MÉTODOS ALTERNATIVOS PARA EVALUAR EXPRESIÓN DIFERENCIAL SIN RÉPLICAS DE LOS TRATAMIENTOS DE MATERIALES *Rubus glaucus* Benth TOLERANTES AL ATAQUE DE *Colletotrichum gloesporiodes* CON EL FIN DE IDENTIFICAR GENES DE IMPORTANCIA ASOCIADOS A TOLERANCIA

JULIANA ARIAS VILLEGAS

JULIÁN ANDRES MUÑOZ RAMIREZ

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERIA INDUSTRIAL

MAESTRIA INVESTIGACIÓN OPERATIVA Y ESTADISTICA

PEREIRA

JULIO 2019

MÉTODOS ALTERNATIVOS PARA EVALUAR EXPRESIÓN DIFERENCIAL SIN RÉPLICAS DE LOS TRATAMIENTOS DE MATERIALES *Rubus glaucus* Benth TOLERANTES AL ATAQUE DE *Colletotrichum gloesporiodes* CON EL FIN DE IDENTIFICAR GENES DE IMPORTANCIA ASOCIADOS A TOLERANCIA

JULIANA ARIAS VILLEGAS

JULIÁN ANDRES MUÑOZ RAMIREZ

DIRECTOR
JUAN CARLOS RINCÓN FLÓREZ Ph.D

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERIA INDUSTRIAL
MAESTRIA INVESTIGACIÓN OPERATIVA Y ESTADISTICA
PEREIRA
JULIO 2019

AGRADECIMIENTOS

A mi familia por su apoyo y amor incondicional, a la Dra. Marta Marulanda, el Dr. Andrés Duque y al grupo de investigación Biodiversidad y Biotecnología por ser mentores, guías e inspiración en mi formación, al Dr. Juan Carlos Rincón por su valiosa asesoría y paciencia, y a mi compañero de tesis Julián Andrés Muñoz por su disciplina y compromiso con el trabajo.

CONTENIDO

1.0 INTRODUCCIÓN

El cultivo de mora de la zona andina, conocido científicamente como *Rubus glaucus* se cultiva masivamente en Colombia donde se pueden encontrar diferentes variedades, los cuales pertenecen a la familia *Rosacea* que incluye otros miembros de gran importancia económica a nivel mundial como la fresa, la pera, la cereza, el durazno, la frambuesa y la rosa, entre otras.

Muchas familias de la zona andina dependen del cultivo de mora (*Rubus glaucus*, BENTH), sin embargo, la productividad de este cultivo no es la ideal debido a las enfermedades que lo afectan, entre las que se encuentra la antracnosis causada por *Colletotrichum gloeosporioides* como una de las de mayor relevancia (López-Vásquez et al., 2013). En la región cafetera de Colombia se ha identificado que en el 52,9% de los casos esta enfermedad afecta la productividad de los cultivos de mora, catalogándola como una de las más importantes (Botero et al., 2002).

Este trabajo se enfoca en identificar los genes en *Rubus glaucus*, BENTH tolerantes al ataque de *Colletotrichum gloeosporioides,* a partir de un análisis de expresión diferencial. Se usaron 3 grupos, un material tolerante inoculado con el patógeno, un material susceptible inoculado con el patógeno y un material susceptible sin inocular. El análisis de expresión diferencial viene después de realizar la secuenciación por RNA-seq y el ensamblaje del transcriptoma de *Rubus glaucus*, BENTH, los cuales fueron desarrollados por el grupo de investigación en Biodiversidad y Biotecnológica de la Universidad Tecnológica de Pereira. Por los altos costos que representan los análisis de secuenciación RNA-seq, sólo se contó con una réplica de este experimento.

El presente proyecto parte con un estudio del estado del arte de los experimentos RNA-seq para identificar métodos que permitan hacer análisis de expresión diferencial cuando sólo se cuenta con una réplica, después de ser identificados, se analizó un experimento de RNA-seq de *Saccharomyces Cerevisiae* con 48

réplicas disponibles (Schurch et al., 2016) del que ya se conocían los genes diferencialmente expresados, se escogió una réplica al azar para correr con los métodos identificados y finalmente validar los resultados obtenidos con los originales (este experimento se realizó 4 veces con réplicas diferentes). El método con el mejor rendimiento (mayor cantidad de genes expresados diferencialmente identificados y menos genes falsos positivos reportados) fue el seleccionado para el análisis de expresión diferencial de *Rubus glaucus*, BENTH, con el fin de tener un acercamiento a los posibles genes diferencialmente expresados que puedan explicar la tolerancia por su correspondencia fisiológica y función.

2.0 PLANTEAMIENTO DEL PROBLEMA

Los análisis de expresión diferencial se basan en experimentos de análisis RNA-seq con al menos tres réplicas por cada uno de los tratamientos; sin embargo, los altos costos asociados a los análisis RNA-seq limitan la posibilidad de acceder a la cantidad de réplicas deseadas estadísticamente. Si bien existen reportes en la literatura de métodos que pueden emplearse cuando se cuenta con un número de réplicas inferior a tres, no hay referencias de cómo impacta el menor número de réplicas la capacidad de predicción de los métodos.

En este contexto, es importante identificar cuáles son los métodos que tienen un mejor desempeño en condiciones de réplicas reducidas. Teniendo en cuenta lo anterior, contar con una réplica de cada uno de los tratamientos, no debería restringir el potencial uso de los resultados con el propósito de identificar genes candidatos en materiales *Rubus glaucus*, Benth que posteriormente pueden ser validados a través de la técnica de PCR en tiempo real, dado que esta réplica por tratamiento puede contener información importante que permita seleccionar genes candidatos a tolerancia a hongos patógenos que pueden ser empleados en futuros programas de mejoramiento genético de la especie.

3.0 OBJETIVOS

3.1 OBJETIVO GENERAL

Evaluar la expresión diferencial a partir de RNA-seq en materiales *Rubus glaucus Benth* susceptibles y tolerantes al ataque de *Colletotrichum gloespoiorides* utilizando diferentes métodos estadísticos que permitan abordar la ausencia de réplicas bilógicas con el propósito de identificar genes de interés.

3.2 OBJETIVOS ESPECÍFICOS

- Realizar una revisión del estado del arte del análisis de expresión diferencial en RNA-seq y métodos estadísticos que permitan evaluar expresión diferencial cuando no se cuenta con réplicas biológicas.

- Comparar diferentes métodos estadísticos para el análisis de expresión diferencial cuando no se cuenta con réplicas, usando datos de un experimento altamente replicado.

- Determinar el método estadístico que permita evaluar con mayor representatividad la expresión diferencial en tratamientos sin réplicas de *Rubus glaucus* Benth susceptibles y tolerantes al ataque de *Colletotrichum gloespoiorides*.

- Identificar y describir los genes expresados diferencialmente con posible asociación a la tolerancia de *Rubus glaucus* Benth al ataque de *Colletotrichum gloesporioides.*

4.0 JUSTIFICACIÓN

Estudios previos han identificado y caracterizado algunas especies de *Colletotrichum* como los responsables de la enfermedad en los cultivos de *Rubus glaucus*, Benth, en la zona cafetera de Colombia, dónde se identificó que en el 81% de los casos fueron causados por *gloeosporioides* (Ramírez et al., 2007). De la misma manera se ha realizado la caracterización molecular de variedades de *Rubus glaucus* (Marulanda et al., 2012). Sin embargo, el conocimiento del genoma y transcriptoma de esta planta es limitado y la mayoría se centran en otros miembros de la familia Rosácea, lo que presenta un reto para desarrollar trabajos que permitan entender mejor la genética de *Rubus glaucus*, Benth, y de esta manera poder identificar los genes asociados a la tolerancia de enfermedades que afectan su productividad y calidad.

Sin embargo, es complejo establecer modelos *in vitro* de inoculación de plantas y los costos de los análisis reducen la posibilidad de contar con un número significativo de réplicas; pese a esto, la información que se deriva de las réplicas reducidas, puede evidenciar información útil para identificar genes candidatos a tolerancia a hongos patógenos que pueden ser empleados en futuros programas de mejoramiento genético de la especie, contribuyendo así a mejorar el acceso a mercados donde está restringido el uso de agroquímicos.

Lo anterior implica, identificar métodos de análisis de expresión diferencial que tengan la capacidad de predecir genes diferencialmente expresados a partir de pocas réplicas, mediante de la evaluación del impacto del menor número de réplicas en la capacidad de predicción de los métodos.

Este trabajo está orientado a contribuir a la identificación de genes de *Rubus glaucus*, Benth asociados a la tolerancia de la enfermedad *Colletotrichum gloeosporioides* a partir del análisis de expresión diferencial de una réplica por cada tratamiento con el método que a partir de análisis previos demuestre la mejor

capacidad de predicción, brindando así una alternativa a aquellos investigadores que también tengan un número reducido de réplicas para sus análisis de expresión diferencial.

5.0 ESTADO DEL ARTE

La célula es la unidad básica de la vida. Hay dos tipos de células, las eucariotas y las procariotas. Una célula eucariota tiene tres principales componentes: membrana celular, citoplasma y núcleo. La membrana celular es la encargada de preservar la integridad de la célula y regular el flujo de nutrientes y proteínas. El núcleo contiene los cromosomas, donde se encuentran los genes, que son las unidades hereditarias de los organismos vivos. Finalmente, en el citoplasma se encuentran diferentes moléculas como el ARN (ácido ribonucleico), proteínas, carbohidratos, entre otros. El ARN contiene información transcrita del ADN, el cual es sintetizado en proteínas por los ribosomas. El proceso de producción de una molécula de ARN desde al ADN se llama expresión genética y el número de moléculas de ARN sintetizadas de cierto gen basado en las condiciones de la célula, se conoce como el nivel de expresión de ese gen. El nivel de expresión de los genes varía en respuesta a condiciones ambientales externas (Tarazona et al., 2014).

Los experimentos de expresión genética diferencial tienen el objetivo de identificar los genes que se expresan bajo ciertas condiciones experimentales, constituyendo una herramienta útil para establecer la relación entre genes y patologías. Algunas técnicas para medir la expresión diferencial incluyen: microrrays, real time polymerase chain reaction (RT-PCR), y serial analisys of gene expresión (SAGE) y los métodos de secuenciación masiva entre los que encontramos los análisis RNA-Seq como uno de los más usados (Tarazona et al., 2014).

Los métodos de secuenciación masiva o Next Generation Sequencing (NGS) han creado grandes posibilidades para la caracterización de los genomas y han avanzado significativamente en el entendimiento de su organización. Hoy en día las tecnologías NGS pueden ser usadas para informar las diferencias individuales del genoma dentro de individuos de la misma especie, caracterizar el espectro de interacción de las proteínas de unión al ADN y crear perfiles genéricos de

modificaciones epigenéticas (Tarazona et al., 2011). En los últimos años, el uso de RNA-Seq ha resultado en un incremento de nueva información que ha diseccionado la isomorfa y la expresión alélica (Carninci et al., 2005). RNA-Seq, también se usa cada vez más para cuantificar la expresión genética (Marioni et al., 2008).

Los métodos de expresión diferencial también han evolucionado con las tecnologías NGS. Los métodos tradicionalmente utilizados para microarrays han allanado el camino a otros enfoques que tienen en cuenta la naturaleza discreta de la cuantificación de la expresión y utilizan diferentes distribuciones de probabilidad para modelar datos (Marioni et al., 2008). La mayoría de las metodologías propuestas hasta ahora se basan en suposiciones paramétricas y usan distribuciones Poisson o binomiales negativas (NB) para modelar recuentos de características, siguiendo el razonamiento del procedimiento de muestreo en la secuenciación del ARN (Tarazona et al., 2011). Sin embargo, la confirmación posterior de las suposiciones de distribución es importante, ya que es posible que no siempre sean ciertas. Además, por lo general, hay muy pocas réplicas disponibles, lo que dificulta la estimación de los parámetros del modelo. Igualmente, los enfoques paramétricos tienden a ser problemáticos para evaluar la expresión diferencial en las características de bajo conteo (Bullard et al., 2010).

Un factor subyacente que se relaciona con varios de los problemas mencionados en el análisis RNA-seq es la cantidad de secuencias generadas en un experimento dado. Cuanto más se secuencia el objetivo, más transcripciones se identifican y mayor es el valor del nivel de expresión (Tarazona et al., 2011). Aunque la mayoría de los métodos de análisis existentes abordan este problema al incluir un factor de corrección relacionado con el tamaño de la biblioteca (Bullard et al., 2010), las tasas de secuenciación más altas presumiblemente resultarán en una estimación más precisa del nivel de expresión y de forma concomitante, los métodos inferenciales disfrutarán de mayor poder para identificar genes expresados diferencialmente (Tarazona et al., 2011); como resultado la habilidad

para detectar este tipo de genes estará determinada por la profundidad de la secuenciación (SD). La profundidad de la secuenciación se refiere al uso de un mayor número de lecturas únicas de cada región de una secuencia. El conocimiento de la relación entre SD y expresión diferencial es necesario para el propósito de cualquier diseño de experimentos y para entender las características del análisis de los resultados (Tarazona et al., 2011).

Tarazona et al., (2011), analiza el efecto que SD tiene en el análisis estadístico de los datos RNA-seq. Se evalúa cómo este parámetro se relaciona con la identificación de genes expresados, ruido de secuenciación, longitud de transcripción y expresión diferencial. Para lograr ese objetivo se evalúan varios métodos para el cálculo de expresión diferencial: edgeR, DESeq, baySeq, Test de Fisher y proponen un nuevo método llamado NOISeq. El experimento consiste en probar tres conjuntos de datos humanos de RNA-Seq con diferentes SDs y también con datos simulados. Como resultado de este experimento se concluye que a una mayor profundidad de secuenciación, mayor detección de genes, pero también mayor ruido en los datos, lo cual hace más complejo el análisis de expresión diferencial. La efectividad de los diferentes métodos también se mide con respecto a la tasa de falsos positivos detectados en el análisis. Tanto edgeR, DESeq y baySeq muestran un mayor incremento en la tasa de falsos positivos a medida que se incrementa la profundidad de la secuenciación, por otra parte el Test de Fisher muestra un patrón constante pero este es debido a su bajo poder para identificar genes altamente expresados (Tarazona et al., 2011) y NOISeq muestra un mejor desempeño reportando una menor tasa de falsos positivos en comparación a la los demás métodos a medida que se incrementa la profundidad de la secuenciación.

Otro de los aspectos importantes a tener en cuenta dentro de los análisis de expresión diferencial además de la profundidad de la secuenciación, es la cantidad de réplicas bilógicas utilizadas. La cantidad de réplicas dependerá de la precisión que se quiera lograr con el experimento, pero también dependerá de los recursos

disponibles y de las características particulares de las muestras a analizar. Si la variabilidad del material experimental y de la variable respuesta es pequeña se requerirán menos replicas por tratamiento (Mendoza et al., 2002). Yuwen et al., (2014) muestra la compensación explícita entre más replicas biológicas y una secuenciación más profunda en el aumento de la potencia para detectar genes expresados de manera diferencial. Ese análisis permite concluir que secuenciar menos lecturas y realizar más replicación biológica es una estrategia efectiva para aumentar el poder y la precisión en estudios de RNA-seq de expresión diferencial a gran escala, puesto que una secuenciación más profunda para cada muestra genera rendimientos decrecientes para el poder de detención de genes de expresión diferencial, una vez más allá de una determinada profundidad de secuenciación.

(Schurch et., 2016), evalúa el rendimiento de las herramientas de análisis de expresión diferencial a través de un experimento RNA-seq altamente replicado diseñado específicamente para probar tanto las suposiciones intrínsecas a las herramientas RNA-seq como para evaluar su desempeño. El documento se centra en 11 herramientas de análisis de expresión diferencial específicas de RNA-seq populares: baySeq, cuffdiff, DEGSeq, DESeq, DESeq2, EBSeq, edgeR, limma, NOISeq, PoissonSeq y SAMSeq y evalúa su desempeño en función del número de réplicas. En este artículo se concluye que en un experimento de análisis de expresión diferencial al menos seis réplicas por condición son necesarias, y en caso de que el objetivo del experimento sea identificar la mayor cantidad de genes diferencialmente expresados, serán necesarias al menos 12 réplicas por cada condición. Entre las herramientas que presentan un mejor desempeño están edgeR o DESeq2 para experimentos de menos de 12 réplicas por condición, y para experimentos de más de 12 réplicas por condición, los autores recomiendan DESeq. Otra conclusión importante es que incluso las mejores herramientas tienen limitado poder estadístico con muy pocas réplicas por condición. Sin embargo, muchos análisis de RNA-Seq se soportan en un bajo número de réplicas

por condición ($n \lesssim 3$), y para compensar la falta de replicación, herramientas como: DESeq, DESeq2, edgeR y limma, modelan la relación de la varianza media y utilizan información de los genes para reducir la varianza del gen dado hacia el modelo común (De Hertogh et al., 2010).

DESeq2 y edgeR son las herramientas que se pueden emplear cuándo se cuenta con una sola réplica por tratamiento. Otra herramienta importante es NOISeq, la cual contiene un componente llamado NOISeq-sim, que permite trabajar con datos donde no se tienen réplicas, dado que este paquete tiene la capacidad de poder simularlas. Esta simulación se basa en la suposición de que los datos siguen una distribución multinomial, donde la probabilidad para cada gen en esa distribución, es la probabilidad de la lectura de mapear ese gen (Tarazona et al., 2011); pero se debe tener presente que NOISeq-sim no puede simular la variabilidad biológica, la cual es necesaria para análisis inferenciales en la población. Sin embargo muchos experimentos RNA-seq tienen muy bajo número de réplicas y aun así la habilidad de los métodos estadísticos para trabajar con replicas técnicas o sin réplicas en lo absoluto, es relevante (Tarazona et al., 2014).

6.0 MARCO TEÓRICO

Los datos que se obtienen al aplicar RNA-Seq para un análisis de expresión diferencial son datos de conteo, por lo que se hace necesario utilizar modelos estadísticos específicos para datos de conteo. Los modelos que se utilizarán son los Modelos Lineales Generalizados (MLGs), en los cuales se apoyan los software utilizados para el análisis de expresión diferencial con RNA-Seq que se desarrollan en este proyecto.

Los Modelos Lineales Generalizados son una extensión de los modelos lineales clásicos los cuales tratan de explicar la variable objetivo como una combinación lineal de una función de un conjunto de variables explicativas. Dos casos particulares de los MLG son el modelo de regresión de Poisson y el modelo de regresión Binomial Negativa. Los software utilizados para los análisis RNA-Seq utilizados en este proyecto, se basan en esos dos modelos.

6.1 MODELO DE REGRESION DE POISSON.

El modelo de regresión de Poisson, es un MLG en el cual la variable respuesta Y sigue una distribución de Poisson de media μ, con μ > 0, lo que se denota como $Y \sim Po$ (μ).

Esta distribución se suele utilizar para modelar la probabilidad de que durante un determinado periodo de tiempo ocurra un cierto número de eventos, por ejemplo, el número de personas que llaman a una centralita de teléfono en un periodo de tiempo determinado. También se suele utilizar para modelar datos de conteo en un cierto espacio, por ejemplo, el número de errores que comete una fotocopiadora en dos impresiones. Se podría decir que la distribución de Poisson corresponde a

datos de conteos en la misma forma que la distribución normal lo es para datos continuos. La función de probabilidad de esta distribución es:

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!};\ para\ y \in \mathbb{Z}_{o}^{+};\ \mu > 0$$

Y su media y varianza son:

$$E(Y) = Var(Y) = \mu$$

Como se muestra, la distribución de Poisson, es una distribución discreta que cumple la propiedad de equi dispersión, es decir, que la varianza y la media son iguales. Por esta propiedad, se tiene que mientras mayor sea el valor esperado, mayor dispersión tendrán los valores que puede tomar la variable que se distribuya así.

6.1.2 Modelo de Regresión de Poisson como un MGL.

Si tomamos logaritmo neperiano de esta distribución se puede comprobar que este modelo pertenece a la familia exponencial, y además podremos identificar los elementos necesarios de un MGL.

$$\ln P[Y = y] = -\mu + yln(\mu) - \ln(y!) = yln(\mu) - \mu - \ln(y!)$$

Los elementos serían:

- La función link, es:

$$\theta = \theta(\mu) = \ln(\mu)$$

- El parámetro de dispersión:

$$\varphi = 1$$

- Función cumulante:

$$b(\theta) = \mu$$

- Función c(y, φ), específica de cada distribución, como se había explicado en MLG:

$$c(y, \varphi) = \ln(y!)$$

Como se puede ver, es un modelo que pertenece al MLG, ahora, una vez que hemos identificado estos elementos, vamos a estudiar el modelo de regresión.

Para obtener la función que predice la media en este modelo vamos a seguir los siguientes pasos:

- Obtenemos μ a partir de la función link:

$$\mu = e^{\theta}$$

- Puesto que θ(μ) = ln(μ):

$$ln(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- La función que predice la media sería:

$$\mu = e^{v}; \quad donde\ v = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Interpretación de los coeficientes:

Veamos la interpretación de los coeficientes considerando sólo una variable independiente X, es decir se predice:

$$\theta^* = \beta_0 + \beta_1 X$$

Suponiendo que se han observado los valores de *X*, x_1 y x_2 para los individuos 1 y 2, respectivamente, se tendría:

$$\theta_1^* = \theta^*(X = x_1) = \beta_0 + \beta_1 x_1$$

$$\theta_2^* = \theta^*(X = x_2) = \beta_0 + \beta_1 x_2$$

La diferencia entre las predicciones sería:

$$\theta_2^* - \theta_1^* = \beta_1(x_2 - x_1)$$

Teniendo en cuenta la expresión de la función link en este modelo, θ = ln(μ), se obtiene que:

$$ln(\mu_2^*) - ln(\mu_1^*) = \beta_1(x_2 - x_1) \Leftrightarrow ln\left(\frac{\mu_2^*}{\mu_1^*}\right) = \beta_1(x_2 - x_1)$$

$$\Leftrightarrow \frac{\mu_2^*}{\mu_1^*} = e^{\beta_1(x_2 - x_1)}$$

En el caso particular en que el predictor aumente en una unidad, se tiene que:

$$\frac{\mu^*(x+1)}{\mu^*(x)} = e^{\beta_1} \Leftrightarrow \mu^*(x+1) = \mu^*(x)e^{\beta_1}$$

Se puede observar de esta manera que el aumento de una unidad en la variable *X* es múltiplo de la media anterior, siendo el factor de proporcionalidad e^{β_1} .

6.2 MODELO DE REGRESION BINOMIAL NEGATIVA.

Se puede obtener siguiendo dos métodos diferentes:

- Derivación en términos de una mixtura Poisson-Gamma.
- Derivación en términos de la expresión clásica de la función de probabilidad de la Binomial Negativa.

6.2.1 Binomial Negativa como mixtura Poisson-Gamma.

Suponemos que la variable respuesta Y_i sigue una distribución de Poisson con media μ_i, es decir:

$$Y_i \sim Po(\mu_i); \quad donde \ \mu_i = \lambda_i u_i$$

$ln\,\mu_i$ se modeliza como un modelo lineal:

$$\ln \mu_i = ln\lambda_i + ln\mu_i = \underline{X'_i\beta} + \epsilon_i$$

Donde μ_i es el efecto observado, error o ruido para el que se propone una distribución Gamma de media 1:

$$\mu_i \sim Ga(v, v); \quad por\ tanto \quad E[\mu_i] = \frac{v}{v} = 1; \ Var[\mu_i] = \frac{v}{v^2} = \frac{1}{v}$$

Con esta notación se tiene que:

$$Y_i\big|_{\lambda_i\ \mu_i} = Y_i\big|_{\underline{x}_i, \mu_i} \sim Po\ (\mu_i)\ \ donde \ \ \mu_i = \lambda_i \mu_i, \ \ \mu_i \ \sim \ \ Ga\ (v, v)$$

y la función de probabilidad de Y_i se puede probar que es:

$$fY_i(y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\,\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1 + \alpha\mu_i}\right)^{y_i}$$

$con\ y_i \in \mathbb{Z}_0^+, \ \alpha = 1/v$

6.2.2 Modelo de Regresión Binomial Negativa como un MGL.

De la expresión de la función de probabilidad del punto anterior se deduce que:

- La función vínculo es:

$$\theta = ln\left(\frac{\alpha\mu}{1 + \alpha\mu}\right)$$

- El parámetro de dispersión de la familia exponencial en forma canónica:

$$\varphi = 1$$

- La función cumulante:

$$b(\theta) = -\frac{1}{\alpha}\, ln\left(\frac{1}{1+\alpha\mu}\right)$$

Interpretación de los coeficientes.

Considerando el estudio de una sola variable independiente *X*, y suponiendo que se han observado los valores x_1 y x_2 de dicha variable tenemos que la razón que mide el cambio de un valor observado a otro es el mismo que para el modelo de regresión de Poisson. En el caso particular en que el predictor aumente en una unidad, al igual que en el modelo de Poisson, se tendría que:

$$\mu^*(x+1) = \mu^*(x)e^{\beta_1}$$

6.2.3 Modelo de Regresión Binomial Negativa cuando se reduce a la comparación de dos grupos independientes.

Cuando comparamos dos grupos independientes, es como si se considerara el modelo anterior con sólo una variable independiente, *X*, siendo ésta una variable indicadora, definida de la siguiente forma:

$$X = \begin{cases} 1, & si\ un\ individuo\ pertenece\ al\ grupo\ tratado \\ 0, & si\ uninvibiduo\ pertenece\ al\ grupo\ control \end{cases}$$

El estudiar la diferencia de estar en un grupo u otro, es equivalente a estudiar el cambio que sufre la variable respuesta al aumentar la variable independiente en una unidad, pues si $x_B = 0$, y estudiamos el cambio al aumentar una unidad esta variable, entonces $x_A = x_B + 1 = 1$, lo que en la comparación entre grupos significará el cambio de un grupo a otro. Si recordamos, en la sección 6.2.1.2,

obtuvimos como resultado que el logaritmo neperiano del cociente de medias puede expresarse como:

$$ln\left(\frac{\mu_A^*}{\mu_B^*}\right) = b_1(x_A - x_B) = b_1(1-0) = b_1$$

Por tanto, tenemos que b_1 proporciona la información sobre la consecuencia de pertenecer a un grupo u otro.

A la razón entre las medias, se le denomina generalmente *fold-change*. En la literatura es usual trabajar con el logaritmo en base 2 de este cociente. En este caso la relación entre las medias debe interpretarse como sigue:

$$Logaritmo\ en\ base\ dos: log_2\left(\frac{\mu_A^*}{\mu_B^*}\right) = b1 \implies \mu_A^* = \mu_B^*\, 2^{b_1}$$

Obsérvese que esta relación es análoga a la que tendríamos utilizando un modelo de Regresión Binomial Negativa:

$$Logaritmo\ neperiano: ln\left(\frac{\mu_A^*}{\mu_B^*}\right) = b1 \implies \mu_A^* = \mu_B^*\, e^{b_1}$$

Aun así, hoy en día, se sigue utilizando log2fold−change para estudiar el efecto de pertenecer a un grupo u otro, pero las consideraciones anteriores deben ser tenidas en cuenta.

6.2.4 Comparación de la expresión diferencial de genes en dos grupos independientes.

Cuando se estudia la expresión diferencial de genes en dos grupos independientes A y B, el contraste se reduce a una comparación de medias, por lo que las hipótesis planteadas para un gen g tendrá la forma:

$$H_0^g: \mu_A^g = \mu_B^g$$

$$H_1^g: \mu_A^g \neq \mu_B^g$$

Donde μ_A^g y μ_B^g corresponden a las medias de los conteos del gen g en las condiciones A y B, respectivamente. El contraste anterior se suele formular en términos de *fold-change* definido en la sección anterior como:

$$fold - change = \frac{\mu_A}{\mu_B}$$

Pudiendo tomar este parámetro los siguientes valores:

$$fold - change = \begin{cases} > 1 \; si \; \mu_A > \mu_B \\ = 1 \; si \; \mu_A = \mu_B \\ < 1 \; si \; \mu_A < \mu_B \end{cases}$$

si $\mu_B = 0$ el fold change es infinito.

Si $\mu_A = \mu_B = 0$ tendremos una indeterminación.

Debido a la magnitud y gran variabilidad de los datos de conteo que se tienen en los estudios de RNA-Seq, no se trabaja directamente con el *fold-change*, sino con el logaritmo en base dos de este parámetro:

$$log_2 fold - change = log_2 \left(\frac{\mu_A}{\mu_B}\right) = log_2 \mu_A - log_2 \mu_B$$

$$log_2 fold - change = \begin{cases} > 0 \; si \; \mu_A > \mu_B \\ = 0 \; si \; \mu_A = \mu_B \\ < 0 \; si \; \mu_A < \mu_B \end{cases}$$

Por tanto, el contraste planteado inicialmente para la comparación de los conteos medios del gen g en los grupos A y B sería:

$$H_0^g: \mu_A^g = \mu_B^g$$

$$H_1^g: \mu_A^g \neq \mu_B^g$$

Es equivalente al contraste

$$H_0^g: log_2\left(\frac{\mu_A^g}{\mu_B^g}\right) = 0$$

$$H_1^g: log_2\left(\frac{\mu_A^g}{\mu_B^g}\right) \neq 0$$

Y este último a su vez a

$$H_0^g:\ log_2(\mu_A^g) - log_2(\mu_B^g) = 0$$

$$H_1^g:\ log_2(\mu_A^g) - log_2(\mu_B^g) \neq 0$$

Existen diversos estadísticos que se pueden utilizar para proceder con este contraste. DEseq2, uno de los paquetes analizados en este proyecto, utiliza el *estadístico de Wald*:

$$t_g = \frac{log_2(\hat{\mu}_A) - log_2(\hat{\mu}_B)}{\sqrt{\widehat{Var}[log_2(\hat{\mu}_A) - log_2(\hat{\mu}_B)]}}$$

Cuya distribución asintótica es *N(0,1)*.

Para un contraste de la forma:

$$H_o:\ \theta = \theta_0$$

$$H_1:\ \theta \neq \theta_0$$

El estadístico de Wald se define como:

$$T = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\widehat{Var}\left(\hat{\theta}\right)} \simeq X_1^2$$

ó equivalente como:

$$T = \frac{(\hat{\theta} - \theta_0)}{\sqrt{\widehat{Var}(\hat{\theta})}} \simeq N(0,1)$$

6.3 TIPO DE ERRORES A CONTROLAR.

Son dos los posibles errores que se pueden cometer al realizar un test de hipótesis, el primero sucede cuando el test resulta significativo, por lo que se rechaza la hipótesis nula siendo cierta, al que se le denomina como error de tipo uno y el segundo sucede cuando el test nos lleva a no rechazar la hipótesis nula y esta es falsa, denominado como error de tipo dos.

Tabla 1. Errores en contrastes de hipótesis.

Hipótesis	No significativo	Significativo
H_0 Cierta	Decisión correcta	Error tipo I
H_1 Cierta	Error tipo II	Decisión correcta

Las pruebas estadísticas por lo general se basan en el control de la tasas de error tipo I, para ello al realizar un test se fija un nivel se significancia, α, normalmente con un valor de 0.05, que mide la probabilidad de rechazar incorrectamente la hipótesis nula, es decir, la significancia es la probabilidad de obtener un error tipo I. Usualmente el p-valor, definido como la probabilidad de observar un resultado igual o más extremo que el estadístico del test, se compara con α para tomar la decisión de rechazar o no la hipótesis nula.

Si se considerara ahora el problema de contrastar m hipótesis nulas dadas por H_o^1, ..., H_o^m, en cada una de las m pruebas podrían ocurrir tanto errores tipo I como errores tipo II. Sumando sobre todas las pruebas se obtienen los siguientes resultados:

Tabla 2. Errores en contrastes de hipótesis múltiples.

Hipótesis	No significativo	Significativo	
H_0 Cierta	U	V	m_0
H_1 Cierta	T	S	$m - m_0$
	m - R	R	m

U y S corresponden al número de veces que se han tomado decisiones correctas en los m contrastes, V es el número de errores tipo I cometidos y T el número de errores tipo II.

La gran parte de los métodos de ajuste para comparaciones múltiple recurren a las tasas **FWER** (family-wise error rate) y **FDR** (false discovery rate), las cuales, al igual que en los tests de hipótesis simples, tratan de controlar la tasa de error tipo I.

6.3.1 FWER (Family-wise error rate)

Esta tasa de error se define como la probabilidad de cometer al menos un error tipo I, por lo que tenemos que:

$$FWER = Pr\,(V \geq 1)$$

Método Bonferroni para controlar el FWER.

El método más conocido para controlar el FWER es el de Bonferroni. Por este método, si se quiere tener un nivel global máximo de α, se debe tomar para el contraste de cada gen un nivel corregido $\alpha* = \alpha/m$, siendo m el número total de contrastes o equivalentemente el número de genes existentes. Esta propuesta puede ser demasiado conservadora y cuando la cantidad de contrastes es grande

los niveles corregidos resultan demasiado bajos. Esto significa que se seleccionarán pocos genes como candidatos a estar DE.

6.3.2 FDR (False discovery rate).

FDR se define como la proporción esperada de hipótesis nulas que son verdaderas entre las que son declaradas como significativas.

Volviendo a la notación definida en la tabla 2, asumiendo que R, número de pruebas que han resultado significativas, es mayor que 0, FDR se define como:

$$FDR = E\left(\frac{V}{R}\right) \quad donde\ E(\cdot)\ es\ la\ esperanza\ matemática$$

Para *R = 0*, se define la razón $\frac{V}{R}$ como 0 ya que si no existen pruebas significativas R no pueden ocurrir falsos rechazos. Por tanto FDR puede expresarse como:

$$FDR = E\left(\frac{\frac{V}{R}}{R > 0}\right) Pr(R > o) + E\left(\frac{\frac{V}{R}}{R = 0}\right) Pr(R = 0)$$

$$= E\left(\frac{\frac{V}{R}}{R > 0}\right) P\,r(R > 0)$$

Bajo el supuesto de que todas las hipótesis nulas son ciertas, en la tabla 2 tendríamos que *V = R* y *V/R* será:

$$\frac{V}{R} = \begin{cases} 0 & si\ V = 0 \\ 1 & si\ V \geq 1 \end{cases}$$

Entonces:

$$FDR = E\left(\frac{V}{R}\right) = 0 * Pr(V = 0) + 1 * Pr(V \geq 1)$$

$$= Pr(V \geq 1) = FWER$$

Esto significa, que si todas las hipótesis nulas son ciertas, entonces la FDR es igual a FWER.

Si se supiera ahora que no todas las hipótesis nulas con ciertas, entonces se tendría que V < R, por lo que $\frac{V}{R} < 1$ y:

$$FDR = E\left(\frac{V}{R}\right) = \left(\frac{V}{R}\right) * Pr(V \geq 1) + \left(\frac{0}{R}\right) * Pr(V = 0)$$

$$= \left(\frac{V}{R}\right) * Pr(V \geq 1) < Pr(V \geq 1) = FWER$$

Es decir: *FDR < FWER*

Como consecuencia, se tiene el resultado general de que FDR es menor o igual que FWER, lo que implica que cualquier enfoque que controle FWER también controlará FDR. El contrario, sin embargo, no es cierto.

La elección de la medida de error a usar se basa en gran medida en el objetivo científico y las expectativas de la investigación. Cuando existen un gran número de variables para analizar, cada vez es más frecuente controlar el FDR, esto se debe, en gran parte, a que se espera que muchas de las correspondientes hipótesis sean falsas. Si el número de hipótesis nulas verdaderamente falsas es pequeño o la consecuencia de equivocarse por la prueba significativa es grave, entonces la FWER puede ser una medida más apropiada.

6.3.3 Método de Benjamini y Hochberg (B-H) para controlar el FDR.

El FDR es una medida atractiva de la tasa de error en el contexto de las investigaciones genéticas. Uno de los enfoques más utilizados para controlar el FDR, es el ajuste de Benjamini y Hochberg (B-H), propuesto por Benjamini y Hochberg (1995) (Y. Benjamini et al., 1995), basado en el supuesto de independencia de los p-valores. A diferencia del método de Bonferroni este método considera un nivel de significación diferente para cada contraste.

Vamos a considerar contrastar la serie de hipótesis nulas independientes dadas por H_o^1, ..., H_0^m, donde se han obtenido los p-valores p_1, ..., p_m y además, supongamos que queremos controlar el FDR a un nivel q. Los pasos que sigue el método B-H son los siguientes:

- Sean $p_{(1)}$,…, $p_{(m)}$ los p-valores observados ordenados de tal manera que:

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

Y sean las correspondientes hipótesis nulas dadas por $H_0^{(1)}, \ldots, H_0^{(m)}$

- se define k como:

$$k = max\left\{i: p_{(i)} \leq \frac{1}{m} q\right\}$$

- Rechazamos las hipótesis nulas $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(k)}$

6.4 HERRAMIENTAS PARA EL ANALISIS RNA-Seq.

RNA-Seq es una metodología que permite evaluar el perfil de expresión del ácido ribonucleico (ARN) y es utilizada entre muchos otros usos para evaluar la importancia de un gen o conjunto de genes en ciertas enfermedades, la eficacia de un fármaco, o los grados de patogenicidad de ciertas cepas de bacterias.

6.4.1 EdgeR

Es un paquete escrito para R que realiza expresión diferencial genética utilizando datos de conteos bajo un modelo binomial negativo (Yunshun et al., 2008). A continuación se explicarán las funciones más importantes del paquete y el procedimiento para realizar el análisis de expresión diferencial con el mismo.

Clase DGEList.

Cada paquete tiene sus propias clases o condiciones para poder trabajar con él, este paquete necesita la clase *DGEList* para poder realizar cualquier análisis. Para crearla se utiliza la propia función *DGEList*:

DGEList(counts = matrix(0, 0, 0), lib.size = colSums(counts), norm.factors = rep(1,ncol(counts)), group = rep(1,ncol(counts)), remove.zeros = FALSE)

En el atributo *count* irá la matriz con los conteos brutos, donde se debe de tener por filas los genes y por columnas las muestras o individuos, en *group* la condición de cada muestra bajo estudio, por ejemplo la condición tratados o no tratados que corresponda a cada muestra, y por último en *gene* se puede incluir algún vector o data.frame que indique alguna información de los genes, no es necesario incluir este atributo, de hecho, por defecto será *genes = NULL*. A través del atributo *remove.zeros* se pueden eliminar los genes que tienen siempre conteos nulos, es decir, que en todas las columnas, la fila que corresponde a ese gen tiene valores ceros.

Filtrado de los genes.

Es muy importante, antes de comenzar con el análisis diferencial de los genes, eliminar aquellos que no tienen conteos, o que apenas tienen. Pues de antemano se sabe que estos genes no se van a expresar de forma distinta en las diferentes condiciones que se estén estudiando, puesto que al no tener conteos para ninguna condición su estudio será nulo. Por ejemplo, si tuviéramos un gen que no

se ha expresado, es decir, que ni en una condición ni en otra ha tenido conteos, al realizarle el estudio para saber si es diferencialmente expresado en las debidas condiciones, como resultado tendremos que no lo es, pues en todas las condiciones ha tenido los mismos conteos, es decir, cero.

En síntesis se trata de una depuración de datos como se haría en cualquier estudio estadístico, donde se analiza que hacer con aquellos datos que se identifican como outlier o que representan algún tipo de problema al estudio.

Para eliminar estos genes basta con identificar aquellos que no tienen conteos y eliminarlos del objeto *DGEList* creado y ya se podrá seguir con el estudio.

Normalización

Este paquete, por defecto, trabaja la normalización de los conteos brutos a través del método TMM (Trimmed Mean of M-values), es decir, media truncada de M – valores [28].

Sean Y_{gk} y μ_{gk} el número de conteos observado y el nivel de expresión para el gen g en la muestra k, respectivamente. El nivel de expresión será desconocido y vendrá dado en términos del número de transcripciones. Sean L_g y N_k la longitud del gen g, es decir, el número de bases que éste adquiere y el número total de lecturas en la muestra k.

El valor esperado de Y_{gk} es:

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

Siendo

$$S_k = \sum_{g=1}^{G} \mu_{gk} L_g$$

Donde S_k representa la salida de RNA total de la muestra. Este método, aunque es bastante recomendable, tiene la dificultad de que el valor S_k es desconocido y se estima a partir de la muestra, por lo que puede variar drásticamente de una muestra a otra.

En edgeR existe una función que calcula unos factores de normalización para escalar el tamaño bruto de las librerías o muestras:

calcNormFactors(object, method=c("TMM","RLE","upperquartile","none"), ...)

En este caso solamente se indicará el atributo *object*, donde irá el objeto de clase *DGEList* para normalizarlo, los demás atributos se pueden dejar con los valores asignados por defecto.

Estimación de la dispersión.

Este paquete calcula la dispersión común o variabilidad total para todos los genes mediante el método qCML. Para ello emplea la función estimateCommonDisp(·):

estimateCommonDisp(object, verbose=FALSE,...)

En *object* se dará el objeto creado anteriormente con la clase *DGEList*, y el atributo *verbose* sirve para que muestre, tras aplicar la función, la dispersión común estimada y el coeficiente de variación biológica, que es la raíz cuadrada de la dispersión.

Sin embargo para genes con bajo nivel de expresión la dispersión común estimada supone una mayor variabilidad de la que realmente presentan, por ello es recomendable estimar la dispersión gen a gen o dispersión tagwise, donde a partir de la dispersión común y de la variabilidad real de cada gen se le asignará a cada uno de ellos una dispersión más propia de este, por lo que para calcular estas estimaciones previamente se deberá haber estimado la dispersión común. Estas estimaciones se obtendrán con la función estimateTagwiseDisp( ):

estimateTagwiseDisp(object, prior.df=10, method="grid", ...)

De nuevo en *object* citaremos el objeto al que se le está realizando el estudio, y los demás atributos se dejarán por defecto.

Pruebas

En este paso ya se puede proceder a identificar los genes diferencialmente expresados. Los datos están bajo la clase *DGEList*, están normalizados, por lo que se pueden comparar entre sí y se han ajustado a una distribución teórica, la distribución Binomial Negativa, para la cual ya se han estimados los parámetros.

El paquete edgeR incluye la función *exactTest*, el test exacto basado en la distribución Binomial Negativa (test exacto binomial negativo) [29], que realiza las pruebas por parejas para la expresión diferencial entre dos grupos.

exactTest(object, pair=1:2, dispersion = "auto", big.count=900, prior.count=0.125)

Solo se nombrará el atributo *object*, donde se volverá a poner nuestro objeto, de clase *DGEList*, ya normalizado y con las estimaciones calculadas.

Con esta función se obtendrán tres elementos:

- Bajo el dominio...*$table* habrá un data.frame donde se pueden observar por columnas el log2fold − change (logFC), los logaritmos en base 2 de la media de los cpm (conteos por millón) de cada gen (logCPM) y el p-valor bilateral (PValue).

- Bajo...*$comparison* se puede ver un vector de caracteres con las

condiciones de los dos grupos que se comparan.

- Bajo...*$genes* un data.frame donde aparecerá la información que se le haya dado, en caso de haberlo hecho, de los genes al crear el objeto de la clase *DGEList*.

Sin embargo, tras calcular las pruebas exactas, se aplicará a estos resultados la función *topTags(·)*, pues con esta función también se tendrán varios elementos, entre ellos una tabla, igual que la obtenida con la función anterior, sin embargo tendrá una columna en la que encontraremos el p-valor ajustado según el método que elijamos, que en nuestro caso será el método de B-H, el cual EdgeR asigna por defecto, explicado anteriormente, y en caso de que al crear el objeto *DGEList* se haya incluido alguna información de los genes, también nos aparecerá una columna con dicha información.

topTags(object, n=10, adjust.method="BH", sort.by="PValue")

Por otro lado, aunque internamente calcule estos valores para todos los genes solo aparecerán por defecto, los datos referidos a los primeros 10 genes con menor p-valor, si se desea que muestre más genes se tendrá que cambiar este valor por defecto, poniendo el número en el atributo *n* de la función.

6.4.2 DESeq2

Al igual que edgeR, DESeq2 permite realizar el análisis de expresión diferencial, basándose, en la distribución Binomial Negativa. A continuación se explicarán las funciones más importantes del paquete y el procedimiento para realizar el análisis de expresión diferencial con el mismo.

Clase DESeqDataSet.

DESeq2 trabaja bajo la clase DESeqDataSet. Para crear un objeto de esta clase está la función DESeqDataSetFromMatrix (·), la cual necesitará la matriz de conteos.

DESeqDataSetFromMatrix(countData, colData, design, ignoreRank = FALSE, ...)

En *countData* irá la matriz de conteos, en *colData* se incluirá la información acerca de las muestras o los individuos, en el que al menos existirá una columna, y cada fila corresponderá a cada columna de la matriz nombrada en *countData*. Dentro de esta información deberá estar la condición que separa a las muestras para el análisis que realizaremos y además se podrá tener más información sobre éstas.

Filtrado de los genes.

Es conveniente realizar un filtrado previo de los genes, pues existen genes que no tienen conteos en ninguna de las muestras, o que apenas se expresan y lo único que hace es dificultar las interpretaciones.

Esta información debe eliminarse manualmente, pues no existe una función específica que lo realice.

Normalización.

Con este paso se busca minimizar el ruido técnico introducido en los datos durante el proceso de secuenciación con el fin de volverlos comparables entre sí, no se pretende transformarlos para que sigan una distribución normal, que es lo que se suele entender como *normalización* en términos estadísticos.

Lo que hace este paquete es dividir los conteos que tenemos para cada gen en

una muestra por el número total de lecturas en dicha muestra. Este proceso es realizado internamente por el método.

Estimación de la dispersión.

La estimación de la dispersión de cada gen se basa en 3 pasos y es calculada internamente por DESeq2. En primer lugar se calcula la media de los conteos de cada gen para estimar la dispersión gen a gen, luego se ajusta una curva a través de esas estimaciones, por último se asigna a cada gen un valor de dispersión, por defecto DESeq2 elige el mayor valor entre la estimación por gen y el valor ajustado. Cabe decir que este paquete toma una postura muy conservadora y por eso mismo asigna el mayor valor posible a cada gen. La dispersión se puede entender como el cuadrado del coeficiente de variación biológica. Cuanta más dispersión o variación biológica tenga un gen, mayor deberá de ser la diferencia entre los conteos de las distintas condiciones tratadas para que ese gen se tome como significativo en la prueba de expresión diferencial.

Pruebas

Las pruebas se realizan bajo test Binomial Negativo, pero antes de realizar el análisis deberemos modificar el orden de los niveles del factor con la condición que agrupa las muestras o individuos. DESq2 utiliza la siguiente función para ese propósito:

relevel(x, ref, ...)

En *x* se incluirá el factor a tratar y en *ref* citaremos el nivel de referencia, que será el que R use como denominador del estadístico log2fold − change. Una vez hecho el cambio de los niveles, el análisis se realizará mediante la función DESeq(·), la cual engloba los pasos de normalización, cálculo de dispersión y la prueba del test binomial negativo, que por defecto es el estadístico de Wald.

DESeq(object, test = c(“Wald”, “LRT”), fitType = c(“parametric”, “local”, “mean”), quiet = FALSE, ...)

En *object*, irá el objeto de clase *DESeqDataSet*. El atributo *quiet*, por defecto, hace que mientras trabaje la función nos vaya indicando los pasos que va siguiendo hasta terminar el análisis.

Desde este nuevo objeto creado ya se pueden observar los estadísticos, y el p-valor que nos indica si aceptar o rechazar que los genes estén diferencialmente expresados, sin embargo no se han calculado los p-valores ajustados por el método B-H por lo que se ejecutará también la función results(·):

results(object, contrast, lfcThreshold = 0, altHypothesis = c(“greaterAbs”, “lessAbs”, “greater”, “less”), pAdjustMethod =“BH”, ...)

Únicamente se nombrará el objeto bajo estudio y los demás atributos se dejarán por defecto. R utiliza para calcular los p-valores ajustados el método B-H por defecto, sin embargo, si se quisiera utilizar otro método habría que definir cualquiera de los siguientes métodos en el atributo *pAdjustMethod*:

“holm”, “hochberg”, “hommel”, “bonferroni”, “BH”, “BY”, “fdr”, “none”

La tabla de resultados presenta el error estándar del estadístico log2fold − change y el estadístico de Wald con el que realiza el contraste, además podrían observarse en algunos casos NA como resultado, esto se debe a alguna de las siguientes razones:

- Si para un gen, todas las muestras tienen conteos cero, la columna *baseMean* será cero para ese gen y la estimación del log2fold − change, el *pval* y el *padj* obtendrán el valor NA. Esto se debe a que el fold-change resultaría ser $\frac{0}{0} =$ IND y a partir de ahí 0 no sería posible realizar

los cálculos. Pero al hacer el filtrado de los genes previamente esto no sucederá.

- Si un gen contiene un conteo extremo, outlier, el *pval* y el *padj* serán NA.
- Si un gen tiene una baja frecuencia de conteos normalizados, entonces únicamente el *pval* será NA.

6.4.3 NOISeq

El paquete NOISeq de Bioconductor permite identificar aquellos genes diferencialmente expresados mediante una aproximación no paramétrica de los datos de conteo disponibles, es decir, dicho paquete permite computar la expresión diferencial entre dos condiciones mediante la identificación del nivel de expresión de los distintos genes bajo dichas condiciones.

NOISeq analiza y calcula la correspondiente expresión diferencial de datos con réplicas técnicas y/o biológicas (NOISeq-real) o sin réplicas (NOISeq-sim). Es necesario conocer el tipo de réplica con la que se trabaja. Existen dos tipos de réplicas, por un lado están las biológicas, donde en general, cada individuo perteneciente al estudio representaría una muestra, y por otro lado las réplicas técnicas, que son muestras tomadas de un mismo individuo, por lo que finalmente se trabajará con la media de todas ellas.

En este último tipo de réplica no es posible extraer conclusiones generales, sino que sólo son válidas para muestras que puedan ser directamente comparables.

Para cada gen, NOISeq calcula dos estadísticos de expresión diferencial: M (el logaritmo en base 2 del ratio entre dos condiciones) y D (el valor absoluto de la diferencia entre condiciones).

Este método se basa, fundamentalmente, en la hipótesis de que cambios en la expresión de un gen entre dos condiciones no necesariamente deben deducir que dicho gen esté diferencialmente expresado bajo dichas condiciones experimentales. Esto ocurre cuando la magnitud de dicho cambio coincide con los cambios en la expresión de ese gen entre réplicas de una misma condición. Por ello, además de computar los cambios en la expresión de cada gen entre dos condiciones experimentales, se debe obtener mediante la comparación de todas las réplicas dentro de la misma condición, la distribución del ruido, es decir, la distribución de los cambios en los valores de expresión de los genes cuando comparamos réplicas de la misma condición.

Se denota C_{gj} al número de reads (counts) para cada gen i en la j-ésima muestra o réplica de la condición experimental g (para NOISeq, g=1 ó 2), donde j varía de 1 hasta el número de muestras o réplicas de la condición g. Para realizar el cálculo de los valores M y D, NOISeq realiza la normalización de los counts de cada muestra o réplica por la media truncada de los M valores (trimmed mean of M values, TMM). Esta corrección para eliminar el ruido introducido por las variaciones biológicas naturales entre las muestras de cada condición no atribuibles a las condiciones en sí. Este método de normalización se basa en la aceptación de que la mayoría de los genes no están diferencialmente expresados.

Además, antes de proceder a la normalización, los niveles de expresión iguales a 0 son reemplazados por una constante dada $k > 0$, con el objetivo de evitar valores M infinitos o indeterminados. Los nuevos valores corregidos se denotan $x_{gj.}$ Para el cálculo de los dos estadísticos de expresión diferencial, M y D, x_g se define, en el caso de disponer de réplicas técnicas, como la suma de los valores de todas las muestras o réplicas de cada gen i para la condición experimental g;

en el caso de disponer de réplicas biológicas, como la media de todos los valores; y, en el caso de no disponer de réplicas, se define como los valores de cada gen i para la condición experimental g. Luego, para cada gen i se calculan los estadísticos de expresión diferencial como:

$$M^i = \log_2\left(\frac{x_1^i}{x_2^i}\right) \; y \; D^i = \left|x_1^i - x_2^i\right|$$

Como bien comentábamos anteriormente, la distribución del ruido se obtiene mediante la comparación de todas las réplicas dentro de la misma condición. Para su creación, en el caso de disponer de réplicas, se acumulan todos los valores (*M,D*) de todas las comparaciones entre réplicas de los distintos genes para cada condición. En el caso de no existir réplicas, éstas son simuladas y se sigue el mismo procedimiento que en el caso de disponer de réplicas.

NOISeq considera que un gen está diferencialmente expresado si sus correspondientes valores (*M, D*) son más altos que el ruido. La probabilidad de expresión diferencial de cada gen se obtiene a través de la comparación de los valores (*M,D*) de dicho gen y la distribución del ruido. Si el cociente entre la probabilidad de expresión diferencial y la probabilidad de expresión no diferencial es mayor que un determinado umbral q, es decir:

$$\frac{P(expresión\ diferencial)}{P(expresión\ no\ diferencial)} > q$$

Dicho gen se considera diferencialmente expresado entre las dos condiciones. En general, el valor de dicho umbral será q=0.8, dado que es equivalente a considerar que un gen está diferencialmente expresado cuando existe un odds ratio de 4:1, es decir, cuando podemos concluir que es 4 veces más probable que dicho gen esté expresado de forma diferencial entre ambas condiciones a que no lo esté. Por consiguiente, el valor de q será el umbral por el cual podemos considerar si hay más o menos genes diferencialmente expresados, dado que a mayor valor de q, mayor umbral y, por tanto, menos genes diferencialmente expresados entre

ambas condiciones. De forma viceversa ocurre para menor valor de q: menor umbral y mayor número de genes diferencialmente expresados entre las dos condiciones.

6.5 *Rubus glaucus* Benth.

La familia *Rosácea* presenta una gran importancia económica y amplia distribución, incluye cultivos de relevancia económica mundial como la fresa, la pera, las cerezas, el durazno, la frambuesa, la rosa y la mora, entre otras. Muchos miembros de esta familia se presentan naturalmente en América y algunos son cultivados masivamente en Colombia, como es el caso del cultivo de Mora de la zona Andina (*Rubus glaucus*) donde se encuentran diferentes variedades y se realiza comúnmente propagación vegetativa.

El cultivo de mora (*Rubus glaucus*, Benth) es una actividad económica importante en Colombia y la zona andina, con un mercado creciente en el mundo, debido al descubrimiento de los efectos benéficos de los poli fenoles para los humanos. Muchas familias de la zona andina dependen de esta actividad, sin embargo, el rendimiento productivo aún no es el esperado debido en gran medida a las pérdidas económicas producidas por algunas enfermedades, entre las que sobresale la antracnosis causada por *Colletotrichum gloeosporioides* (López-Vásquez et al., 2013). En la región cafetera de Colombia se han reportado incidencias de 52,9% en promedio (Botero et al., 2002), lo que la postula como la enfermedad más importante para el cultivo.

En estudios previos se identificaron y caracterizaron algunas especies de *Colletotrichum spp* como los agentes causales de la enfermedad en la zona cafetera de Colombia, identificando que el 81% de los casos fueron causados por *C. gloeosporioides*, estos hallazgos permitieron estandarizar el método de aislamiento e inoculación en la planta (Ramirez et al., 2007; Marulanda et al.,

2014), lo que permitió posteriormente identificar y caracterizar material tolerante al patógeno (López-Vásquez et al., 2013). Adicionalmente, se ha realizado la caracterización molecular de las variedades de mora andina incluyendo la tolerante al patógeno (Marulanda et al., 2012). Sin embargo, el conocimiento de la genética, del genoma y del transcriptoma aún es limitado y la mayoría de esfuerzos se han centrado en otros miembros de la familia Rosácea (https://www.rosaceae.org), lo que plantea la necesidad de desarrollar trabajos que permitan dilucidar mejor la composición genética, que en el futuro permita identificar variantes asociadas a características productivas, de calidad de la fruta y de tolerancia a enfermedades como la antracnosis causada por *C. gloeosporioides*, para tener en cuenta en programas de propagación y mejoramiento genético de la mora andina.

En el género *Rubus* se han reportado varios genes asociados a la calidad de la fruta, al metabolismo de poli propanoides y a la resistencia a enfermedades (Zheng and Hrazdina, 2010; Han et al., 2017), pero casi todos evaluados en frambuesa. Recientemente, (García-Seco et al., 2015) realizó el ensamble y análisis del transcriptoma de la fruta de *Rubus sp*. var. Lochness, que es un tipo diferente a las moras que se cultivan comercialmente en la zona andina, además de una parte diferente de la planta y sin considerar las consecuencias de la inoculación con un hongo como *C. gloesporoides*. Por lo anterior se hace necesario realizar estudios para el análisis del transcriptoma en *R. glaucus* Benth que permitan comprender sus particularidades genéticas. Además, recientemente se han liberado nuevas versiones del genoma de referencia de otros miembros de la familia Rosácea (Edger et al., 2018; Saint-Oyant et al., 2018; VanBuren et al., 2018), los cuales pueden ser usados para el análisis de expresión, con el fin de mejorar la comprensión del transcriptoma y la expresión de genes asociados a la tolerancia a patógenos.

7.0 METODOLOGÍA

Se seleccionaron EdgeR, Deseq2 y NOIseq como las herramientas que presentan un mejor rendimiento para detectar genes expresados diferencialmente cuando se tiene un limitado número de réplicas (n < 3). Esta selección se basó en las conclusiones del articulo titulado: "*How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?"* (Schurch et., 2016).

EdgeR, Deseq2 y NOIseq son paquetes de la biblioteca de Bioconductor del lenguaje de programación R.

El análisis de expresión diferencial a partir de RNA-Seq se realizó tres etapas: secuenciación, ensamblaje del transcriptoma y análisis de expresión diferencial, las dos primeras desarrolladas por el grupo de investigación en Biodiversidad y Biotecnología de la Universidad Tecnológica de Pereira.

7.1 MATERIALES *Rubus glaucus* Benth.

Las plantas muestreadas en este trabajo correspondieron a cultivos comerciales localizados en el municipio Guática, Risaralda, Colombia. Las plantas muestreadas en campo fueron propagadas in vitro en el laboratorio de Biotecnología Vegetal de la Universidad Tecnológica de Pereira. Para el estudio se usó material previamente reportadas como tolerantes a *C. gloespoiorides* (López-Vásquez et al., 2013) y material susceptible a este patógeno, los cuales fueron identificados por sus características genéticas, morfológicas y caracterizadas usando marcadores moleculares (López-Vásquez et al., 2013). A partir de estos grupos se logró la inoculación de dos aislados de cada cultivo (susceptible y tolerante). La concentración del inoculo fue de 1.2 X 106 esporas/ml y para el proceso de infección se incubó por 72 horas con cepas (3s1 y 6)

altamente patogénicas de *C. gloespoiorides* previamente aisladas y caracterizada (Ramírez et al., 2007). A partir de los materiales se establecieron tres grupos, el primero conformado por material susceptible a *C. gloespoiorides* no inoculado (SCNI), al cual se le aplicó agua estéril, el segundo correspondió a material susceptible a *C. gloespoiorides* e inoculado (SCI) y el tercero material tolerante al hongo e inoculado (TCI). De los tres grupos de plantas, se realizó una colecta en nitrógeno líquido del tejido inoculado y se procedió a realizar la extracción de RNA mediante el kit comercial FastTrack® kit MAG mRNA isolation kit (ThermoFisher Scientific, Waltham, MA, USA). Las muestras fueron tratadas con DNasas, para su posterior síntesis a cADN y secuenciación.

7.2 SECUENCIACIÓN.

Una vez obtenido el cADN, este fue refrigerado y enviado al Beijing Genomics Institute BGI (Hong-Kong, China) con el fin de preparar las librerías y realizar la secuenciación mediante la tecnología Illumina HiSeqTM 2000 (Illumina Inc., San Diego, CA, USA). La librería fue preparada en paired-end (pareado) siguiendo el protocolo de Illumina TruSeq RNA y una vez enviadas las lecturas estas fueron filtradas con el fin de eliminar adaptadores, lecturas menores de 36pb y de baja calidad (valor phred menor que 20) usando el software trimmomatic v0.38 (Bolger et al., 2014).

7.3 ENSAMBLAJE DEL TRANSCRIPTOMA.

Para el ensamblaje de secuencias se realizó una revisión de la base de datos Genome Database for Rosaceae (GDR) y se encontraron genomas de referencia de diferentes miembros de la familia, incluyendo un ensamblaje del género Rubus. En un trabajo previo se encontró que el usar genoma de referencia de occidentalis v3 (VanBuren et al., 2018) dio mejores resultados en el ensamblaje. A partir de la

información se realizó un alineamiento de lecturas generando un índice de acuerdo a cada genoma de referencia mediante el programa Bowtie 2 v2.3.4 (Langmean and Salzberg, 2012), posteriormente, se mapearon las lecturas al genoma de referencia mediante tophat v2.1.1 (Trapnell et al., 2009).

Tabla 3. Resumen del ensamble *de novo* del transcriptoma de *Rubus glaucus.*

Muestra	**Total de lecturas brutas**	**Longitud de lecturas brutas**	**%GC brutas/ limpias**	**Total de lecturas limpias**	**%Q20**	**Contigs**	**Longitud media**	**N50**	**L50**	**Clusters (UniGenes)**
Infectado tolerante	62 000 000	66.8/90	45/41.5	46,022,646	96.5	121647	317.7	300	81487	95311
Infectado Susceptible	51 516 018	80.1/90	51/43.9	45843084	97.2	58652	464.1	527	45521	46364
Susceptible	54 446 562	77.6/90	52/43.9	46978736	97.7	58636	461.3	522	45353	46382
Total	167 962 580	74.4/90	49.3/44.6	138844466	97.1	55185	443.9	489	42362	43579

7.4 ANÁLISIS DE EXPRESIÓN DIFERENCIAL.

Para el análisis de expresión diferencial se parte del hecho que sólo se cuenta con una réplica del experimento por cada condición. Los métodos propuestos para este análisis son EdgeR, NOI-seq y Deseq2. Sin embargo, para garantizar cual método tiene el mejor desempeño se analizaron datos de un experimento altamente replicado (48 réplicas). Los datos del experimento altamente replicado corresponden a *Saccharomyces cerevisiae* el cual es uno de los organismos mejor estudiados en biología molecular con un transcriptoma relativamente pequeño. Este experimento usa dos condiciones, la primera denominada como WT en condición normal tipo salvaje y snf2 mutante, ambas con el mismo background genético. Estos datos están disponibles en el Repositorio del Archivo Europeo de Nucleótidos (PRJEB5348, http:// www.ebi.ac.uk/ena/data/view/ERX425102).

Se escogieron 4 réplicas al azar (44, 32, 48 y 11), se analizaron con cada uno de los métodos, los resultados se contrastaron con el reporte general de los genes

expresados diferencialmente cuando se corren las 48 réplicas, para así determinar cuántos genes expresados diferencialmente encontrados con cada método coincidían con los originales. Posteriormente con base en la cantidad de genes reportados como expresados diferencialmente, se comparó con los que realmente habían sido reportados como diferencialmente expresados y se calculó la verdadera tasa de falsos positivos encontrados con cada método y se comparó con la estimación. A partir de esto, se escogió el mejor método para analizar los datos de ***Rubus glaucus*** **Benth** y determinar la expresión diferencial.

7.4.1 Análisis con *Saccharomyces Cerevisiae*.

Cada análisis comienza con la matriz de conteos de cada una de las réplicas, la cual contiene tres columnas. La primera columna corresponde al nombre del gen y las otras dos columnas a las condiciones a analizar con sus respectivos conteos. La tabla de conteo de este experimento tiene 7131 genes. La siguiente tabla muestra los primeros 6 genes de la tabla de conteo para la réplica 11.

Tabla 4. Tabla de Conteos *Saccharomyces Cerevisiae* Réplica 11.

Nombre del Gen	Condición 1 (Snf2)	Condición 2 (WT)
15S_rRNA	3	4
21S_rRNA	24	70
alignment_not_unique	0	0
ambiguous	544028	803015
HRA1	1	1
ICR1	145	106

Como ya se explicó anteriormente, tanto EdgeR como DESeq2 asumen que los datos de conteo siguen una distribución binomial negativa y NOIseq se basa en una aproximación no paramétrica de los datos de conteo. A continuación se presentan los gráficos de densidad de los datos para entender mejor su comportamiento.

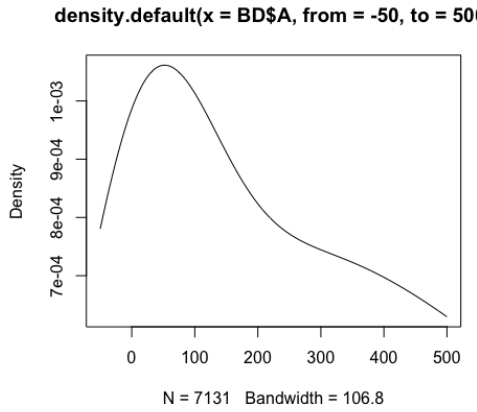


Ilustración 1. Gráfico de Densidad Réplica 11 Condición 1 (Snf2). (WT).

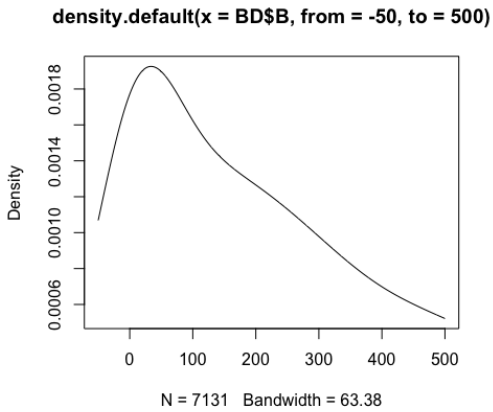


Ilustración 2. Gráfico de Densidad Réplica R11 Condición 2

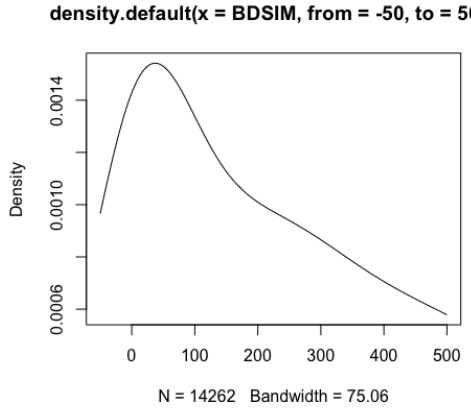


Ilustración 3. Gráfico de Densidad Réplica 11 Condición 1 y 2. (Snf2).

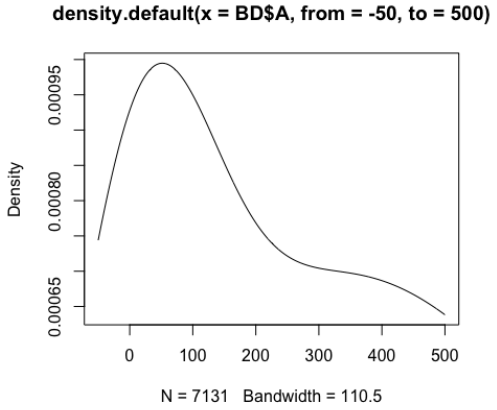


Ilustración 4. Gráfico de Densidad Réplica 32 Condición 1

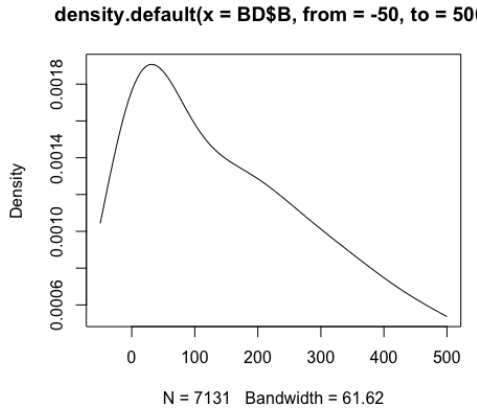


Ilustración 5. Gráfico de Densidad Réplica 32 condición 2 (WT).

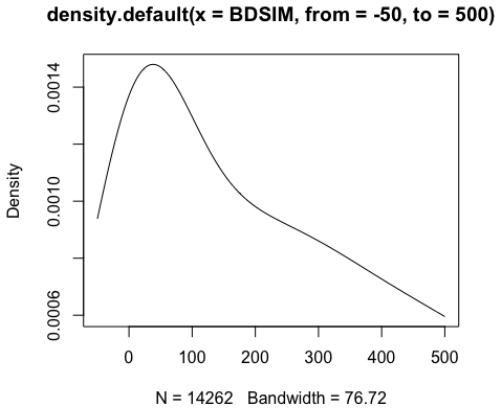


Ilustración 6. Gráfico de Densidad Réplica 32 Condición 1 y 2.

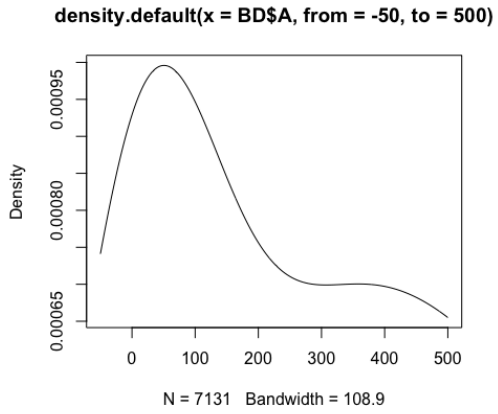


Ilustración 7. Gráfico de Densidad Réplica 44 Condición 1 (Snf2). (WT).

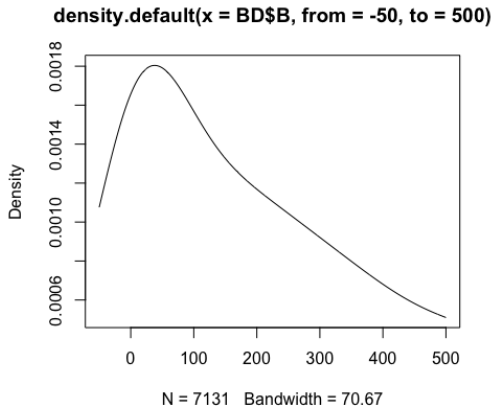


Ilustración 8. Gráfico de Densidad Réplica 44 Condición 2 (WT).

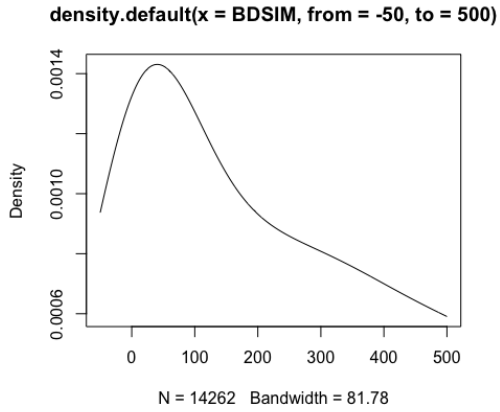


Ilustración 9. Gráfico de Densidad Réplica 44 Condición 1 y 2. (Snf2).

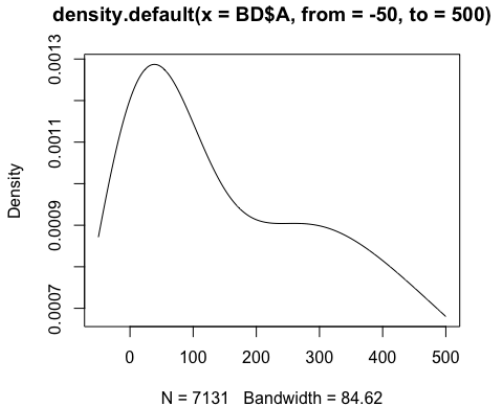


Ilustración 10. Gráfico de Densidad Réplica 48 Condición 1 (Snf2).

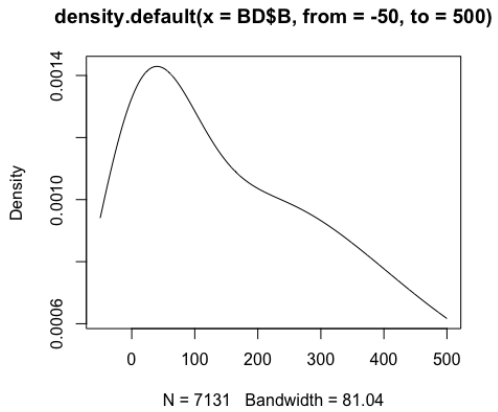


Ilustración 11. Gráfico de Densidad Réplica 48 Condición 2 (WT).

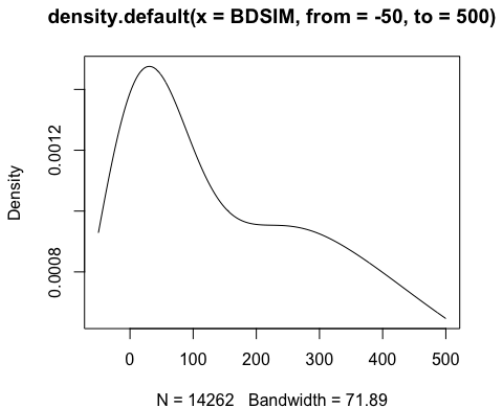


Ilustración 12. Gráfico de Densidad Réplica 48 Condición 1 y 2.

A partir de los gráficos anteriores podemos concluir que los datos presentan mayor concentración a la izquierda con una asimetría positiva, compatibles con una distribución de conteo de Poisson.

7.4.1.1 Edge R.

Para iniciar el análisis con EdgeR lo primero que se hizo fue importar la tabla conteo a R para cada réplica. El siguiente paso fue el filtrado de los genes. Se eliminaron los genes que tenían muy baja expresión, manteniendo los que tenían un nivel razonable. Puesto que se tenía un grupo de dos condiciones, se mantuvieron los genes que lograron al menos un cpm (conteos por millón) de dos.

La función *cpm(·)* calculó los conteos por millón de cada muestra, es decir, la suma de cada columna tras aplicar esta función fue de un millón.

Se impuso la condición de cpm(naseqMatrix) > 1, la cual creó una matriz con valores lógicos donde aparecía TRUE si la condición se cumplía y FALSE si no se cumplía. En general, a los valores TRUE se les asigno unos y a los valores FALSE se les asigno ceros, por lo que al calcular luego la suma por filas con la función *rowSums(·)* lo que se hizo realmente fue contar cuantos valores positivos se tenían por filas en la matriz de valores lógicos construida anteriormente. Tras hacer esto, sólo se mantuvieron las filas, es decir, los genes, en los que esta suma o recuento fue mayor o igual a dos.

El siguiente paso fue estimar los factores de normalización mediante la función *calcNormFactors*, la cual proporciona un conjunto de factores de normalización que minimizan el *logfold-change.* El cálculo de dichos factores de normalización se realiza por la media truncada de los valores M (trimmed mean of M values, TMM) entre cada par de genes.

EdgeR se basa en el supuesto de los datos de conteo se modelan mediante una Binomial Negativa, por lo que antes de proceder a realizar el correspondiente test para determinar los genes diferencialmente expresados, se estimó el parámetro de dispersión. En primer lugar, se estimó la dispersión común para todos los genes a través de la función *estimateCommonDisp,* es decir, dicho parámetro proporciona una idea general de la variabilidad de los datos. Esta función calcula el coeficiente de variación biológica (BCV), que constituye la raíz cuadrada de dicho parámetro de dispersión y representa el coeficiente de variación entre réplicas de la misma condición. Sin embargo, al no tener réplicas para este análisis no se pudo determinar este parámetro. Teniendo en cuenta la naturaleza de los datos se decidió asignar un parámetro de dispersión de 0.2.

Finalmente, se realizó el test para determinar la expresión diferencial de los genes a través de la función *exactTest.*

Tras calcular las pruebas exactas, se aplicó a estos resultados la función *topTags(·)*, pues con esta función se obtiene una tabla, igual que la obtenida con la función *exactTest*, sin embargo tiene una columna en la que se encuentra el p-valor ajustado y la columna FDR que proporciona la probabilidad de error de tipo I. El criterio para seleccionar los genes que se expresan diferencialmente fueron aquellos con FDR < 0.1. Este procedimiento se repitió para cada una de las réplicas.

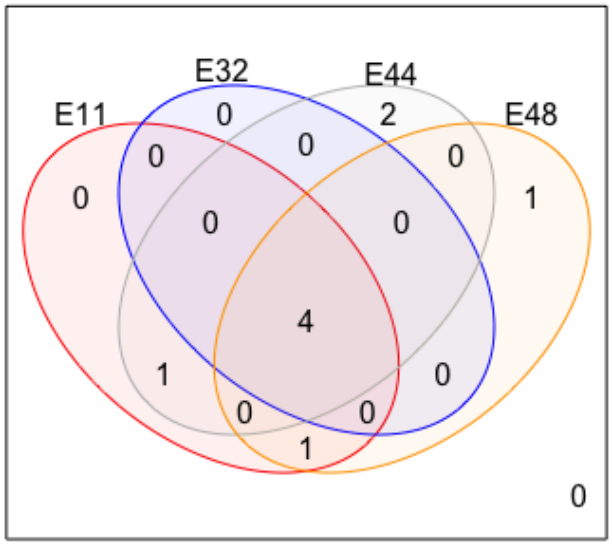


Ilustración 13. Diagrama de Venn entre los genes diferencialmente expresados en S *Cerevisiae* con el método EdgR para las réplicas 11, 32, 44 y 48.

7.4.1.2 Noiseq

Con la tabla de conteo importada a R se creó un data frame con estos datos y otro con las condiciones de los genes, condición 1 (snf2) y condición 2 (WT). Para calcular los genes expresados diferencialmente el primer paso fue invocar la función *noiseq* que incluye como criterios los dos date frame anteriores y para este caso como no se disponían de réplicas se introdujo el criterio replicates = "no". La función *noiseq* permite obtener los valores de los dos estadísticos de expresión diferencial para cada tránscrito: M (el logaritmo en base 2 del ratio entre dos condiciones) y D (el valor absoluto de la diferencia entre condiciones). Una vez obtenidas dichas estimaciones, utilizamos la función *degenes* para conocer aquellos genes en los cuales el cociente entre su probabilidad de expresión diferencial y su probabilidad de no expresión diferencial fuera mayor que el umbral *q*. Cuando se usa NOISeq con réplicas, el manual del usuario de Bioconductor recomienda que el valor de *q* esté en torno a 0.8, para propósitos de este proyecto

que hace uso del análisis sin réplicas, se determinó que *q* fuera 0.95 para hacerlo más exigente en la selección de los genes expresados diferencialmente.

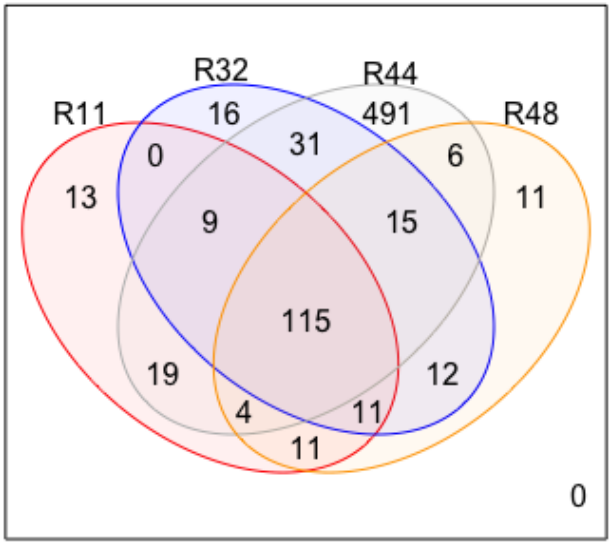


Ilustración 14. Diagrama de Venn entre los genes diferencialmente expresados en *S Cerevisiae* con el método Noiseq para las réplicas 11, 32, 44 y 48.

7.4.1.3 DESeq2.

A diferencia de los dos primeros métodos, al intentar correr cada una de las réplicas con este paquete, no se obtuvieron genes diferencialmente expresados. El paquete no arrojó ninguna salida. Esto se puede explicar por la ausencia de réplicas al hacer el análisis.

7.4.2 Análisis *Rubus glaucus*, Benth.

Del análisis con *S cerevisiae* se concluyó que EdgeR era el mejor método para el análisis de expresión diferencial de *Rubus glaucus*, Benth.

Como se explicó para *S cerevisiae,* el primer paso fue importar la tabla de conteo a R. Como *Rubus glaucus*, Benth tiene tres condiciones, se crearon 3 tablas de conteo, la primera (suVSco) con la condición 1 (Susceptible) y condición 2 (Control), la segunda tabla de conteo (suVSto) con la condición 1 y condición 3 (Tolerante) y una tabla de conteo (coVSto) con las condiciones 2 y 3.

Para entender mejor el comportamiento de los datos a continuación se presentan las gráficas de densidad.

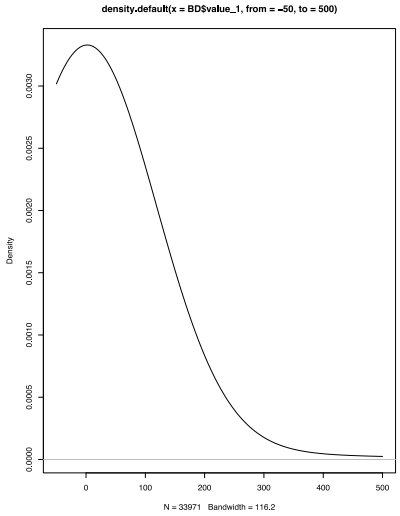


Ilustración 15. Gráfica densidad Condición 1 (Susceptible).

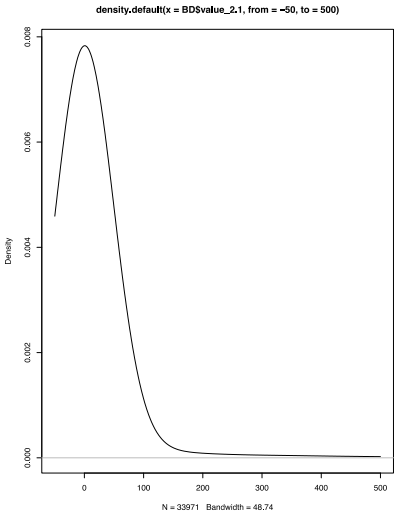


Ilustración 16. Gráfica Densidad Condición 2 (Control)

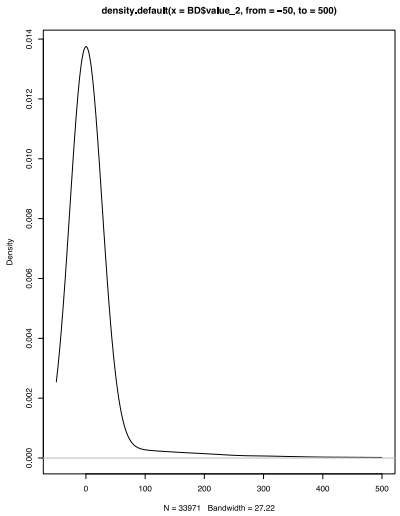


Ilustración 17. Gráfica Densidad Condición 3 (Tolerante).

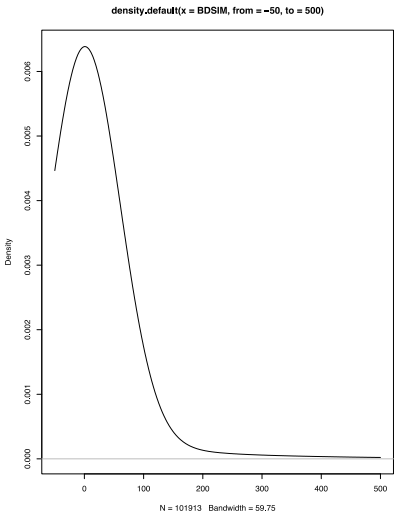


Ilustración 18. Gráfica Densidad condiciones 1,2 y 3.

A partir de los gráficos anteriores podemos concluir que los datos presentan mayor concentración a la izquierda con una asimetría positiva, compatibles con una distribución de conteo de Poisson.

Ya con las tablas de conteo el siguiente paso consistió en el filtrado de los genes. Previamente se eliminaron manualmente los genes con conteos de cero en cada par de condiciones, pasando de 33.971 genes a sólo 2.478. Lo que se hizo después fue eliminar los que tenían muy baja expresión, manteniendo los que tenían un nivel razonable. Para este propósito se utilizó la función *cpm(·)* que calcula los conteos por millón de cada muestra. Se definió como criterio que se mantendrían los genes que lograrán al menos un cpm (conteos por millón) de dos. Después de aplicar esta función se pasó de 2.478 genes a 1.827.

Al igual que con *S cerevisiae* no se pudo calcular el parámetro de dispersión por no disponer de réplicas. Se decidió asignar un parámetro de dispersión de 0.2 teniendo en cuenta la naturaleza de los datos.

Finalmente, se procedió a realizar el test para determinar la expresión diferencial de los genes a través de la función *exactTest.* Tras calcular las pruebas exactas, se aplicó a estos resultados la función *topTags(·)*.

Tabla 5. Resultados función *topTags(·)* en *Rubus glaucus*, Benth. Condición 1 y 3.

	row.names.suVSto.	GEN	value_1	value_2	sampleA	sampleB	logFC	logCPM	PValue	FDR
1	1	GEN01	81.084	234.813	value_1	value_2	-1.518903482	7.873688169	0.118583386	0.382102021
2	10	GEN10	336.61	209.887	value_1	value_2	1.64471611	8.603497211	0.090673453	0.331984768
3	100	GEN111	2127.55	807.146	value_1	value_2	0.516318559	10.28284159	0.583851364	0.816006103
4	1000	GEN1332	52.445	77.7339	value_1	value_2	-1.474440426	5.661444445	0.163249144	0.454807308
5	1001	GEN1335	724.943	1892.64	value_1	value_2	0.095917535	7.086814068	0.934240929	0.982081805

La tabla 5 presenta los 5 primeros genes, la columna FDR proporciona la probabilidad de error de tipo I, la columna logFC, correspondiente al $\log_2$fold-change, muestra el cambio en la proporción de los conteos para ambas condiciones en función del $\log_2$.

El criterio para seleccionar los genes que se expresados diferencialmente fueron

aquellos con FDR < 0.1.

7.4.2.1 Anotación de los genes diferencialmente expresados

Una vez identificados haciendo uso de EdgeR los 16 genes asociados a la tolerancia de los tratamientos *Rubus glaucus* Benth, se procedió a realizar la anotación funcional que consiste en la localización de los genes en un genoma de referencia, para este paso se usó **JBrowse** herramienta disponible en **GDR** (Genome Database for Rosaceae). El genoma de referencia que se empleó es el de la especie *Rubus occidentalis* v3.0

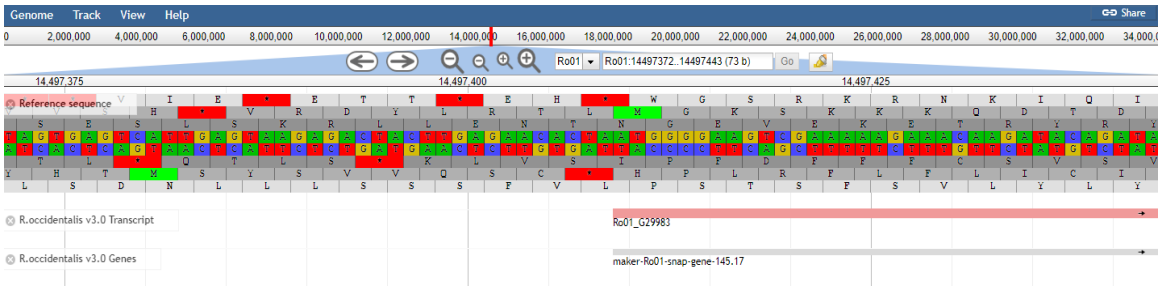


Ilustración 19. Localización del gen Ro01_G29983

La herramienta JBrowse nos proporciona la secuencia del gen, lo que nos permitiré identificarlo a través de la información disponible en Blast

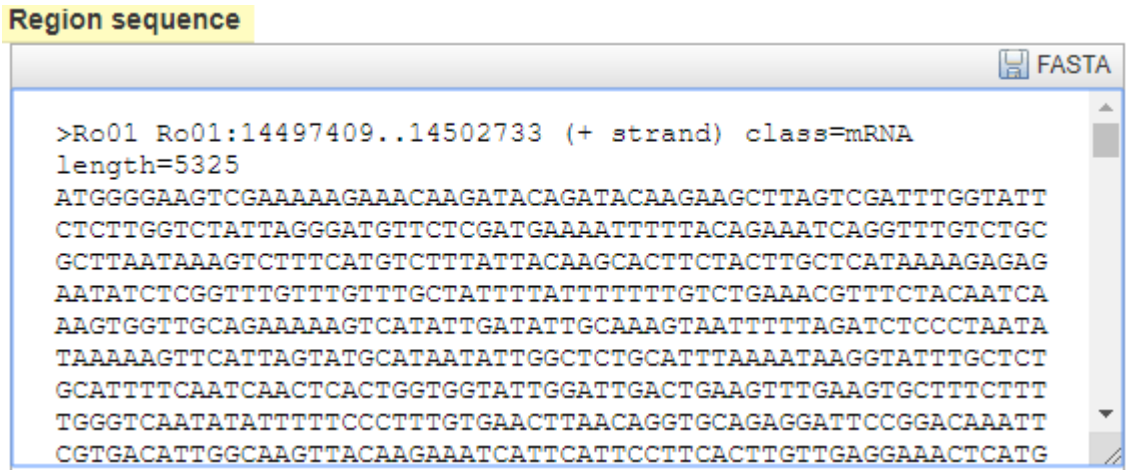


Ilustración 20. Región de secuencia del gen Ro01_G29983

Blast es una herramienta de alineamiento local por pares que opera a través de un algoritmo que permite realizar una búsqueda preliminar de similitud entre una secuencia problema y las bases de datos disponibles.

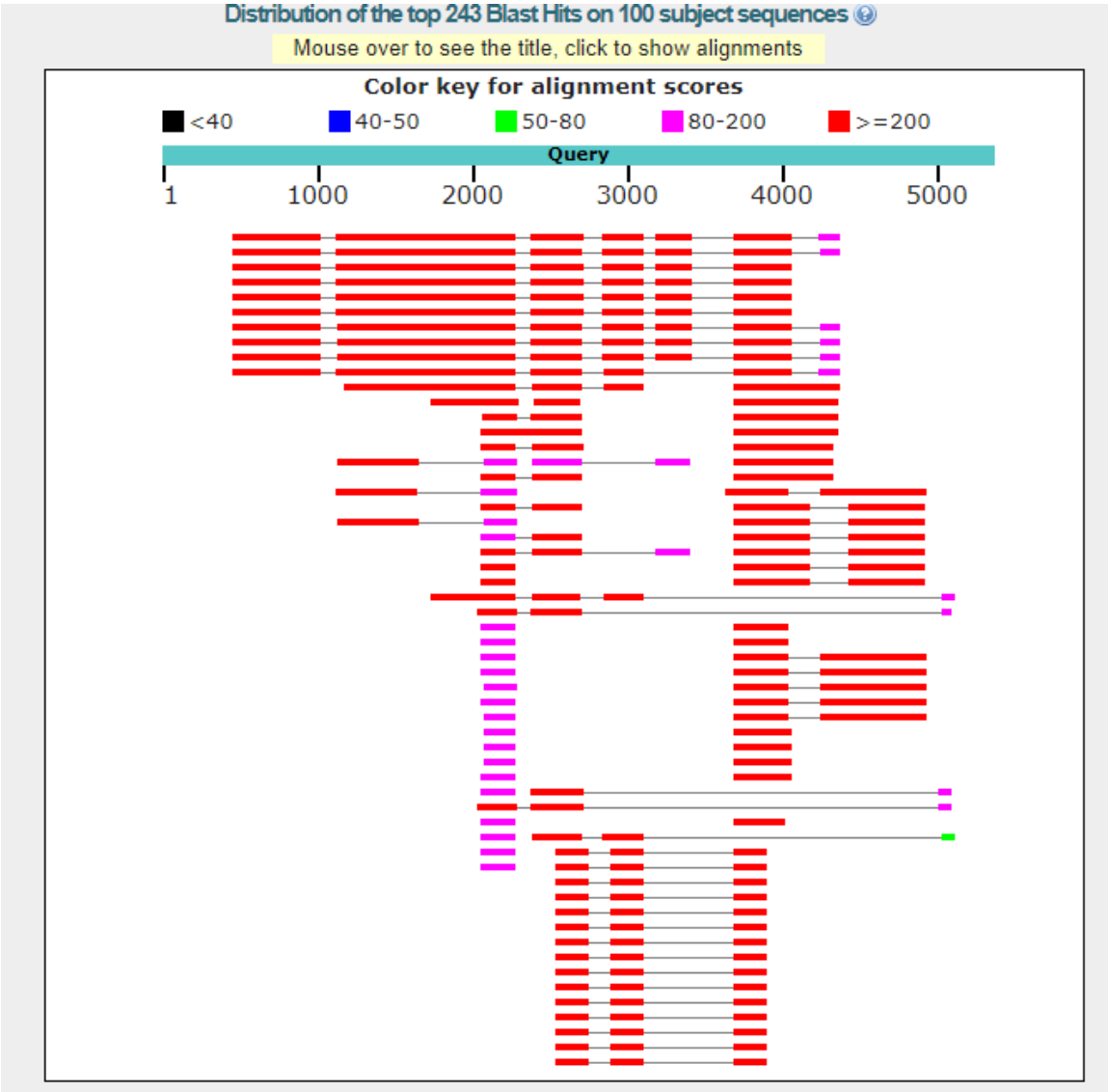


Ilustración 21. Región de secuencia del gen Ro01_G29983

Blast proporciona información de los alineamientos más significativos identificados en las bases de datos. Esta información permite identificar el nombre con el que se encuentra reportado el gen, a partir de lo cual es posible determinar la proteína que está regulando y su función.

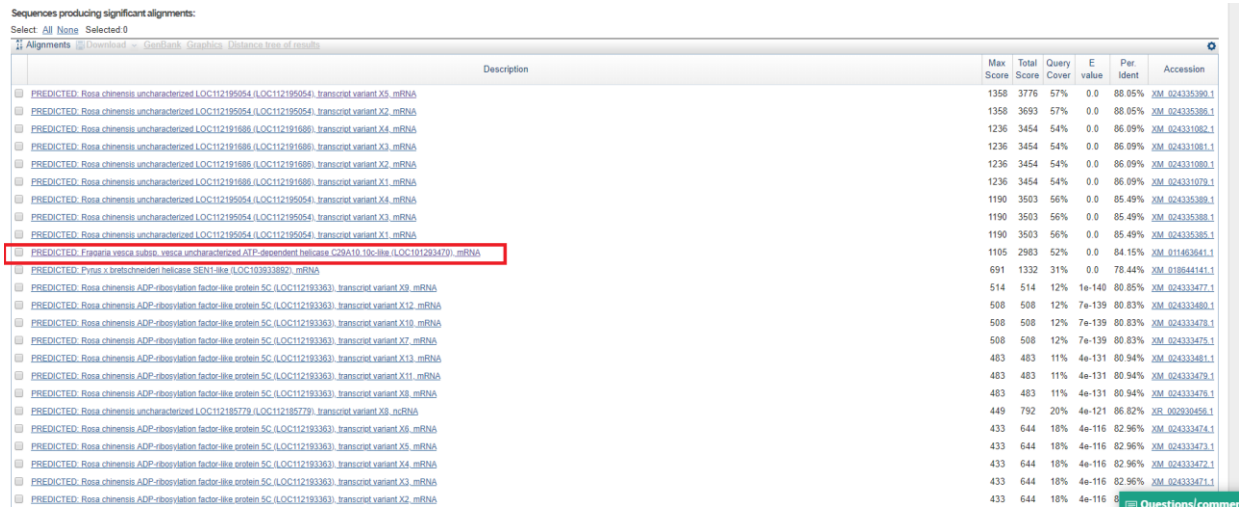


Ilustración 22. Alineaciones significativas del gen Ro01_G29983

En este caso específico, el gen Ro01_G29983 ha sido reportado en *Fragaria vesca* con el nombre de ATP-dependent helicase C29A10.10c. Haciendo uso de **Uniprot** una herramienta que proporciona información de las secuencias las proteínas y su función, se estableció que este gen participa en los procesos de duplicación y reproducción celular, transcripción, recombinación y reparación del ADN, y la biogénesis de ribosoma.

La descripción ontología de los genes se realizó a través de the gene ontology research (http://geneontology.org/). El proyecto de ontología de genes (Ashburner, et al. 2000; Gene Ontology Consortium 2015) es la mayor iniciativa bioinformática para desarrollar una representación computacional del conocimiento de la evolución de cómo los genes codifican las funciones biológicas. El proyecto ha desarrollado ontologías formales que representan más de 40.000 conceptos biológicos y se están revisando constantemente para reflejar nuevos descubrimientos.

8.0 RESULTADOS

***Saccharomyces Cerevisiae*.**

Al correr cada una de las distintas réplicas en EdgeR, se encontraron siempre 4 genes expresados diferencialmente comunes entre réplicas (Ilustración 13). En promedio, este método encontró 6 genes expresados diferencialmente por réplica con reporte de un solo gen como falso positivo en la réplica 44, sin embargo, la cantidad total de genes expresados diferencialmente cuando se corren las 48 réplicas fue 4763. Este resultado es completamente entendible, puesto que este paquete no está diseñado para trabajar con una sola réplica, lo único que se buscaba era entender el comportamiento de EdgeR bajo ese escenario.

Cuando se analizaron los mismos datos con NOI-seq se encontraron siempre 115 genes expresados diferencialmente comunes entre réplicas (Ilustración 14). Para la réplica 44 se encontraron 800 genes diferencialmente expresados con 110 genes falsos positivos, la réplica 32 reportó 214 genes diferencialmente expresados con 5 genes falsos positivos, la réplica 48, 189 genes diferencialmente expresados con 4 genes falsos positivos, finalmente la réplica 11 reportó 188 genes diferencialmente expresados con 6 genes reportados como falsos positivos.

Finalmente, al intentar correr cada una de las réplicas con Deseq2, no se obtuvieron genes diferencialmente expresados, el paquete no arrojó ninguna salida. Como se ha explicado anteriormente, la limitación en el número de réplicas disponibles para el análisis compromete los resultados del análisis de expresión diferencial, en este caso Deseq2 al no disponer de réplicas para el análisis, no pudo determinar cuáles genes se expresaban diferencialmente.

La ilustración 23 muestra la intersección de los genes diferencialmente expresados en común entre los diferentes métodos a través del análisis de las distintas réplicas. EdgeR encontró un total de 9 genes diferencialmente expresados que son diferentes entre si en las cuatro réplicas analizadas, mientras

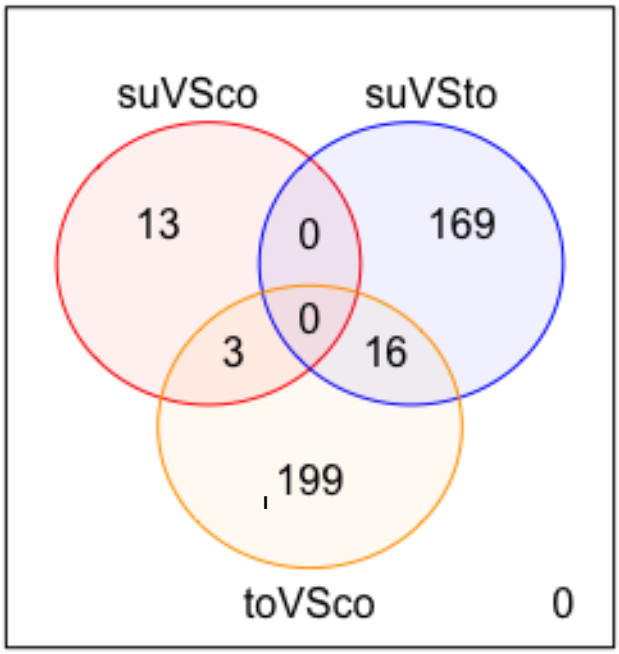
NOI-seq identificó 764 genes. Al cruzar ambos resultados se puede observar que los genes detectados por EdgeR fueron igualmente reportados como genes diferencialmente expresados con NOI-seq.

Ilustración 23. Diagrama de Venn entre los genes diferencialmente expresados en *S Cerevisiae* usando el método EdgeR vs NOI-seq.

Con base en los resultados anteriores se determinó EdgeR como el método más adecuado para calcular la expresión diferencial de los tratamientos *Rubus glaucus*, BENTH. Aunque NOI-seq encontró una mayor cantidad de genes expresados diferencialmente en comparación a EdgeR, este último sólo reportó un gen como falso positivo, de la misma manera los genes reportados por EdgeR como diferencialmente expresados, también fueron reportados por NOI-seq. Además, NOI-seq tiene la limitación de no poder simular la variabilidad bilógica, la cual es necesaria para análisis inferenciales en la población. En este caso se prefirió ser prudente con la selección del método, donde prima la calidad sobre la cantidad de los genes diferencialmente expresados.

***Rubus glaucus*, Benth.**

Después de hacer el análisis de expresión diferencial para los distintos tratamientos de *Rubus glaucus*, Benth con EdgeR, se identificaron 16 genes expresados diferencialmente para susceptible versus control, 218 genes para tolerante versus control y 185 genes diferencialmente expresados para susceptible versus tolerante.



lustración 24. Diagrama de Venn entre los genes diferencialmente expresados en *Rubus glaucus*, Benth para los distintos tratamientos.

La ilustración 17 muestra que existen 16 genes asociados a la tolerancia de los tratamientos *Rubus glaucus*, Benth. Esto se determinó al cruzar los resultados obtenidos en los tres grupos de análisis, encontrado que los 16 genes fueron detectados en los dos grupos que estaban asociados al análisis de genes tolerantes contrastados con los susceptibles y de control.

8.1 ANOTACIÓN DE LOS GENES DIFERENCIALMENTE EXPRESADOS

A continuación, se presentan en la tabla 6 las funciones identificadas para los 16 genes expresados diferencialmente:

Tabla 6. Genes diferencialmente expresados en material de *R. glaucus* tolerante a *C. gloesporoides* con la anotación funcional

ID	Loci	Nombre Gen	Ruta metabólica y Función	Descripción de la ontología del gen	Regulación de la muestra tolerante	Referencias reportadas
GEN1117	Ro01:14497408-14502733	ATP-dependent helicase C29A10.10c	Participa en los procesos de duplicación y reproducción celular, transcripción, recombinación y reparación del ADN, y de biogénesis de ribosoma	Proteólisis del ATP dependiente	Sobreexpresado (Up)	----------
GEN1416	Ro01:21520523-21522304	3'-N-debenzoyl-2'-deoxytaxol N-benzoyltransferase	Cataliza el acoplamiento estereoselectivo del sustrato sustituto N-debenzoil- (3'RS) -2'-deoxitaxol con benzoil-CoA para formar predominantemente un 3'-epímero de 2'-deoxtaxol	Función molecular	Sobreexpresado (Up)	Asociado con la acumulación de axol y taxanos en plantas (Onrubia et al, 2011)
GEN1988	Ro01:1467234-1468914	Phosphatidylinositol/phosphatidylcholine transfer protein SFH12-like	Requerido para el transporte de proteínas secretoras del complejo de Golgi (por similitud). Cataliza la transferencia de fosfatidilinositol y fosfatidilcolina entre membranas in vitro.	Proceso biológico	Sobreexpresado (Up)	Asociado con actividad antifúngica en *Saccharomyces cerevisiae* (Pries et al, 2017)
GEN153	Ro01:1671742-1679324	Ethanolamine kinase	Participa en la biosíntesis de fosfolípidos. Cataliza el primer paso en la biosíntesis de fosfatidiletanolamina.	Función molecular	Sobreexpresado (Up)	Asociada a tolerancia al estrés ER inducido en *Arabidopsis* (Lyn et al, 2019)

GEN1980	Ro01:1317634-1325513	Trichome differentiation protein GL1	Activador de transcripción, cuando se asocia con BHLH2 / EGL3 / MYC146 o BHLH12 / MYC1. Participa en la especificación del destino de las células epidérmicas en las hojas. Junto con TTG1 y GL3, promueve la formación de tricomas y la endoreplicación. Regula la producción de una señal que induce células precursoras del cabello (tricoma) en los primordios de las hojas para diferenciarse. Se une a las regiones promotoras de los sitios de unión a WER (WBS) y activa la transcripción de genes diana	----------	Sobreexpresado (Up)	Asociada a formación y ubicación de tricomas en las hojas en *Arabidopsis thaliana* (Liu et al, 2016)
GEN410	Ro01:4322460-4326741	Suppressor of mec-8 and unc-52 protein homolog 2	Proteína espliceosomal auxiliar involucrada en el empalme de pre-ARNm específicos que afectan múltiples aspectos del desarrollo	----------	Sobreexpresado (Up)	Asociado a el empalme de pre-ARNm específicos que afectan múltiples aspectos del desarrollo de Zea mays y *Arabidopsis thaliana* (Chung et al, 2009)
GEN438	Ro01:4554148-4557146	Thioredoxin reductase NTRB	Posee actividad tiorredoxina-disulfuro reductasa	----------	Sobreexpresado (Up)	----------
GEN652	Ro01:7068306-7070215	Lyrata crooked neck-like protein 1	Involucrado en el proceso de empalme pre-ARNm	----------	Sobreexpresado (Up)	----------
GEN948	Ro01:11321455-11321875	Co-CT7 sequence	No reportado	----------	Sobreexpresado (Up)	----------
GEN113	Ro01:1225991-1229775	Proline-rich protein-like	Inhibidor bifuncional / proteína de transferencia de lípidos de plantas	Función molecular	Subexpresado (Down)	Asociado a controla el alargamiento de los pelos radiculares en *Arabidopsis*

						thaliana (Boron et al, 2014)
GEN1884	Ro01:236593-246286	sm-like protein LSM6A	Componente de los complejos de proteínas LSM, que están involucrados en el procesamiento de ARN. El complejo citoplásmico LSM1-LSM7 regula la expresión de los genes del desarrollo mediante el destape de las transcripciones específicas relacionadas con el desarrollo. Juega un papel crítico en la regulación de la expresión de genes relacionados con el desarrollo.	Componente celular	Subexpresado (Down)	----------
GEN1946	Ro01:870562-875704	Protein phosphatase 2C 27	Confiere tolerancia a la sal al desencadenar la expresión de genes sensibles al estrés.	Proceso biológico	Subexpresado (Down)	----------
GEN2127	Ro01:3011787-3014146	Inositol oxygenase 4	No reportado	----------	Subexpresado (Down)	Se expresa bajo condiciones de restricción de nutrientes o baja energía en *Arabidopsis thaliana* (Alford et al, 2012)
GEN2145	Ro01:3232537-3238797	K(+) efflux antiporter 4-like	El movimiento dirigido de iones de potasio (K +), fuera o dentro de una célula, o entre células, por medio de algún agente como un transportador o poro.	Proceso biológico	Subexpresado (Down)	----------
GEN424	Ro01:4478822-4484797	Protein FATTY ACID EXPORT 2	Puede estar involucrado en la exportación de ácidos grasos libres de los plastidios	----------	Subexpresado (Down)	Asociado al desarrollo de semillas en *Arabidopsis* (Tia et al; 2019)

9.0 CONCLUSIONES

- Los métodos para el análisis de expresión diferencial requieren una cantidad de réplicas dentro del experimento para que los resultados sean estadísticamente significativos. La revisión del estado del arte de los métodos de expresión diferencial permitió identificar que los métodos más adecuados cuando se cuenta con una cantidad de réplicas inferior a tres son Deseq2, EdgeR y NOI-seq.

- La comparación de los métodos de análisis de expresión diferencial Deseq2, EdgeR y NOI-seq, a partir del análisis de datos de *Saccharomyces Cerevisiae* permite concluir que EdgeR tiene un desempeño satisfactorio en la detección de los genes diferencialmente expresados con bajas de tasas de falsos positivos, por tanto, se puede emplear cuando se tengan réplicas reducidas y aun cuando no se cuente con réplicas de los tratamientos

- NOI-seq tiene un bajo desempeño en la detección de los genes diferencialmente expresados, dado que reporta altas tasas de falsos positivos.

- Deseq2 es un método muy sensible a la ausencia de réplicas, por tanto no es recomendable emplearlo cuando se cuente con un número reducido de réplicas o en ausencia de ellas.

- Los 16 genes candidatos identificados a través de EdgeR se podrían validar con posterioridad a través de la técnica de PCR en tiempo real, la cual es muy sensible para la detección de ácidos nucleicos. A partir de los resultados obtenidos sería posible identificar SNP’s asociados a

características de interés de *Rubus glaucus* Benth aportando así a un futuro programa de mejoramiento genético de la especie.

- Las funciones de los genes candidatos identificados a través de EdgeR están asociados a procesos celulares tales como la reparación del ADN, la transferencia de lípidos y la biosíntesis de fosfolípidos, el empalme de pre-ARNm, tolerancia a la sal y regulación en la producción de una señal que induce células precursoras de (tricoma) en los primordios.

10. BIBLIOGRAFÍA

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29. doi: 10.1038/75556

Bolger, a. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible read trimming tool for Illumina NGS data. Bioinformatics 30(15): 2114–2120. http://bioinformatics.oxfordjournals.org/content/30/15/2114.

Botero, M., G. Rios, G. Franco, M. Romero, J. Pérez, J. Morales, J. Gallego, and D. Echeverry. 2002. Identificación y especialización de enfermedades asociadas a los cultivos de mora (*Rubus glaucus* Benth) en el eje cafetero. p. 87–92.

Bullard, E. Purdom, K.D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics, 11(1):94, 2010.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.

Carrasco Sara. Técnicas de análisis de expresión diferencial basadas en conteos para el estudio de datos de RNA-seq usanso R y bioconductor. En: https://idus.us.es/xmlui/bitstream/handle/11441/40563/Carrasco%20Carrasco%20Sara%20TFG.pdf?sequence=1

Conesa, A., and S. Götz. 2008. Blast2GO: A comprehensive suite for functional

analysis in plant genomics. Int. J. Plant Genomics. doi: 10.1155/2008/619832.

De Hertogh B, De Meulder B, Berger F, Pierre M, Bareke E, Gaigneaux A, Depiereux E. 2010. A benchmark for statistical micro- array data analysis that preserves actual biological and technical variance. BMC Bioinformatics 11: 17

Edger, P.P., R. VanBuren, M. Colle, T.J. Poorten, C.M. Wai, C.E. Niederhuth, E.I. Alger, S. Ou, C.B. Acharya, J. Wang, P. Callow, M.R. McKain, J. Shi, C. Collier, Z. Xiong, J.P. Mower, J.P. Slovin, T. Hytönen, N. Jiang, K.L. Childs, and S.J. Knapp. 2018. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. Gigascience. doi: 10.1093/gigascience/gix124.

Fatemeh Seyednasrollah, Asta Laiho, Laura L. Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics, Volume 16, Issue 1, 1 January 2015, Pages 59–70.

Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT. 2014. Differential expression analysis of RNA-seq data at single-base resolution. Biostatistics 15: 413–426.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. doi:10.1093/bioinformatics/bts565.

García-Seco, D., Y. Zhang, F.J. Gutierrez-Mañero, C. Martin, B. Ramos-Solano, F. Gutiérrez, C. Martin, and B. Ramos-Solano. 2015. RNA-Seq analysis and transcriptome assembly for blackberry (Rubus sp. Var. Lochness) fruit. BMC Genomics 16(1): 1–11. doi: 10.1186/s12864-014-1198-1.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Di Palma, B.W. Birren, C. Nusbaum, K.

Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. doi: 10.1038/nbt.1883.

Han, Y., H. Wan, T. Cheng, J. Wang, W. Yang, H. Pan, and Q. Zhang. 2017. Comparative RNA-seq analysis of transcriptome dynamics during petal development in Rosa chinensis. Sci. Rep. 7(February): 1–14. doi: 10.1038/srep43382.

Hardcastle TJ, Kelly KA. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11: 422J.H.

Kvam VM, Liu P, Si Y. 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 99: 248–256.

Langmean, B., and S.L. Salzberg. 2012. Bowtie 2. Nat. Methods 9(4): 357–359. doi: 10.1038/nmeth.2250.Digestion.

López-Vásquez, J.M., J. Castaño-Zapata, M.L. Marulanda-Ángel, and A.M. López-Gutiérrez. 2013. Caracterización de la resistencia a la antracnosis causada por Glomerella cingulata y productividad de cinco genotipos de mora (Rubus glaucus Benth.). Acta Agron. 62(2): 174–185.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517.

Marulanda, M., A.M. López, and M. Uribe. 2012. Molecular characterization of the Andean blackberry, Rubus glaucus, using SSR markers. Genet. Mol. Res. 11(1): 322–331. doi: 10.4238/2012.February.10.3.

MD. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11, 2010.

Mendoza, H, Bautista, G. (2002). Diseño Experimental. Universidad Nacional de Colombia, http://www.virtual.unal.edu.co/cursos/ciencias/2000352/. Licencia: Creative Commons BY-NC-ND.

Michael Love, Simon Anders, Wolfgang Huber. 2019. DESeq2: Differential gene expression analysis based on the negative binomial distribution. Refrence Manual. http://www.bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf

Ramírez, M., M. Marulanda, and L. Isaza. 2007. Identificación de la especie de Colletotrichum responsable de la antracnosis en la mora de Castilla en la región cafetera. Systematics (37): 585–590.

Robinson, Mark D.; McCarthy, Davis J; Smyth, Gordon K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. En Bioinformatics, vol. 26 no 1, 2010. p. 139-140.

Saint-Oyant, L.H., T. Ruttink, L. Hamama, I. Kirov, D. Lakhwani, N.N. Zhou, P.M. Bourke, N. Daccord, L. Leus, D. Schulz, H. Van De Geest, T. Hesselink, K. Van Laere, K. Debray, S. Balzergue, T. Thouroude, A. Chastellier, J. Jeauffre, L. Voisine, S. Gaillard, T.J.A. Borm, P. Arens, R.E. Voorrips, C. Maliepaard, E. Neu, M. Linde, M.C. Le Paslier, A. Bérard, R. Bounon, J. Clotault, N. Choisne, H. Quesneville, K. Kawamura, S. Aubourg, S. Sakr, M.J.M. Smulders, E. Schijlen, E. Bucher, T. Debener, J. De Riek, and F. Foucher. 2018. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. Nat. Plants. doi: 10.1038/s41477-018-0166-1.

Salzman, J., H. Jiang, W. Wong. 2011. Statistical modeling of RNA- Seq data. Statistical Science 226: 62–83.

Sánchez Sara. Análisis de datos RNA-seq: comparación de métodos para el estudio de expresión genética diferencial. https://idus.us.es/xmlui/bitstream/handle/11441/40809/Sánchez%20Santana%20Sara%20del%20Carmen%20TFG.pdf?sequence=1

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22, 839–851.

Srivastava S, Chen L. 2010. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res 38: e170.

S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in RNA-seq: A matter of depth. Genome Research, 21: 2213 - 2223, 2011.

Tarazona Sonia. 2014. Statistical methods for transcriptomics: from microarrays to RNA-seq. https://riunet.upv.es/bitstream/handle/10251/48485/Tarazona%20-%20Statistical%20methods%20for%20transcriptomics%3A%20From%20microarrays%20to%20RNA-seq.pdf?sequence=1

Tarazona Sonia, Pedro Furió, María José Nueda, Alberto Ferrer, Ana Conesa. 2016. NOISeq: Differential Expression in RNA-seq. User´s guide. http://www.bioconductor.org/packages/release/bioc/vignettes/NOISeq/inst/doc/NOISeq.pdf

Trapnell, C., L. Pachter, and S.L. Salzberg. 2009. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. doi: 10.1093/bioinformatics/btp120.

VanBuren, R., C.M. Wai, M. Colle, J. Wang, S. Sullivan, J.M. Bushakra, I. Liachko, K.J. Vining, M. Dossett, C.E. Finn, R. Jibran, D. Chagné, K. Childs, P.P. Edger, T.C. Mockler, and N. V Bassil. 2018. A near complete, chromosome-

scale assembly of the black raspberry (Rubus occidentalis) genome. Gigascience. doi: 10.1093/gigascience/giy094.

VanBuren, R., D. Bryant, J.M. Bushakra, K.J. Vining, P.P. Edger, E.R. Rowley, H.D. Priest, T.P. Michael, E. Lyons, S.A. Filichkin, M. Dossett, C.E. Finn, N. V. Bassil, and T.C. Mockler. 2016. The genome of black raspberry (Rubus occidentalis). Plant J. doi: 10.1111/tpj.13215.

Y. Benjamini, and Y. Hochberg (1995). Controlling the False Discovery Rate: a Practical an Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B, 57, 289-300.

Ye, J., Y. Zhang, H. Cui, J. Liu, Y. Wu, Y. Cheng, H. Xu, X. Huang, S. Li, A. Zhou, X. Zhang, L. Bolund, Q. Chen, J. Wang, H. Yang, L. Fang, and C. Shi. 2018. WEGO 2.0: A web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Res. doi: 10.1093/nar/gky400.

Yuwen Liu, Jie Zhou, Kevin P. White; RNA-seq differential expression studies: more sequence or more replication?, Bioinformatics, Volume 30, Issue 3, 1 February 2014, Pages 301–304.

Yunshun Chen, Davis McCarthy, Matthew Ritchie, Mark Robinson, Gordon Smyth. 2008. EdgeR: differential expression analysis of digital gene expression data. User's Guide. http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Zheng, D., and G. Hrazdina. 2010. Cloning and characterization of an expansin gene, RiEXP1, and a 1-aminocyclopropane-1-carboxylic acid synthase gene, RiACS1 in ripening fruit of raspberry (Rubus idaeus L.). Plant Sci. doi: 10.1016/j.plantsci.2010.04.001.

11. ANEXOS

A continuación se muestra el código R utilizado para los paquetes EdgeR, NOI-seq y DESeq2 en la sección 7.1.

EdgeR

```
# Importar las tablas de conteo a R por cada condición

 > library(readr)

> Snf2<-
read_delim("~/Desktop/Snf2_countdata3/Snf2_rep48_MID55_allLanes_tophat2.0.5
.bam.gbgout","\t", escape_double = FALSE, col_names = FALSE, trim_ws =
TRUE)

> WT <-
read_delim("~/Desktop/WT_countdata/WT_rep48_MID85_allLanes_tophat2.0.5.ba
m.gbgout", "\t", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE)


 # Crear una sóla tabla de conteos con las dos condiciones

> junta<-merge(Snf2,WT, by.x="X1", by.y="X1")
```

```
> colnames(junta)<-c("gene", "h", "WT")


# Filtrado de los datos

> col_ordering = c(2,3)

> rnaseqMatrix = junta[,col_ordering]

> rnaseqMatrix = round(rnaseqMatrix)

> rnaseqMatrix = rnaseqMatrix[rowSums(cpm(rnaseqMatrix) > 1) >= 2,]


# Creación del Objeto DGEList

> conditions = factor(c(rep("h", 1), rep("WT", 1)))

> exp_study = DGEList(counts=rnaseqMatrix, group=conditions)


# Normalización y pruebas

> exp_study = calcNormFactors(exp_study)

> et = exactTest(exp_study, pair=c("h", "WT"), dispersion=0.2)

> tTags = topTags(et,n=NULL)

> result_table = tTags$table

> result_table = data.frame(sampleA="h", sampleB="WT", result_table)

> result_table$logFC = -1 * result_table$logFC
```

Para unir los resultados con los nombres de los genes y seleccionar los genes diferencialmente expresados

> junta2<-data.frame(row.names(junta),junta)

> reasult_table2<-data.frame(row.names(result_table),result_table)

> fin<-
merge(junta2,reasult_table2,by.x="row.names.junta.",by.y="row.names.result_table.",all.y=T)

> pos<-fin[which(fin$FDR<0.1),]

Importar los resultados originales del experimento que presenta los genes expresados diferencialmente.

genes<-read_delim("~/Desktop/genescuffdiff_full.txt", "\t", escape_double =FALSE, trim_ws = TRUE)

dif<-genes[which(genes$significant=="yes"),]

Contrastar los genes diferencialmente expresados con cada réplica con los resultados originales.

> fin2<-merge(pos,dif,by.x="gene", by.y="gene_id", all.x=T)

NOIseq

Importar las tablas de conteo a R por cada condición

```
 > library(readr)

> Snf2<-
read_delim("~/Desktop/Snf2_countdata3/Snf2_rep48_MID55_allLanes_tophat2.0.5
.bam.gbgout","\t", escape_double = FALSE, col_names = FALSE, trim_ws =
TRUE)

> WT <-
read_delim("~/Desktop/WT_countdata/WT_rep48_MID85_allLanes_tophat2.0.5.ba
m.gbgout", "\t", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE)


 # Crear una sóla tabla de conteos con las dos condiciones

> junta<-merge(Snf2,WT, by.x="X1", by.y="X1")

> colnames(junta)<-c("gene","A", "B")

> readcount<-data.frame(junta[,c(2,3)])

> colData <- data.frame(Conditions=names(readcount))

> row.names(colData)<-c("A","B")

> len<-matrix(c("A","B",7131,7131),2)

> len<-data.frame(len)


# Análisis de expresión diferencial

> library(NOISeq)

> mydata<-readData(readcount, factors=colData)

> mydata<-addData(mydata)
```

```
> myresults <- noiseq(mydata,factor="Conditions", replicates = "no")

> myresults@results[[1]]

> myresult1<-degenes(myresults,q=0.95)


# Para unir los resultados con los nombres de los genes

> junta2<-data.frame(row.names(junta),junta)

> reasult_table2<-data.frame(row.names(myresult1),myresult1)

> fin<-
merge(junta2,reasult_table2,by.x="row.names.junta.",by.y="row.names.myresult1."
,all.y=T)


# Contrastar los genes diferencialmente expresados con cada réplica con los
resultados originales.

> fin2<-merge(fin,dif,by.x="gene", by.y="gene_id", all.x=T)


# Excluyendo los genes falsos positivos.

> fin3<-fin11[which(fin11$significant=="yes"),]
```

DESeq2

```
# Importar las tablas de conteo a R por cada condición
```

```
 > library(readr)

> Snf2<-
read_delim("~/Desktop/Snf2_countdata3/Snf2_rep48_MID55_allLanes_tophat2.0.5
.bam.gbgout","\t", escape_double = FALSE, col_names = FALSE, trim_ws =
TRUE)

> WT <-
read_delim("~/Desktop/WT_countdata/WT_rep48_MID85_allLanes_tophat2.0.5.ba
m.gbgout", "\t", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE)


# Crear una sóla tabla de conteos con las dos condiciones

junta<-merge(Snf2,WT, by.x="X1", by.y="X1")

colnames(junta)<-c("gene", "A", "B")


# Análisis de expresión diferencial

> library(DESeq2)

> colnames(junta)<-c("gene","A", "B")

> readcount<-data.frame(junta[,c(2,3)])

> colData <- data.frame(condition=factor(c("A","B")))

> row.names(colData)<-c("A","B")

> dds <- DESeqDataSetFromMatrix(countData=readcount,colData =
as.data.frame(colData), design=~condition)

> dds$condition<-relevel(dds$condition, "A")
```

```
> dds$condition<-droplevels(dds$condition)

>dds <- DESeq(dds)

> res <- results(dds)
```