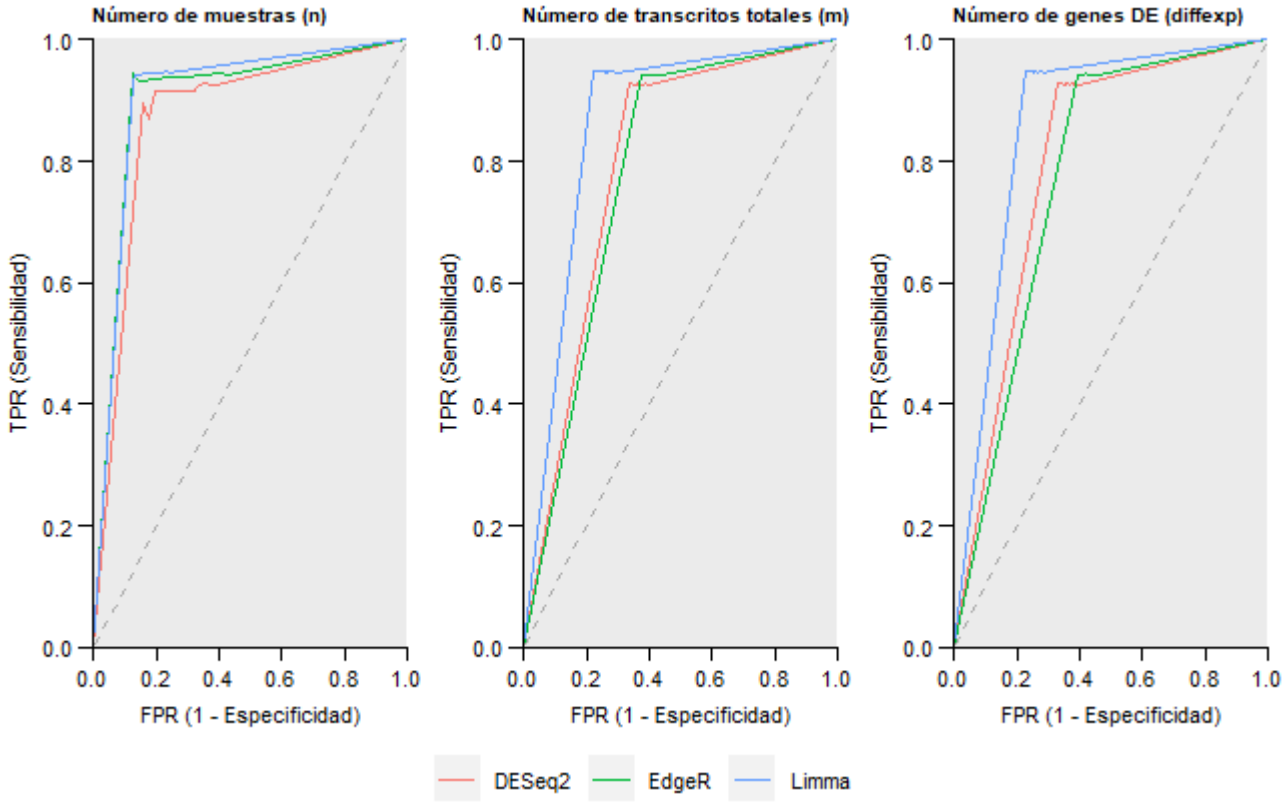
4.1 Parámetros óptimos para *DESeq2, Limma* y *EdgeR*

En una primera instancia, se realizó el experimento con el número de muestras (n), con tal de observar cuál de los tres paquetes era el mejor para este parámetro, considerando el mejor como el número más alto. Posteriormente, se realizó un barrido del número de transcritos (m) de igual manera considerando para cada paquete cual es el mejor, para finalmente considerar el número de genes DE (diffexp). Se tomó este parámetro al último con tal de ser coherentes, en el sentido que se desconoce de antemano cuántos genes DE habrá en un conjunto de datos reales.

A partir de los diferentes barridos en función del número de muestras (n), los transcritos totales (m), y los genes DE (diffexp) con los datos simulados matemáticamente se computaron curvas ROC para evaluar el rendimiento de cada paquete e identificar correctamente los genes DE (verdaderos positivos) y al mismo tiempo minimizar el número de genes no DE, identificados erróneamente como DE (falsos positivos). Por lo tanto, para la simulación propuesta se determina en la Figura 6 que la curva de *Limma*; seguida de *EdgeR* y *DESeq2*; se encuentra para todas las variables analizadas por encima de las otras, lo que indicaría los mejores resultados, es decir, irrespectivamente del número de muestras, *Limma* clasifica mejor los genes DE.

Figura 6.

Comparación de curvas ROC en función al número de muestras (n), transcritos totales (m), y los genes DE (diffexp) con datos RNA-Seq simulados matemáticamente para los tres paquetes.



Fuente. Elaboración propia.

Además, se determinaron los valores óptimos para *DESeq2*, *Limma* y *EdgeR*. Los puntos de corte óptimos para las diferentes variables se calcularon obteniendo el mayor índice de Youden, los cuales se pueden ver a continuación en la Tabla 1:

Tabla 1.

Comparación de variables, valores de corte óptimos para cada variable, métricas correspondientes a falsos positivos (FPR) y verdaderos positivos (TPR) con los diferentes paquetes empleados en el análisis DGE.

Variables	**Valores óptimos**	**FPR**	**TPR**	**Paquete**
n	48	0,384286	0,926580	*DESeq2*
m	120000	0,391429	0,926697	*DESeq2*
diffexp	500	0,390000	0,929063	*DESeq2*
n	48	0,278571	0,948932	*Limma*
m	120000	0,290000	0,946756	*Limma*
diffexp	100	0,300000	0,947706	*Limma*
n	48	0,424286	0,943131	*EdgeR*
m	72000	0,437143	0,942567	*EdgeR*
diffexp	400	0,457500	0,941257	*EdgeR*

Fuente. Elaboración propia.

Por lo que respecta a los índices de Youden, en la Tabla 1 se pueden observar los parámetros óptimos de cada variable, dependiendo el paquete utilizado para el análisis DGE. Por lo tanto, en relación con el número de muestras (n) se evidenció que mientras mayor es el número de muestras, mejor es el desempeño de los tres paquetes, esto sucede en todo modelo estadístico, ya que mientras más muestras existan, la ley de los grandes números entrará más en vigor. Dicha ley nos dice, que cuanto más alto es el número de muestras en cualquier prueba estadística, la distribución en la que los datos yacen, será cada vez más parecida a una distribución normal o Gausiana (Dinov, Christou, & Gould, 2009).

Por lo que concierne al barrido en el número de transcritos totales (m); independientemente de *Limma*; *DESeq2* clasifica mejor la tasa de verdaderos positivos. Y, por último, fijando los valores óptimos encontrados para cada paquete en n y m; precedido de *Limma*; *DESeq2* clasifica mejor a los verdaderos positivos respecto a *EdgeR*. Por otra parte, Flicek et al. (2012) mencionan que, aunque hay menos de 22 000 genes que codifican proteínas conocidas en el genoma humano, se transcriben en más de 140 000 transcritos diferentes. Por lo tanto, respecto al número de transcritos totales (m) se puede mencionar que *EdgeR* considera un valor óptimo con un número mucho menor. Es decir, *EdgeR* sería un buen candidato para un método en el que las lecturas hayan sido pobres en el sentido, que no se haya alineado de manera correcta o en el que las condiciones del experimento no determinen un número de transcritos alto. No obstante, la tasa de falsos positivos es muy alta. Comparando con *DESeq2* y *Limma*, ambos paquetes son óptimos si, por ejemplo, después del filtrado de genes, se obtiene un número de transcritos alto, siendo en este caso mejor *Limma* con una tasa de falsos positivos que no alcanza al 30%.

Cabe mencionar, que un barrido en el número de genes diferencialmente expresados (diffexp) se puede determinar dos aspectos. El primero, que *Limma* es más conservador que *DESeq2* y *EdgeR*, siendo estos últimos más laxos, a la hora de identificar genes DE. La segunda observación, es que *Limma* tiene una tasa de falsos positivos menor, lo cual tiene sentido al ser más conservador. No obstante, *DESeq2*, con solamente un 10% más de falsos positivos obtiene un número más alto de genes DE.

4.2 Análisis DGE con datos reales

Mediante el análisis DGE con el conjunto de datos reales se pudo obtener una comparativa de los tres métodos. Se clasificaron los genes como sobreexpresados e infraexpresados, respecto al nivel

de expresión en el grupo de referencia o control. La comparativa permitió determinar que el paquete con el que se obtuvo un mayor número de genes DE fue *DESeq2*. Por el contrario, se obtuvo un número bajo de genes con el paquete *Limma*.

Tabla 2.

Número total de genes diferencialmente expresados, sobreexpresados e infraexpresados obtenidos del análisis de expresión diferencial con los paquetes DESeq2, Limma y EdgeR.

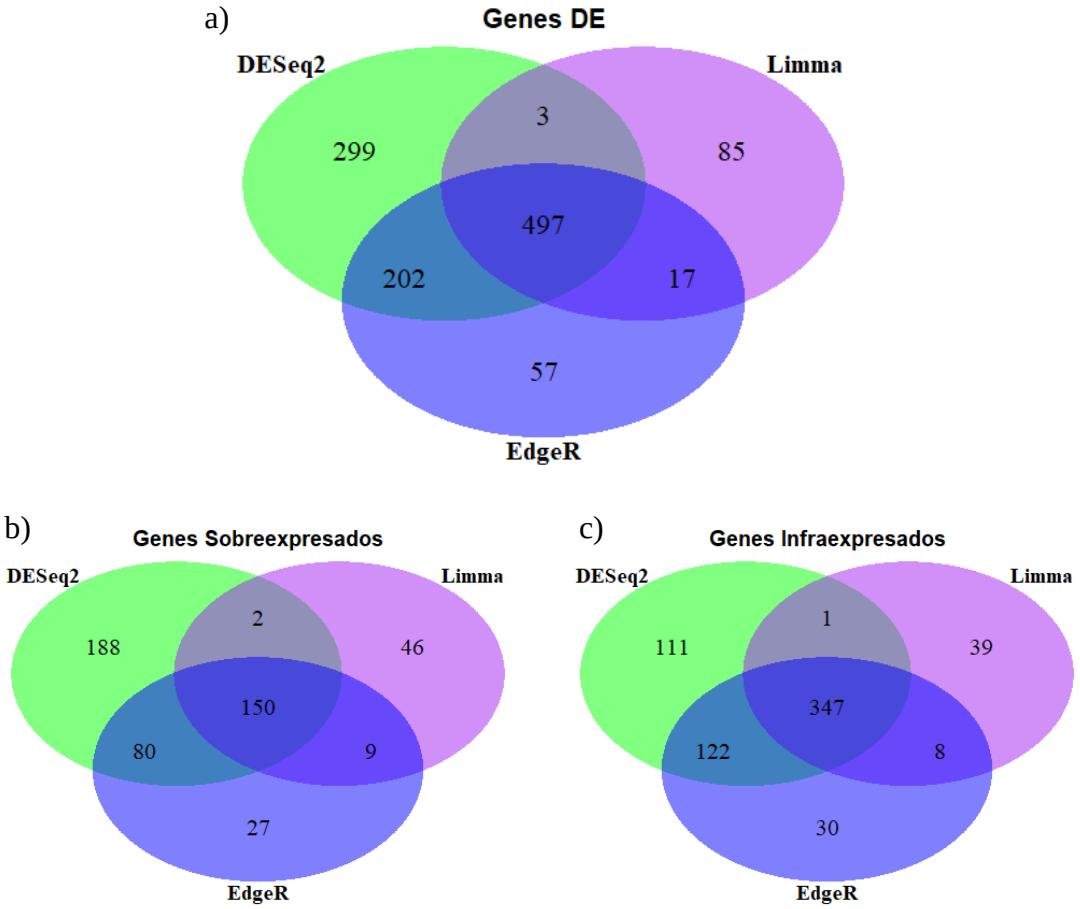
Genes	***DESeq2***	***Limma***	***EdgeR***
DE	1001	602	773
Sobreexpresados	420	207	507
Infraexpresados	581	395	266

Fuente. Elaboración propia.

Para comparar los datos obtenidos a partir del análisis DGE con los tres paquetes, se realizaron diagramas de Venn, en los cuales se pueden observar los genes DE comunes y distintos obtenidos por cada método.

Figura 7.

Diagramas de Venn para la visualización de los genes diferencialmente expresados comunes entre DESeq2, Limma y EdgeR.



Nota. Ilustración de genes DE comunes entre los tres paquetes mediante diagramas de Venn. a) Genes DE totales. b) Genes sobreexpresados. c) Genes infraexpresados. Fuente: Elaboración propia.

La región de intersección de los óvalos indica la presencia de genes compartidos en los conjuntos de datos analizados. Se puede observar que existe un gran número de genes comunes entre los tres paquetes, tanto para los genes DE, los sobreexpresados y los infraexpresados. Además, *DESeq2* identifica un mayor número de genes, en comparación con *EdgeR* y *Limma*. Por otro lado, se puede determinar que *Limma* se basa en un enfoque que conserva la mayoría de los

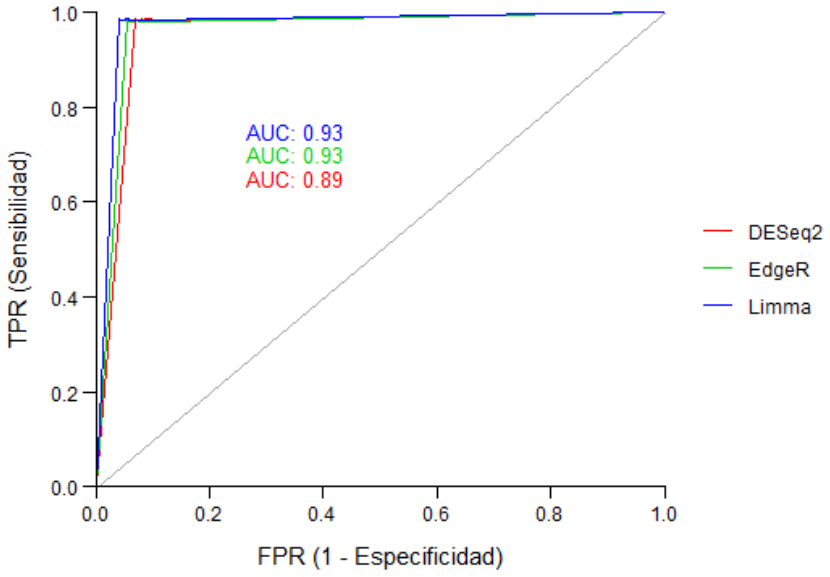
genes DE que se encuentran presentes en *DESeq2* y *EdgeR*. Mientras que *DESeq2* en un enfoque más amplio y laxo en el cual identifica más genes DE y *EdgeR* en una posición intermedia.

4.3 Análisis de Curvas ROC y el área bajo la curva (AUC)

A partir de la simulación con un barrido de genes DE, obtenidos de los datos reales se realizaron curvas ROC y se calcularon valores de AUC para los tres paquetes. El barrido se realizó considerando el menor y mayor valor de genes DE correspondientes a *Limma* con 602 y *DESeq2* con 1001, respectivamente. Para lo cual, se pudo observar los valores de AUC más altos con *Limma* y *EdgeR*, los cuales representan el mejor método para el análisis con el conjunto de datos reales, teniendo en cuenta el número de genes totales y el número de observaciones. Además, se puede observar que *Limma*, representa el punto de corte óptimo más alto, es decir, más cercano a la esquina superior izquierda, por lo tanto, se establece un equilibrio para la correcta identificación de verdaderos positivos y el rechazo de falsos positivos.

Figura 8.

Curva ROC de datos RNA-Seq simulados a partir de datos DE obtenidos de datos reales.



Fuente. Elaboración propia.

4.4 Discusión

El paquete con el que se identificó un mayor número de genes DE fue *DESeq2*; lo cual concuerda, con los estudios realizados por Chamorro (2019), en donde se mencionó que los paquetes que obtuvieron un mayor número de transcritos totales fueron los que se basaban en una distribución binomial negativa (NB), es decir, *DESeq2* y *EdgeR*. Por otra parte, en este estudio con el análisis de datos reales, se evidenció que *Limma* obtuvo un menor número de genes DE.

En cuanto al análisis de los parámetros óptimos obtenidos, tales como el número de muestras (n), número de transcritos totales (m) y genes diferencialmente expresados (DE), mediante una simulación de matrices de conteos; *Limma* resultó ser el método más apropiado debido a la baja tasa de falsos positivos que se ilustró en la Tabla 1.

Con respecto al conjunto de datos reales, *Limma* es el método más apropiado cuando se considera el número de transcritos totales y el número de muestras, puesto que al observar las curvas ROC se determinó que el punto de corte óptimo más cercano a la esquina superior izquierda es representado por *Limma*. De modo que, la tasa de verdaderos positivos es mayor, por tanto, la sensibilidad se maximiza. Por el contrario, la tasa de falsos positivos es menor, así se reduce la posibilidad de clasificar incorrectamente los casos negativos.

Además, se debe considerar que *DESeq2* posee un flujo de trabajo más optimizado, es decir, se realizan menos pasos para el análisis DGE en comparación con los otros paquetes. Sin embargo, el tiempo de cómputo es mayor. Por otro lado, el tiempo de ejecución *Limma* fue considerablemente menor, mientras que *EdgeR* se ubica en una posición intermedia, lo cual concuerda con Seyednasrollah, Laiho, & Elo (2015).

De hecho, también es importante mencionar que de acuerdo con la literatura del estudio realizado por Chamorro (2019), *Limma* es más conservador, en el sentido que identifica menos

genes DE. Esto puede indicar que *Limma* sería el método apropiado, si se quiere aplicar a la clínica diaria, mientras que *DESeq2*, es más apropiado para estudios de investigación

Se hace un especial énfasis en que cada método utiliza normalizaciones diferentes, modelos y pruebas estadísticas distintas. Si bien es cierto, que no se tomaron en cuenta las normalizaciones, o los distintos tipos de dispersiones o formas, que puedan tener las dispersiones de los datos, el presente trabajo, es una aproximación que los bioinformáticos deberían de tener en cuenta a la hora de analizar los datos de ésta ómica.

También cabe recalcar, que la simulación matemática realizada, se inspiró en la distribución NB, es decir, en el número de veces que no se acierta a ver un gen. No obstante, pueden existir otros autores que consideren una mejor forma de representar los datos de *RNA-Seq*, pero para este estudio en particular, se escogió esta distribución por el hecho de ser la más aceptada por la comunidad científica.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

Como conclusión del presente estudio, los paquetes empleados para el análisis DGE (*DESeq2*, *Limma* y *EdgeR*) son eficientes para identificar correctamente los genes diferencialmente expresados, aunque con diferentes tasas de verdaderos positivos y negativos. Cabe mencionar, que mientras se disponga de un mayor número de muestras, mejor será el estudio, irrespectivamente del método; sin embargo, esto se ve limitado por motivos económicos, poblacionales, etc. También se debe considerar el número de transcritos totales, siendo *DESeq2* y *Limma* los mejores para una amplia gama de transcritos. Aunque el número de transcritos en un organismo pueda ser conocido, debido a las variabilidades del estudio, ya sean técnicas o biológicas, dicho número cambiará.

Finalmente, con lo que respecta al número de genes DE, *Limma* es más conservador, obteniendo un menor número genes DE; mientras que *DESeq2* tiene un enfoque amplio y laxo, siendo el más apropiado para fines de investigación. Entre los aspectos a considerar se encuentra el tiempo de cómputo, para lo cual se evidenció que *Limma* fue considerablemente menor en comparación con *DESeq2*, sin embargo, *DESeq2* presenta un flujo de trabajo más optimizado.

En el presente estudio, se han considerado los paquetes más comunes para el análisis DGE *(DESeq2, Limma, EdgeR)*, sin embargo, se recomienda realizar estudios tomando en cuenta otros métodos de análisis existentes. Además, en cuanto a la simulación de datos *RNA-Seq* se recomienda establecer un mayor rango de barrido en relación a los diferentes parámetros de n, m y diffexp, para considerar otras condiciones de estudio. Es imprescindible remarcar, que cada estudio y análisis es diferente, y dicho trabajo, podría ayudar a futuros investigadores a realizar

protocolos en el ámbito de la expresión genética mediante la secuenciación masiva en paralelo, de los cuales se carece o no se tiene conocimiento.

El presente proyecto abre la puerta a un protocolo bioinformático, del cual se evidencian pocos en la literatura. Si bien existen protocolos, como por ejemplo para el análisis de variantes en genomas o exomas, estos se reducen al flujo bioinformático, más no al análisis estadístico. Sin embargo, en este estudio se propone un protocolo estadístico, para futuros investigadores, a los cuales les podrá servir de mucha ayuda, el hecho de discernir con un código, cuál será el mejor paquete para el estudio de futuros datos.

BIBLIOGRAFÍA

Ayala, S. (20 de Octubre de 2020). *RSG-Ecuador:un esfuerzo por contribuir con el desarrollo de la Bioinformática en el país y la región*. Obtenido de https://www.catalisisec.com/post/rsg-ecuador-un-esfuerzo-por-contribuir-con-el-desarrollo-de-la-bioinform%C3%A1tica-en-el-pa%C3%ADs-y-la-regi%C3%B3n

Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., . . . Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and Microarrays. *PloS one, 3*. doi:https://doi.org/10.1371/journal.pone.0017820

Castañeda, P. (2021). *Análisis de metodologías estadísticas en RNA-Seq, con aplicación a cáncer de pulmón.* Obtenido de https://repositorio.unal.edu.co/bitstream/handle/unal/81575/1016060566.2021.2022.pdf?sequence=3&isAllowed=y

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Rev Chil Infect*.

Chamorro, C. (04 de Junio de 2019). *Análisis de datos de RNA-Seq empleando diferentes paquetes desarrollados dentro del proyecto Bioconductor para estudios de expresión génica diferencial.* Obtenido de https://openaccess.uoc.edu/bitstream/10609/96466/6/cchamorropTFM0619memoria.pdf

Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., & Smyth, G. (2022). *EdgeR: differential analysis of sequence read count data User's Guide.* Obtenido de bioconductor.org: https://www.bioconductor.org/packages/devel/bioc/vignettes/*EdgeR*/inst/doc/*EdgeR*UsersGuide.pdf

Corchete, L. (2019). Expresión génica en mieloma múltiple: Análisis de datos RNA-Seq y Microarrays en combinación con estudios de metaanálisis y predicción de respuesta al tratamiento. *IBSAL*.

Costa-Silva, J., Domingues, D., & Lopes, F. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE*. Obtenido de https://doi.org/10.1371/journal.pone.0190152

Di, Y., Schafer, D. W., Cumbie, J. S., & Chang, J. H. (2011). The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology, 10(1). doi:10.2202/1544-6115.1637

Düntsch, I., & Gediga, G. (2020). Indices for rough set approximation and the application to confusion matrices. *International Journal of Approximate Reasoning*, 118, 155–172. https://doi.org/10.1016/j.ijar.2019.12.008

Gaur, P., & Chaturvedi, A. (2017). A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis. En D.-Q. Wei, Y. Ma, W. Cho, Q. Xu, & F. Zhou, *Translational Bioinformatics and Its Application.* Springer. Obtenido de https://bibliotecas.ups.edu.ec:2582/10.1007/978-94-024-1045-7_10

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

Huber, W., Carey, V., Gentleman, R., Anders, S., Carlson, M., Carvalho, B., . . . Irizarr, R. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *HHS Author Manuscripts* .

Kukurba, K., & Montgomery, S. (2015). RNA Sequencing and Analysis. *HHS Author Manuscrpits*. doi:10.1101/pdb.top084970

Law, C., Chen, Y., Shi, W., & Smyth, G. (2014). voom: precision weights unlock linear model analysis tools for RNA-Seq read counts. *Genome Biology*. doi:https://doi.org/10.1186/gb-2014-15-2-r29

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general-purpose read summarization program. *Bioinformatics*, 923–930.

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*.

Lindén, A., & Mäntyniemi, S. (2011b). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7), 1414–1421. https://doi.org/10.1890/10-1831.1

Liu, R., Heo, I., Liu, H., Shi, D., & Jiang, Z. (2023). Applying Negative Binomial Distribution in Diagnostic Classification Models for Analyzing Count Data. *Applied psychological measurement*, 47(1), 64–75. https://doi.org/10.1177/01466216221124604

Love, M., Anders, S., & Huber, W. (03 de 09 de 2023). *Analyzing RNA-Seq data with DESeq2*. Obtenido de http://bioconductor.org/packages/devel/bioc/vignettes/*DESeq2*/inst/doc/*DESeq2*.html#session-info

Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with *DESeq2*. *Genome Biol.* doi: 10.1186/s13059-014-0550-8.

Martínez, J., & Pérez, P. (2022). La curva ROC. *Medicina de Familia. SEMERGEN*. Obtenido de https://www.sciencedirect.com/science/article/pii/S1138359322001952

McCarthy, D. J., and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25, 765-771. http://bioinformatics.oxfordjournals.org/content/25/6/765

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008): "Mapping and quantifying mammalian transcriptomes by RNA-Seq," Nat Methods, 5, 621–628.

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean journal of anesthesiology*, 25-36. doi:10.4097/kja.21209.

Narkhede, S. (2021). Understanding Confusion Matrix - towards Data science. *Medium*. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Pan, Y., Landis, J. T., Moorad, R., Wu, D., Marron, J. S., & Dittmer, D. P. (2023). The Poisson distribution model fits UMI-based single-cell RNA-Sequencing data. *BMC Bioinformatics*, 24(1). https://doi.org/10.1186/s12859-023-05349-2

Phipson, B, Lee, S, Majewski, IJ, Alexander, WS, and Smyth, GK (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics* 10, 946-963. http://projecteuclid.org/euclid.aoas/1469199900

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). *Limma* powers differential expression analyses for RNA-Sequencing and microarray studies. Nucleic acids research, 43(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). *EdgeR*: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 139–140. doi:https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol* 11(3).

Robinson, M. D. and G. K. Smyth (2007): "Moderated statistical tests for assessing differences in tag abundance," Bioinformatics, 23, 2881–2887.

Robinson, M. D. and G. K. Smyth (2008): "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," Biostatistics, 9, 321–332.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural Processes*, 148, 56–62. https://doi.org/10.1016/j.beproc.2018.01.004

Saeed, U., & Usman, Z. (2019). Chapter 4: Biological Sequence Analysis. En H. Husi. Obtenido de https://www.ncbi.nlm.nih.gov/books/NBK550342/

Seyednasrollah, F., Laiho, A., & Elo, L. (2015). Comparison of software packages for detecting differential expression in RNA-Seq studies. *Brief Bioinform.* doi:10.1093/bib/bbt086

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, Article 3. http://www.statsci.org/smyth/pubs/ebayes.pdf

Tao, P., L, L. S., Sun, Y., Wang, Y., Yang, Y., Yang, B., & Li, F. (2022). ISG15 is associated with cervical cancer development. *Oncology letters*. Obtenido de https://doi.org/10.3892/ol.2022.13500

Tong, Y. (2021). The comparison of *Limma* and *DESeq2* in gene analysis. *E3S Web Conf*. Obtenido de https://doi.org/10.1051/e3sconf/202127103058