UNIVERSIDAD NACIONAL DE COLOMBIA

# Análisis de metodologías estadísticas en RNA-seq, con aplicación a cáncer de pulmón.

**Paula Fernanda Castañeda Valderrama**

Universidad Nacional de Colombia
Facultad de ingenieria, Ingenieria de sistemas y computación
Bogotá, Colombia
2021

# Análisis de metodologías estadísticas en RNA-seq, con aplicación a cancer de pulmón.

**Paula Fernanda Castañeda Valderrama**

Trabajo de grado presentada(o) como requisito parcial para optar al título de:
**Magister en Bioinformática**

Director(a):
Ph.D. Estadístico Carlos Eduardo Alonso Malaver .

Línea de Investigación:
Bioinformática funcional y estructural

Universidad Nacional de Colombia
Facultad de ingeniería, Departamento de Ingeniría industrial y de Sistemas.
Ciudad, Colombia
2021

A mi familia:

"Bendito el hombre que confia en el señor y pone su confianza en él. Será como un árbol plantado junto al agua que extiende sus raíces hacia la corriente; no teme que llegue el calor y sus hojas están siempre verdes. En época de sequia no se angustia y nunca deja de dar fruto"

Jeremías 17: 7-8

# Agradecimientos

Le agradezco principalmente a Dios y a mi familia, sin ellos no hubiese sido posible terminar esta etapa profesional y personal tan importante. A mi director Carlos Eduardo, quien creyó siempre en mi y me apoyó y a la Profesora Liliana López Kleine quien me orientó en el desarrollo de esta investigación.

Pero sobre todo le doy gracias a mi motor, mi esposo y mi hija Maria Paz, ella, sin saberlo, me da la alegria y la fuerza que necesito para alcanzar mis metas.

# Resumen

**Análisis de metodologías estadísticas en RNA-seq, con aplicación a cáncer de pulmón.**

La identificación de genes expresados diferencialmente en un grupo de pacientes que padezcan una enfermedad, es un primer paso en el desarrollo de procedimientos para la detección temprana de dicha enfermedad en los pacientes control.

Bajo esta premisa, un proposito deseable de algunas metodologías desarrolladas, es identificar los genes diferencialmente expresados de un conjunto de pacientes con cancer de pulmón[19], enfermedad que cobra más de 150.000 vidas cada año en los Estados Unidos y cerca de 1.76 millones de muertes a nivel mundial[26].

Existen diferentes metodologías que permiten hallar los genes diferencialmente expresados, en la presente investigación se comparan tres de ellas Limma[22], Deseq2 [16] y Noiseq [10], haciendo uso de un conjuntos de datos simulados por medio del paquete de Bioconductor del software R de nombre compcodeR[27], que permite simular muestras de genes con características similares a un conjunto real, donde se conoce el porcentaje y el conjunto de genes que está diferencialmente expresado dentro de la matriz de conteos.
Se generan tres matrices de conteos con 20 y 30 muestras, 30000 genes, 300 y 500 genes diferencialmente expresados, para las tres matrices se ejecutan las tres metodologías mencionadas, encontrando que es Deseq2 el método que encuentra el mayor número real de genes diferencialemente expresados y tiene un mejor porcentaje de Recall.

Finalmente se aplica el método seleccionado, Deseq2, a las muestras recolectadas de pacientes con cancer de pulmón, análisis de muestras de pacientes, de tejido epitelial pulmonar normal y tumores pulmonares de carcinoma de células escamosas[19]. La aplicación encuentra 7506 genés diferencialmente expresados.

**Palabras clave: Cáncer de pulmón, Noiseq, Deseq2, Limma, RNA-seq, Bioconductor, Bioinformática.**

# Abstract

**Analysis of statistical methodologies in RNA-seq, with application to lung cancer**

The identification of differentially expressed genes in a group of patients who have a disease is a first step in the procedures development for the early detection of the disease in control patients.

Under this premise, a desirable purpose of some developed methodologies is to identify the differentially expressed genes of a set of patients with lung cancer[18], the disease claims more than 150.000 lives each other in the United States and almost of 1.76 million deaths in the world.

There are different methodologies that allow to finding the differentially expressed genes.
In this investigation, Limma [22], Deseq2 [16] and Noiseq [10] methodologies are compared, using a set of data simulated by means of the Bioconductor package of the R software called compcodeR citeAG49, which allows to simulate gene samples with similar characteristics to a real set, where the percentage and the set of diferentially expressed genes in the count matrix are known.
Three counting matrices are generated with 20 and 30 samples, 30,000 genes, 300 and 500 differentially expressed genes, for the three matrices the three methodologies are executed, finding that Deseq2 is the method that finds the highest actual number of differentially expressed genes and has a better percentage of Recall.

Finally, Deseq2 was apply into the clinical samples collected from patients with lung squamous cell carcinoma[19]. We found 7506 differentially expressed genes.

**Keywords: Lung cancer, Noiseq, Deseq2, Limma, RNA-seq, Bioconductor, Bioinformatics**

.

Este Trabajo Final de maestría fue calificado en mayo de 2022 por el siguiente evaluador:

Emiliano Barreto Hernández MSc.
Profesor Instituto de Biotecnología
Universidad Nacional de Colombia

# Contenido

# 1. Abreviaturas

| **Abreviatura** | **Término** |
|---|---|
| *RNA-seq* | Secuenciación del transcriptoma entero |
| *ARN* | Ácido Ribonucleico |
| *ADN* | Ácido desoxirribonucleico |
| *CPM* | Mínimo de conteos por millón |
| *TMM* | Media truncada de M-valores |
| *RPKM* | Lecturas por millón de kilobases |

# 2. Introducción

El cáncer de púlmon es una enfermedad que causa un alto número de muertes anuales en todo el mundo. Existen estudios que aseguran que cerca de 1.76 millones de muertes se dan por esta enfermedad cada año y que además son los hombres quienes tienen un porcentaje mayor de muertes por esta causa[26].

Lo que sucede normalmente es que los pacientes esperan exámenes médicos para conocer si presentan o no la enfermedad, sin embargo, existen otros métodos que podrían dar un dictamen preliminar sobre la posibilidad de llegar a padecer cáncer de púlmon.

Cobra sentido, entonces, hablar sobre metodologías de hallazgo de genes diferencialmente expresados. A lo largo de los años han existido diferentes métodos tales como los microarreglos[1] que fueron evolucionando e introduciendo pruebas $t$ con estimaciones agrupadas de la varianza de las muestras, luego hace una aparición la estadística bayesiana que calcula la distribución marginal, haciendo uso de la distribución a priori y a posteriori[9].
El análisis de RNA-seq (secuenciación del transcriptoma entero) fue introducido con el proposito de mejorar el método de microarreglos, que permite revelar la presencia y cantidad de ARN en muestras biológicas generando variables de conteos. EL calculo de RNA-seq hace uso de herramientas estadísticas que permiten determinar la significancia estadística de un conjunto de genes expresado diferencialmente.

Una de las caracterísicas mas relevantes del RNA-seq, es el análsis de calidad de las muestras recolectadas, a continuación se realiza el alineamiento y conteo de los genes a estudiar, finalmente se aplica una metodología de análisis de expresión diferencial de genes, tales como EdgeR, NOIseq, Deseq, Deseq2, Limma entre otros.[23]

Los métodos que se utilizan y comparan dentro de esta investigación son, Deseq2, Limma y Noiseq.

# 3. Antecedentes

El carcinoma de pulmón cobra más de 150.000 vidas cada año en los Estados Unidos y cerca de 1.76 millones de muertes a nivel mundial, superando así la mortalidad por cáncer de mama y próstata [26]. La clasificación actual del cáncer de pulmón se basa en características clínicas [3].

Basándose en las estadísticas presentadas a lo largo de los años tal como lo hace Siegel[24], dónde se observa que el 29 % de las muertes por cáncer de hombres son causadas por cáncer de pulmón y para las mujeres es del 26 % y que además su participación en nuevos casos mapeados es de 14 % y 13 % respectivamente; se encuentra la necesidad de aplicar una metodología genética que alerte la presencia de la enfermedad cobrando sentido la opción de hallar los genes diferencialmente expresados en un conjunto de datos de pacientes con cáncer de pulmón con el fin de concluir con el hallazgo temprano de la enfermedad[14].

En el 2002 se introduce la tecnología de microarreglos que mide la cantidad de RNA mensajero de todos los genes de un organismo (miles de genes) en un solo experimento. Sobre los datos de expresión de genes o transcriptómicos es de interés identificar genes diferencialmente expresados (individuos estadísticos), así como clasificar las muestras (variables). Para esto se realizan modelos de aprendizaje automático con el fin de clasificar las muestras de acuerdo al tipo de cáncer de pulmón según los tipos de histopatología y se comparan a partir del rendimiento por precisión en la clasificación de salida [1].

En el análisis de microarreglos se presenta el uso de pruebas t con estimaciones agrupadas de la varianza de la muestra, basada en genes expresados de manera similar para identificar los genes influyentes en la clasificación de las muestras (variables), sin embargo, estos métodos no muestran un comportamiento coherente en todo el intervalo de agrupación y pueden estar sesgados cuando se especifican los hiperparámetros anteriores de forma heurística. Es ahí donde hacen su aparición los procesos estadísticos bayesianos, que calculan explícitamente en forma analítica la distribución marginal para la diferencia en la expresión media de dos muestras, haciendo uso de la distribución a priori y el conocimiento previo de la varianza [9].

En los microarreglos se introduce como algoritmo de selección de características, el análisis de componentes principales que en algunas ocasiones también presenta modificaciones con el fin de mejorar su precisión [11, 12].

El análisis de RNA-seq o secuenciación de transcriptoma entero aparece como una mejora de los microarreglos, método que ha mostrado diferentes problemas con la diferenciación del transcriptoma, (falsos positivos). Esta metodología, RNA-seq, permite revelar la presencia y cantidad de ARN, en una muestra biológica en un momento dado al igual que los microarreglos. La diferencia con los microarreglos es que el RNA-seq cuantifica la cantidad de RNA mensajero de cada gen por secuenciación directa y no por hibridación conduciendo a una variable de conteos y no una variable continua. En la Figura **C-4**, tomada de [23], se evidencia el proceso a seguir para realizar RNA-seq,

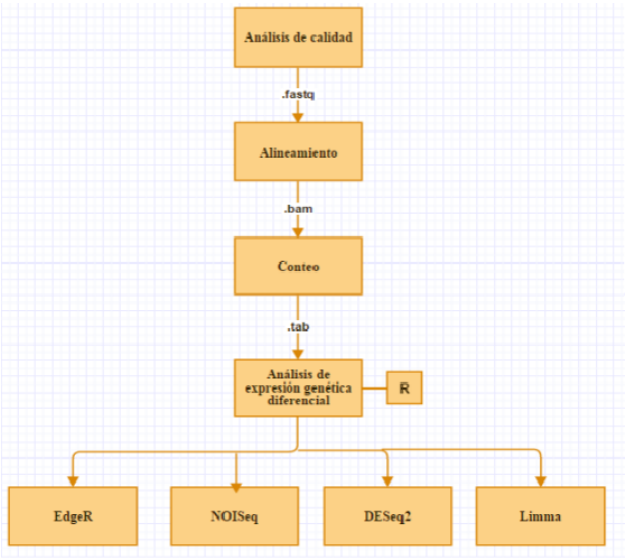



**Figura 3-1**.: Pasos a seguir en el análisis de datos de RNA-Seq para la identificación de genes diferencialmente expresados [23]

El proceso de RNA-seq se puede dividir aproximadamente en dos tipos:

- El primero basado en el genoma de referencia, donde existe un genoma ensamblado para una especie, para la cual se realiza un experimento RNA-seq. Permite que las lecturas se alineen contra el genoma de referencia y mejora significativamente la capacidad de reconstrucción de transcritos. Esta categoría incluye los humanos y la mayoría de los modelos de organismos, pero excluye la mayoría de las especies biológicamente interesantes [15].
- El segundo no tiene en cuenta el genoma de referencia, no se requiere ensamblaje del genoma para las especies de interés disponible. En este caso, se necesitaría ensamblar las lecturas en transcripciones utilizando enfoques de Novo. Para este tipo de RNA-seq el ensamblaje depende en gran medida de los parámetros y es más complejo que el anteriormente nombrado [15].

Entre los diferentes métodos para evaluar la expresión génica, la secuenciación de alto rendimiento del ARN o RNA-Seq es interesante teniendo en cuenta que durante el proceso de RNA-seq, el ARN aislado de muestras de interés se utiliza para generar una biblioteca de ADN, que luego es amplificada y secuenciada. En última instancia, RNA-seq puede determinar que genes se expresan, los niveles de su expresión y la presencia de cualquier transcrito anteriormente desconocido [29].

La secuenciación masiva de ARNs o RNA-seq, tiene un mayor alcance en comparación a los microarreglos y otros métodos, se tiene un mayor volumen de datos y esto genera un desafío bioinformático. Es posible afirmar también que los métodos de análisis no han evolucionado de la misma forma que esta técnica [18].

Inmersas a la secuenciación del transcriptoma se encuentran algunas metodologías estadísticas como: Redes Neuronales, Inducción del árbol de decisión, Máquinas de soporte vectoriales, Redes neuronales de avance, Modelos lineales, Modelos de regresión, métodos no paramétricos [17].

El proceso de RNA-seq ya mencionado contiene diferentes metodologías estadísticas que pueden variar si se tiene en cuenta el tamaño de muestra, los datos y otras consideraciones, estas metodologías son las que finalmente permiten tener precisión y rigurosidad en los resultados, de las que se presenta una breve explicación a continuación:

LIMMA: Metodología que proporciona una solución para el análisis de datos de expresión génica, se tenía originalmente para el proceso de microarreglos, sin embargo y dado el avance que han tenido los análisis genéticos, este método tuvo una mejora, en la cual se les realiza una transformación a los datos. Dicha transformación se basa en la estimación de la relación media varianza de los conteos en escala logarítmica para generar las ponderaciones a nivel de las observaciones para su posterior aplicación y estimación bayesiana. De esta forma el modelo lineal puede ejecutarse normalmente. [6]

Deseq2: Análisis de expresión de transcritos basado en modelos lineales generalizados para una distribución binomial negativa, filtra las transcripciones que cuenta mediante la media de conteos normalizados para cada gen y finalmente, utiliza la prueba de Wald para determinar genes diferencialmente expresados. [15]

NOIseq: Este método tiene enfoques no paramétricos para el análisis de expresión diferencial de datos de RNA-Seq. Se obtiene comparando las estadísticas de expresión diferencial, valores M-D, de esa transcripción o gen contra distribución de cambios en los valores de expresión al comparar réplicas dentro de la misma condición, medido a partir de la probabilidad de diferenciación. [23]

EdgeR: Análisis de expresión diferencial con réplicas biológicas, la metología estadística utilizada está basada en: Distribuciones Bionomiales Negativas, estimación empírica de Bayes, pruebas exactas, modelos lineales generalizados y pruebas de cuasi-verosimilitud. Este análisis se puede aplicar también para el análisis de señales diferencales de otros tipos de datos genómicos que producen los conteos de las lecturas.[7]

En la Figura **3-2** es posible visualizar las características y diferencias de las diferentes metodologías nombradas.

| | **EdgeR** | **NOISeq** | **DESeq2** | **Limma** |
|---|---|---|---|---|
| **Modelo probabilístico** | Binomial Negativa | Aproximaciones no paramétricas | Modelo de regresión binomial negativo | Modelos lineales |
| **Método de normalización** | TMM | RPKM | Normalisation | TMM |
| **Filtrado** | CPM | CPM | Basado en la media de los datos normalizados para cada tránscrito | CPM |
| **Test** | Test exacto basado en la distribución Binomial Negativa | Estimación de la probabilidad de expresión diferencial | Test paramétrico de Wald | Prueba t-Student |

**Figura 3-2**.: Cuadro comparativo métodologías RNA-Seq [23]

Luego de la aplicación del RNA-seq se observa que, aunque la tecnología es cada vez más utilizada y ha ido evolucionando, no se encuentra una metodología estándar para realizar el análisis de los datos. Es posible ver en diversos estudios previos tales como [6] [23] y [31] como unas metodologías se superan a otras por tiempo, capacidad, rendimiento, precisión, entre otros .

Uno de los principales problemas encontrados es que para la realización de los modelos estadísticos aplicados al RNA-seq los datos deben tener distribuciones continuas. Sin embargo, dado que los datos que se obtienen en RNA-Seq están relacionados con la abundancia de transcritos de RNA, son datos discretos [31].

# 4. Plantenamiento del Problema

Bajo la necesidad de analizar los transcritos de un grupo de pacientes con cáncer de pulmón, se encuentran diferentes metodologías [6] todas ellas con algún error acerca de falsos positivos. El problema es reconocer cuáles de estos métodos son los que llevarán al menor número de falsos positivos y la mejor precisión en el hallazgo de la diferenciación del transcriptoma en el carcinoma pulmonar, debido a que si el problema existe se puede detectar la presencia errónea de la enfermedad en los pacientes [13]. Además, dado que todo el proceso está realizado alrededor del RNA-seq, bajo este análisis los métodos estadísticos varían y es necesario conocer muy bien los datos y el proceso de secuenciación del transcriptoma entero antes de tomar una decisión acerca del método estadístico inmerso a elegir[31].

Para el estudio se seleccionan tres metodologías que se describen en [23] teniendo en cuenta la viabilidad del trabajo a realizar, dicho esto la pregunta de investigación que se formula es: Del análisis y comparación de las metodologías Deseq2, Limma y NOIseq que realizan RNA-seq aplicadas al conjunto de datos simulados. ¿Cuál es la metodología que tiene mayor precisión en el momento de hallar los transcritos diferencialmente expresados?

# 5. Justificación

El RNA-seq como una tecnología de secuenciación de próxima generación se está utilizando cada vez más para la creación de perfiles de expresión génica como reemplazo de los microarreglos. Sin embargo, las propiedades de los datos de RNA-seq aún no se han establecido completamente [13]. Se sabe que no provienen de una distribución normal y esto debe corregirse en el análisis de datos.

Se conoce que es importante encontrar una metodología estándar que lleve a la menor cantidad de falsos positivos de expresión debido a que tener un falso positivo significaría decir que un gen está diferencialmente expresado y que realmente no lo esté, lo cual puede llevar a la detección de alguna enfermedad de manera temprana o errónea. Es importante estar seguros de su eficiencia y precisión, realizando una revisión de sensibilidad de los métodos a características tales como el tamaño de muestra, tamaño del efecto, poder, entre otros. Por ejemplo, para evitar supuestos en la distribución de los datos y disminuir la tasa de falsos positivos se pueden realizar transformaciones o emplear programas no paramétricos que no asumen una distribución establecida para los datos [21].

Bajo la premisa de que existen diversas metodologías y modelos a aplicar es importante revisar los resultados que se obtienen de la ejecución, de forma tal que sea posible lanzar con cierta certeza la mejor opción de modelado estadístico para que cualquier investigador realice la toma de decisiones acertadas según sea el caso [31].

Por este motivo, en este trabajo se comparan tres de las metodologías más utilizadas para el proceso que se ha venido describiendo, realizando el análisis de RNA-seq en un conjunto de datos simulados, aplicandolos en datos recolectados de pacientes con cáncer de pulmón.

# 6. Objetivos

## 6.1. Objetivo General

Determinar la metodología estadística que mejor halle los transcritos diferencialmente expresados de un conjunto de datos de pacientes con cáncer de pulmón por medio del proceso de RNA-seq, teniendo en cuenta la comparación de tres paquetes de Bioconductor en la plataforma R.

## 6.2. Objetivos específicos

- Aplicar las metodologías Deseq2, Limma y NOIseq a un conjunto de datos simulados, por medio del paquete Bioconductor en la plataforma R.
- Analizar los resultados obtenidos de cada una de las metodologías, teniendo en cuenta tamaño de muestra, tamaño del efecto, poder, capacidad, rendimiento y precisión.
- Comparar las metodologías aplicadas basándose en los análisis realizados, con el fin de proceder a la toma de decisiones acerca de la metodología de mejor performance aplicándola al conjunto de datos de pacientes con cáncer de pulmón.

# 7. Metodología

Para alcanzar los objetivos presentados, se realiza una investigación descriptiva [8], en la que se prueban y aplican tres teorías objetivas existentes por medio de relación de variables, que finalmente se miden y analizan haciendo uso de procedimientos estadísticos, además este estudio se aplica en primera instancia a un conjunto de datos simulados, luego la metodología seleccionada se aplica a una población real, lo que también asegura el tipo de investigación descriptiva, [4].

Teniendo en cuenta que, "La investigación cuantitativa es aquella donde se recogen y analizan datos cuantitativos" [5] y dado que tanto los datos simulados como los datos recolectados corresponden a conteos de RNA y son utilizados para aplicar métodos cuantitativos y finalmente analizar los resultados, el enfoque es cuantitativo.

El diseño que se realiza es no experimental, se genera un conjunto de datos simulados que cumple con las características necesarias para efectuar análisis de RNA-seq conociendo a priori el número de genes que están diferencialmente expresados, finalmente y bajo diferentes métricas se escoge una de las metodologÃas que se aplica al conjunto real de genes, los cuales son muestras clínicas de pacientes con carcinoma de células escamosas de pulmón, 9 muestras de tumor y 9 de tejido adyacente[3]. La base de datos contiene los conteos de cada uno de los genes, en cada una de las muestras.
El número total de génes es de 26.363.
El trabajo se realiza en tres fases que se tienen en cuenta para el desarrollo de los objetivos de este trabajo:

- **Fase 1:** En esta fase se realiza la revisión de literatura que permite la creación de la introducción, antecedentes y marco conceptual de los términos necesarios para el trabajo.

- **Fase 2:** Se construye el conjunto de datos simulados que proviene del paquete en R compcodeR [28], dicho paquete proporciona una amplia funcionalidad para comparar los resultados obtenidos por diferentes métodos, con el fin de realizar análisis de expresión diferencial de datos de RNAseq, también contiene funciones para simular datos de conteo.

Finalmente, proporciona interfaces convenientes a varios paquetes para realizar el análisis de expresiones diferenciales. Los parámetros de la función de simulación de datos (generateSyntheticData) pueden ser manipulados por el usuario de manera que se pueda configurar y ejecutar un análisis diferencial definido por el mismo. En la Tabla 7-1 se presentan los parámetros mencionados:

**Tabla 7-1**.: Parámetros función de simulación de datos [27]

| **Parámetro** | **Funcionalidad** |
|---|---|
| dataset | Nombre de la matriz de datos a generar. |
| n.vars | Tamaño de la matriz de genes. |
| sample.per.count | Número de muestras. |
| n.diffexp | Número de genes diferencialmente expresados. |
| seqdepth | La profundidad de secuenciación base (número total de lecturas asignadas). Este número se multiplica por un valor dibujado uniformemente entre minfact y maxfact para cada muestra para generar datos con diferentes profundidades de secuenciación real. |
| min. fact and max.fact | El mínimo y máximo para la distribución uniforme utilizada para generar factores que se multiplican con seqdepth para generar profundidades de secuenciación individuales para las muestras simuladas. |
| relmeans | Vector de medias de las muestras, en caso de ser auto los valores serán estimados a partir de conjuntos de datos de RNA-seq. |
| dispersion | Vector de dispersiones de las muestras, en caso de ser auto los valores serán estimados a partir de conjuntos de datos de RNA-seq. |
| fraction.upregulated | Fracción de genes diferencialmente expresados que están sobre expresados en la condición 1 comparado con la condición |
| between.group.diffdisp | Admitir que la dispersión sea diferente entre muestras. |
| filter.threshold.total | Umbral de filtro en el recuento total de un gen en todas las muestras. |
| effect.size | La fuerza de la expresión diferencial, es decir, el tamaño del efecto, entre las dos condiciones. |

El uso específico de este paquete se puede visualizar en el Anexo A. Para la compa-

ración de los conjuntos de muestras se cambia el número de muestras y el número de genes diferencialmente expresado, manteniendo el número total de genes y las demás características, como profundidad de secuenciación y otros.

Los parámetros que se modificaron se pueden visualizar en la tabla 7-2.

**Tabla 7-2**.: Muestras generadas

| **Muestra** | **Parámetros** |
|---|---|
| 1 | 30.000 genes, 20 muestras, 500 genes diferencialmente expresados, no se cambian los demás parámetros. |
| 2 | 30.000 genes, 30 muestras, 500 genes diferencialmente expresados, no se cambian los demás parámetros. |
| 3 | 30.000 genes, 30 muestras, 300 genes diferencialmente expresados, no se cambian los demás parámetros. |

- **Fase 3:** En esta fase se realiza el análisis de los resultados obtenidos haciendo uso de algunas medidas de desempeño:
  1. Número de génes diferencialmente expresado
  2. Falsos Positivos
  3. Verdaderos Positivos
  4. Precisión
  5. Recall

  Las deficiones de dichas medidas se extienden en la sección del Marco conceptual. Una vez realizada la comparación entre los resultados se determina la metodología que mejor encuentra los transcritos diferencialmente expresados y se aplica a los datos de los pacientes con cáncer de pulmón obtenidos de NCBI.

# 8. Marco Conceptual

## 8.1. Metodologías Aplicadas

### 8.1.1. Limma

Limma es un paquete de software de R que permite realizar análisis de datos de expresión génica. Estadísticamente limma opera con una matriz de valores de expresión, ajustando un modelo lineal de la forma:

$$Y_i - \beta_0 + \beta_1 x_{i1} + ... + \beta_1 x_{ip} + \epsilon_i \tag{8-1}$$

a cada fila de datos, Limma proporciona la capacidad de analizar comparaciones entre muchas respuestas de RNA simultáneamente, esta metodología utiliza el modelo lineal y las funciones de expresión diferencial que son aplicables a los datos de cualquier tecnología de expresión génica cuantitativa que incluye microarrays, RNA-seq y PCR.[25]
En la Figura **8-1** se visualiza una representación del modelo lineal y principios estadísticos empleados en limma.
Los métodos anteriores se aplicarán sobre bases de datos simuladas con el propósito de identidificar el método que mejor halla los genes diferencialmente expresados, con el fin de aplicarlo al conjunto de datos de pacientes con cancer de pulmón.

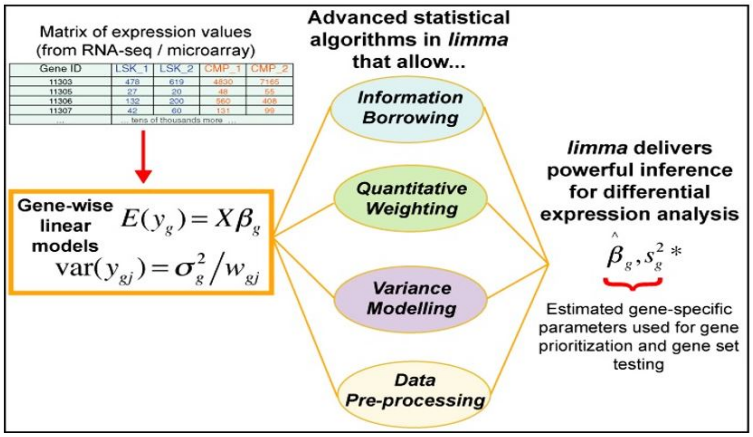



**Figura 8-1**.: Representación del modelo lineal y principios estadísticos empleados en limma. Tomado de [20]

### 8.1.2. Noiseq

Noiseq es otro de los paquetes pertenecientes a Bioconductor que realiza análisis exploratorio y diferencial de datos de RNA-seq, este proceso lo realiza teniendo en cuenta un enfoque no paramétrico, que crea una distribución nula o de ruido de los cambios de conteo comparando el número de lecturas de cada gen en muestras dentro de la misma condición. Esta distribución de referencia se utiliza luego para evaluar si es probable que el cambio en el número de conteos entre dos condiciones para un gen dado sea parte del ruido o representa una expresión diferencial verdadera. Se implementan dos variantes del método: Noiseqreal usa réplicas, cuando están disponibles, para calcular la distribución de ruido y Noiseq-sim simula en ausencia de réplicas[22].

### 8.1.3. Deseq2

Deseq2 es otro método del paquete Bio-conductor que por medio de modelos lineales generalizados (GLM) realiza análisis diferencial de datos de RNA-seq. El punto de partida de un análisis DESeq2 es una matriz de recuento $K$ con una fila para cada gen $i$ y una columna para cada muestra $j$. Las entradas de la matriz $K_{ij}$ indican el número de lecturas de secuenciación que se han asignado a un gen en una muestra. Para cada gen, se ajusta un modelo lineal generalizado (GLM)[16][30]:

$$\begin{cases} y_{tij} \sim BinomialNegativa(\beta, \gamma). \\ log(E(y_{tij})) = \alpha_t + \tau_{it} + \epsilon_{ij} \\ \alpha_t \sim N(0, G) \end{cases} , \tag{8-2}$$

$\tau_{it} \sim N(\tau, \sigma_t^2)$ con $i = 1, 2, 3, 4, 5, .., n$ que son las muestras, $\epsilon_{itj} \sim N(0, \sigma^2)$ es el error del modelo que se asume independiente e idénticamente distribuido.

Se modelan los conteos $K_{ij}$ bajo una distribución binomial negativa de media $\mu_{ij}$ y dispersión $\sigma_i$ la media se deja como una cantidad $q_{ij}$, proporcional a la concentración de fragmentos de ADN del gen en la muestra, escalada por un factor de normalización $s_{ij}$, quiere decir que la media $\mu_{ij}$ será igual al producto de de la concentración de fragmentos $q_{ij}$ y el factor de normalización $s_{ij}$, $\mu_{ij} = q_{ij}s_{ij}$[16]

El modelo lineal generalizado a utilizar, tiene función de enlace logarítmica, la cual se presenta a continuación:

$$log\ q_{ij} = \sum_{r} x_{ij}\beta_{ir} \tag{8-3}$$

con una matriz de diseño de elemento $x_{jr}$ y coeficientes $\beta_{ir}$

En el caso en el que se comparan sólo dos grupos, como muestras caso y control, los elementos de la matriz de diseño indican si una muestra $j$ son casos o no, y el ajuste del modelo lineal generalizado tiene coeficientes que indican la fuerza de expresión general del gen y el $log_2$ fold change entre el caso y el control [10].

## 8.2. Medidas de desempeño

Para tomar decisiones en cuanto a la comparación de los métodos, se tendrán en cuenta algunas medidas de desempeño que se obtienen de la Matriz de confusión (Figura **8-2**):

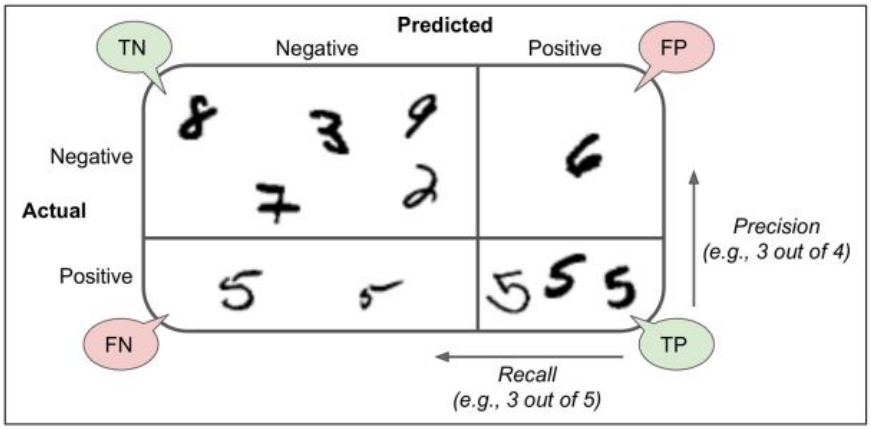



**Figura 8-2**.: Matriz de confusión [2].

La idea general es contar las veces que las predicciones del modelo se realizan bien o mal, teniendo en cuenta el conjunto real de datos. Las filas de la matriz representan la clase real y las columnas las clases predichas.

- **Número de génes diferencialmente expresados**
- **Falsos Positivos**, que corresponde al número de genes que los modelos predicen que están diferencialmente expresados pero en realidad no lo están. **FP**

- **Verdaderos Positivos**, que corresponde al número de genes que los modelos predicen que están diferenciamente expresados y en realidad lo están. **TP**
- **Precisión**, que corresponde a la calidad, quiere decir, porcentaje de génes diferencialmente expresados que realmente lo están.
- **Recall**, que correponde al porcentaje de génes diferencialmente expresados que los modelos son capaces de identificar.

# 9. Discusión y Resultados

### 9.0.1. Simulación

De los tres conjuntos de muestras simuladas por medio del código que se puede ver dentro del Anexo A, presentadas en la metodología (Tabla 7-2: Muestras generadas), se realizan 50 iteraciones por muestra con el fin de obtener los resultados más generales posibles, se habla de generales ya que se quiere visualizar los cambios que pueden haber entre cada iteración.

Por un lado las muestras presentan correlaciones cercanas al 0.70, la mínima correlación encontrada fue del 0.64 y se da dentro de la muestra 1, además se ven dispersas lo que asegura que los datos generalizan bien el comportamiento de genes reales donde las correlaciones se pueden ver cercanas a 1 entre las muestras y son dispersos [15]. En la Figura **9-1** se visualizan los hallazgos mencionados.

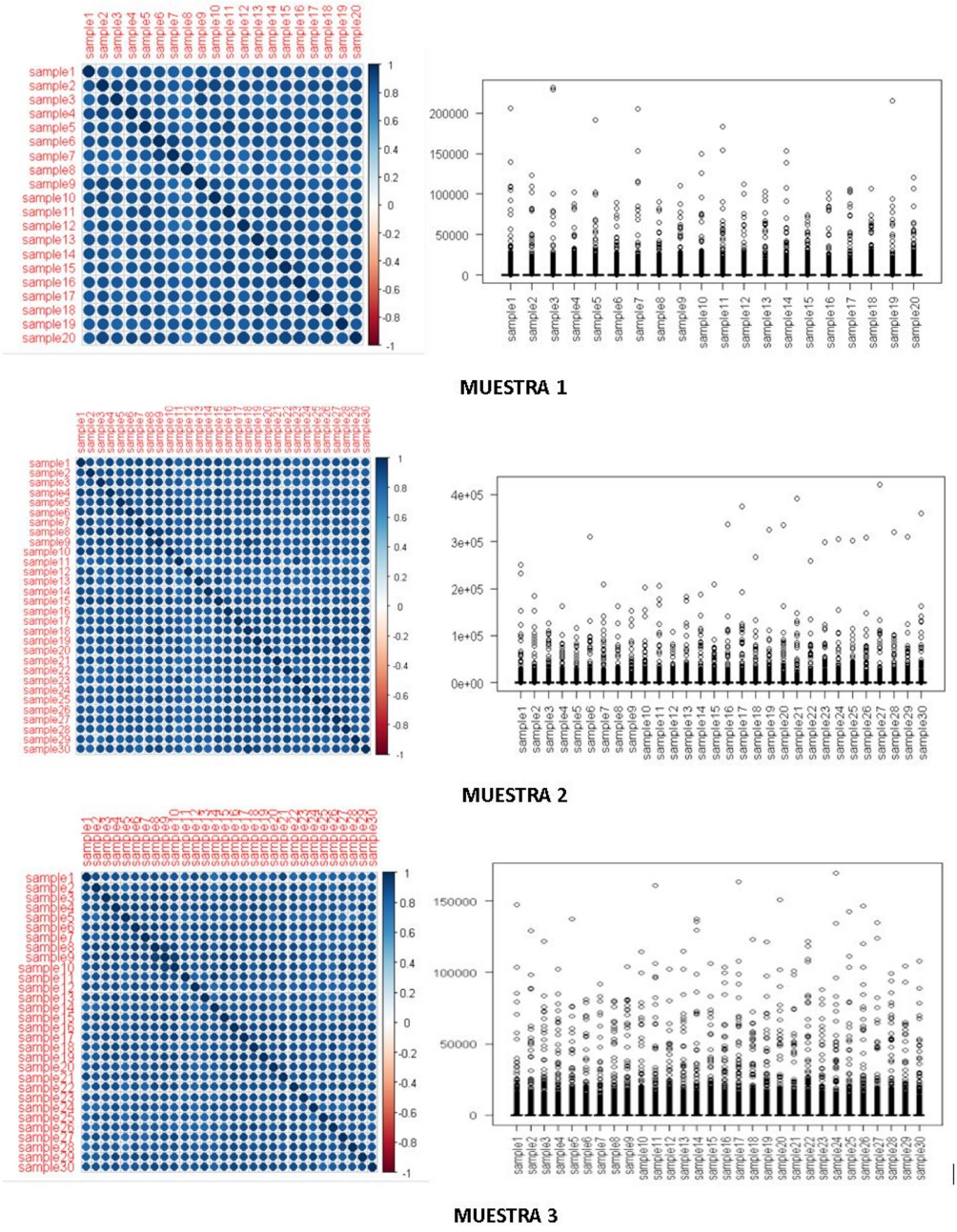



**Figura 9-1**.: Representación de las correlaciones y dispersión de la muestra 1, 2 y 3

Es importante la revisión de los conteos antes de realizar cualquier tipo de transformación. Existen varios métodos para dicha revisión, en este caso se realizan gráficos de dispersión y correlación de manera que se reconozca si exiten muchos genes con conteos en 0 o correlaciones muy bajas. Se espera que las muestras estén correlacionadas.

Con cada una de las matrices de datos simulados se ejecutan las tres metodologías mencionadas dentro del marco conceptual, de este proceso se obtienen los siguientes resultados:

### Genes diferencialmente expresados

Al realizar la comparación de los modelos, los resultados de mayor relevancia están basados en el número de génes diferencialmente expresados que cada una de los modelos encuentra dentro de cada una de las muestras.

A continuación, se presentan una serie de gráficos que permiten análizar y tomar decisiones en cuanto a las metodologías aplicadas, basados en el Número de génes diferencialmente expresados, Falsos Positivos, Verdaderos Positivos, Precisión y Recall.

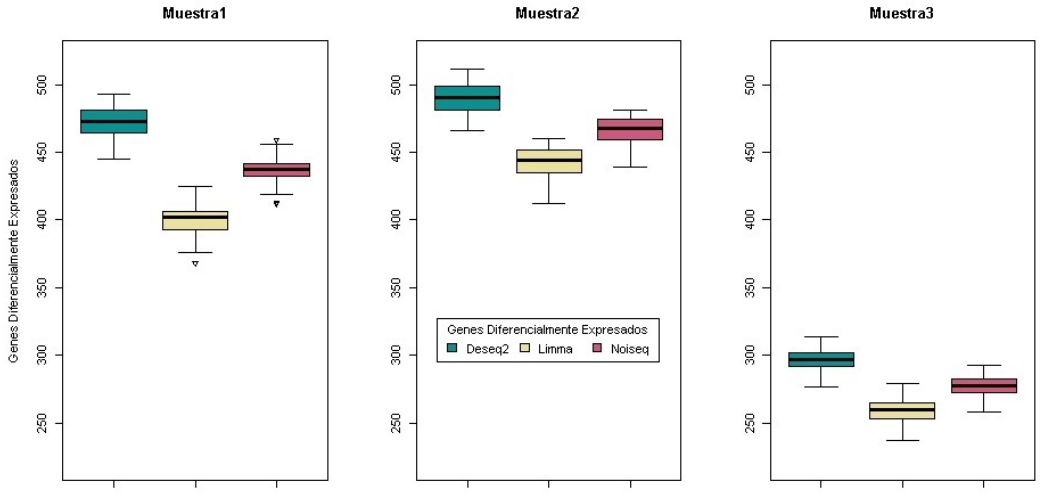



**Figura 9-2**.: Los boxplot representan el número de **genes diferencialmente expresados** por cada una de las muestras, por lo tanto lo que se visualiza es la distribución de los resultados del uso de las metodologías Deseq2, Limma y Noiseq en cada una de las muestras

En la Figura **9-2** se observa que Deseq2 es el método que encuentra el mayor número de genes diferencialmente expresado, es en la muestra dos dónde el número de genes encontrados aumenta y los resultados no presentan ningún dato atípico en comparación con la muestra1, dónde tanto en Limma como en Noiseq se evidencia que una de las simulaciones encontró un número menor de genes al esperado. Para la muestra tres no se observan resultados atípicos y la distribución parece estar menos sesgada que para las demás muestras.

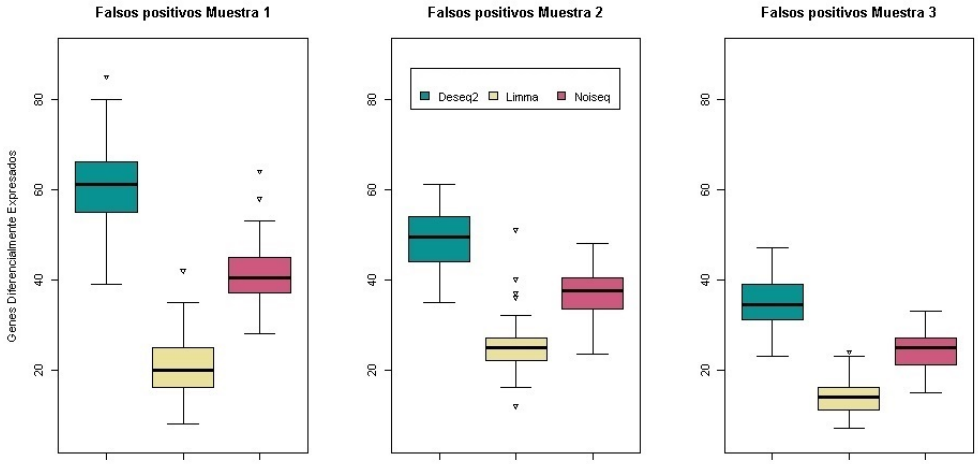



**Figura 9-3**.: **Falsos Positivos.** Los boxplot representan el número de genes que cada modelo en cada simulación predice como diferencialmente expresados y que comparando con el conjunto de genes simulado no coinciden.

El problema de los falsos positivos está en la predicción de un gen que se muestra como diferencialmente expresado, pero que en la realidad no lo está. Dentro de la Figura **9-3** se puede visualizar que para todas las muestras Limma presentan datos atípicos, en algunas de sus simulaciones muestra un número mayor de falsos positivos del esperado. En la muestra tres se observa un mejor comportamiento dentro de cada modelo y entre los modelos.

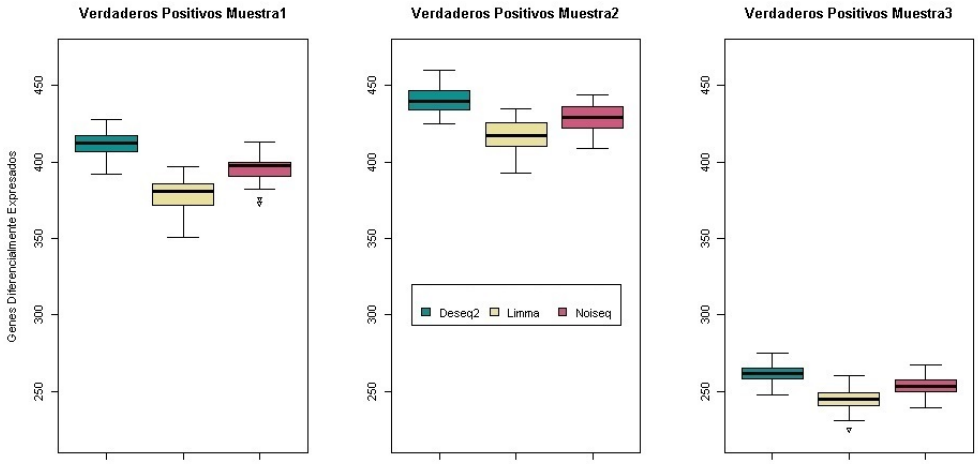



**Figura 9-4**.: **Verdaderos Positivos** Los boxplot representan el número de genes que cada modelo en cada simulación predice como diferencialmente expresados y que comparando con el conjunto de genes simulado coinciden.

El hallazgo real de los genes es de suma importancia, por eso la revisión del conjunto de genes predicho contra el conjunto de genes real, en la Figura **9-4** se observa un comportamiento similar entre los modelos, sin embargo, es Deseq2 el método que encuentra el mayor número de génes reales para todas las muestras y es Limma, el método que menor número de genes reales halla.

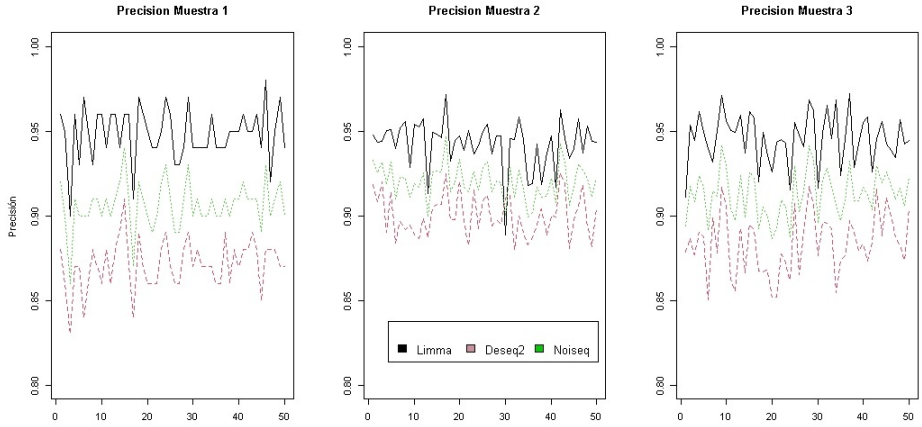



**Figura 9-5**.: **Precisión** Los gráficos representan, la precisión de cada modelo dentro de cada muestra y cada simulación.

La precisión, mide el porcentaje de casos positivos detectados, en otras palabras, mide el porcentaje de genes diferencialmente expresados reales detectados. Se observa en la Figura **9-5** que para esta medida, la metodología Limma tiene una mayor precisión dentro de todas las muestras.

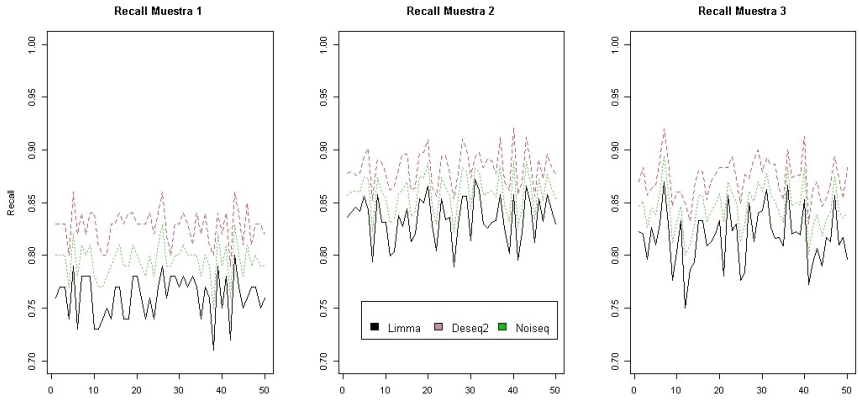



**Figura 9-6**.: **Recall** Los gráficos representan, la sensibilidad(Recall) de cada modelo dentro de cada muestra y cada simulación.

El Recall o sensibilidad mide la capacidad de poder detectar correctamente los genes diferencialmente expresados entre el conjunto real de los genes con esta característica. Se reconoce que este es uno de los resultados mas importante, dada la necesidad del estudio. En la Figura **9-6** se observa que el modelo con el porcentaje de Recall mas alto en todas las muestras es Deseq2.

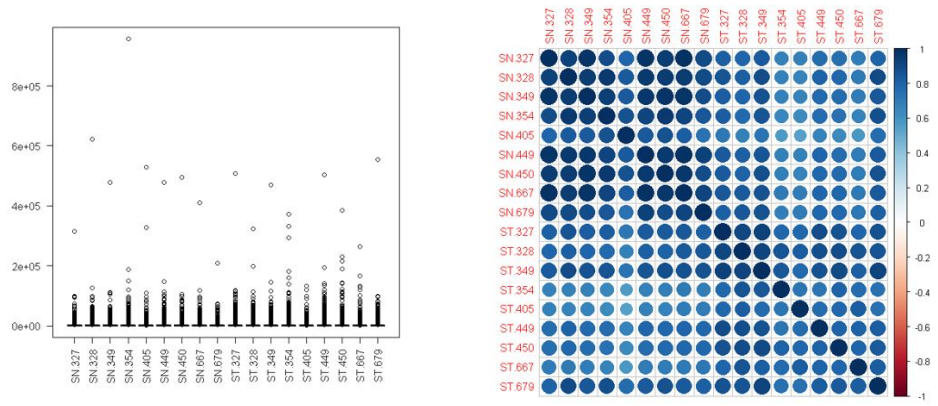
Se observa que para todas las muestras e iteraciones ejecutadas la metodología Deseq2 estima un mayor número de génes diferencialmente expresados Figura **9-2**, sin embargo, ésta misma metodología muestra el mayor número de falsos positivos Figura **9-3**, haciendo dudar de la veracidad de los resultados, por este motivo se decide revisar el número de genes que hacían parte del conjunto de datos con genes expresados, encontrando que es el primer método (deseq2) es el que muestra mayor número de verdaderos positivos Figura **9-4**.

Teniendo en cuenta que los algoritmos inmersos dentro de cada metodología son modelos estadísticos, es posible generar una matriz de confusión con los resultados. Se analizan las métricas de Precisión y Recall de todos los modelos ejecutados para las muestras; encontrando que el modelo de mayor precisión es el modelo lineal que se ejecuta dentro de la metodología Limma Figura **9-5** y para la sensibilidad o recall es el modelo lineal generalizado dentro de Deseq2 el que presenta mayores porcentajes Figura **9-6**.

### 9.0.2. Aplicación cáncer de pulmón

Para la aplicación al conjunto real de genes, se tienen muestras clínicas de pacientes con carcinoma de células escamosas de pulmón, 9 muestras de tumor y 9 de tejido adyacente[3]. La base de datos contiene los conteos de cada uno de los genes, en cada una de las muestras.

El número total de génes es de 26.363, se realizan dos filtros, en el primero se seleccionan los datos de genes con Biotipo: protein_coding y en el segundo se eliminan los genes que tienen conteos en 0, obteniendo así un conjunto de datos de 19.150 génes. En la Figura 8-9 es posible ver el comportamiento de las muestras.

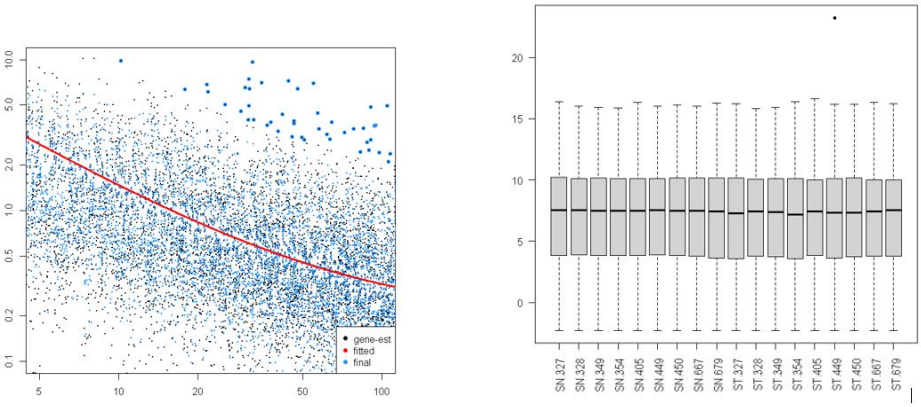



(a) Distribución muestras (b) Correlación muestras

**Figura 9-7**.: Distribución y correlación de conteos crudos

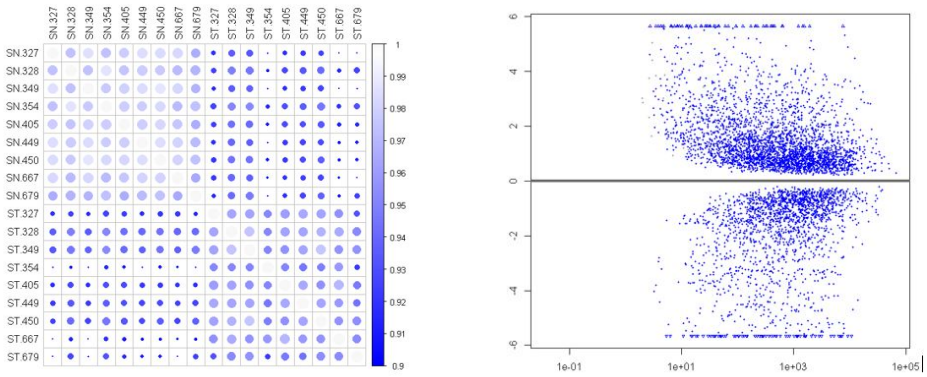
Dentro de las muestras se evidencia que hay correlaciones cercanas al 60 % entre las mismas y son dispersas.

Para continuar con el análisis y realizar el modelo lineal generalizado es necesario hacer una normalización de las muestras, obteniendo como resultado 7.506 genes diferencialmente expresados.(En el Anexo B se pueden ver los primeros 160 génes del conjuto resultante). Los resultados se visualizan en las Figuras 8-10 y 8-11.




(a) Resultado modelo (b) Distribición normalizados

**Figura 9-8**.: Modelo Deseq2 y Normalización de las muestras




(a) Correlaciones muestras normalizadas (b) MA-plot diferencialmente expresados

**Figura 9-9**.: Correlaciones y genes diferencialmente expresados

El porcentaje de genes diferencialmente expresados que se encontraron haciendo uso del método deseq2 corresponde al 39 % del total de genes que hicieron parte del modelado, esto puede deberse al conjuto de datos como tal, así como sucede en [6].
Basado en los resultados obtenidos con las muestras simuladas, se puede pensar que del porcentaje de genes diferencialmente expresado cerca del 11 % son falsos positivos y un 85 % son reales.
Para estar más seguros con los genés resultantes existe la posibilidad de utilizar varios de los métodos y comparar los genes diferencialmente expresados encontrados, realizar una comparación y así reconocer con mayor certeza los genes que realmente están diferencialmente expresados.

# 10. Conclusiones y recomendaciones

## 10.1. Conclusiones

- De los modelos comparados se puede afirmar que el método que encuentra mayor número real de genes diferenciamente expresado es Deseq2, sin embargo, hay que tener especial cuidado con su aplicación debido a que también es el método que genera el mayor número de falsos positivos.
- De la aplicación de los tres métodos a las muestras simuladas, se observó que entre más muestras y menor número de genes se tengan mejor son los resultados para Limma, Deseq2 y Noiseq, en cuanto a precisión, capacidad y rendimiento.
- Como trabajo futuro, sería importante revisar la posibilidad de generar una unión de varios modelos, revisando profundamente sus resultados, de forma que se pueda crear una metodología con mayor efectividad.

## 10.2. Recomendaciones

- Para realizar este tipo de análisis se recomienda revisar la información real, frente a correlaciones, dispersiones, entre otros.
- La decisión sobre Deseq2 como el método que encuentra mayor número de genes diferencialmente expresado se hace a partir de algunas métricas importantes, sin embargo, existen otros indicadores de la matriz de confusión que pueden ser de valor.
- Realizar comparaciónes con otros métodos como EdgeR o Deseq podría ayudar con la creación de nuevas metodologías.

# A. Anexo: Código de simulación de datos

```
B01<- generateSyntheticData(
  dataset= "prueba", n.vars= 30000,
  samples.per.cond=15, n.diffexp = 300, seqdepth=10e6,
  minfact= 1.2, maxfact= 1.4,
  relmeans = "auto", dispersions= "auto",
  fraction.upregulated= 0.5,
  between.group.diffdisp = F,
  filter.threshold.total = 10, effect.size = 4)
```

# B. Anexo: Tabla de Genes diferencialmente expresados aplicación cáncer de pulmón

| Primeros 160 génes con mayor expresión | | | |
|---|---|---|---|
| ENSG00000078898 | ENSG00000162543 | ENSG00000138180 | ENSG00000111669 |
| ENSG00000244067 | ENSG00000185250 | ENSG00000143228 | ENSG00000186409 |
| ENSG00000101448 | ENSG00000179813 | ENSG00000251655 | ENSG00000131747 |
| ENSG00000230657 | ENSG00000157429 | ENSG00000186471 | ENSG00000051180 |
| ENSG00000215182 | ENSG00000176601 | ENSG00000175166 | ENSG00000179071 |
| ENSG00000134827 | ENSG00000164932 | ENSG00000205081 | ENSG00000007174 |
| ENSG00000140534 | ENSG00000154479 | ENSG00000157330 | ENSG00000077684 |
| ENSG00000173610 | ENSG00000154556 | ENSG00000143476 | ENSG00000087586 |
| ENSG00000188039 | ENSG00000120262 | ENSG00000162643 | ENSG00000152766 |
| ENSG00000154099 | ENSG00000183644 | ENSG00000182329 | ENSG00000104237 |
| ENSG00000181092 | ENSG00000155749 | ENSG00000127324 | ENSG00000088325 |
| ENSG00000151023 | ENSG00000077157 | ENSG00000197057 | ENSG00000203780 |
| ENSG00000159763 | ENSG00000127863 | ENSG00000137807 | ENSG00000118492 |
| ENSG00000137691 | ENSG00000029153 | ENSG00000164627 | ENSG00000213085 |
| ENSG00000128536 | ENSG00000101182 | ENSG00000198826 | ENSG00000236882 |
| ENSG00000167419 | ENSG00000169126 | ENSG00000188517 | ENSG00000155085 |
| ENSG00000162078 | ENSG00000149201 | ENSG00000181378 | ENSG00000089101 |
| ENSG00000138294 | ENSG00000167034 | ENSG00000132122 | ENSG00000197653 |
| ENSG00000156042 | ENSG00000163576 | ENSG00000164675 | ENSG00000162599 |
| ENSG00000163491 | ENSG00000165795 | ENSG00000155530 | ENSG00000137494 |
| ENSG00000119147 | ENSG00000124107 | ENSG00000094755 | ENSG00000122952 |
| ENSG00000160862 | ENSG00000159588 | ENSG00000170959 | ENSG00000117983 |
| ENSG00000034971 | ENSG00000166596 | ENSG00000196277 | ENSG00000157856 |
| ENSG00000179902 | ENSG00000170891 | ENSG00000144061 | ENSG00000116117 |
| ENSG00000158486 | ENSG00000204566 | ENSG00000093134 | ENSG00000206530 |
| ENSG00000145491 | ENSG00000250305 | ENSG00000133640 | ENSG00000004478 |
| ENSG00000203963 | ENSG00000106772 | ENSG00000115884 | ENSG00000117477 |
| ENSG00000171962 | ENSG00000118997 | ENSG00000081277 | ENSG00000188732 |
| ENSG00000011426 | ENSG00000197748 | ENSG00000182481 | ENSG00000162814 |
| ENSG00000204361 | ENSG00000145476 | ENSG00000080572 | ENSG00000137804 |
| ENSG00000163071 | ENSG00000188659 | ENSG00000152763 | ENSG00000112539 |
| ENSG00000101003 | ENSG00000186952 | ENSG00000197168 | ENSG00000159166 |
| ENSG00000137098 | ENSG00000129295 | ENSG00000119636 | ENSG00000095637 |
| ENSG00000203734 | ENSG00000117724 | ENSG00000105220 | ENSG00000078098 |
| ENSG00000136231 | ENSG00000213204 | ENSG00000124490 | ENSG00000099942 |
| ENSG00000102547 | ENSG00000169347 | ENSG00000160838 | ENSG00000196419 |
| ENSG00000140057 | ENSG00000155761 | ENSG00000175267 | ENSG00000115423 |
| ENSG00000152936 | ENSG00000175084 | ENSG00000172403 | ENSG00000168959 |
| ENSG00000118193 | ENSG00000180263 | ENSG00000135951 | ENSG00000165923 |
| ENSG00000172771 | ENSG00000161800 | ENSG00000130413 | ENSG00000110400 |

# C. Anexo: Resultados simulación

**Tabla C-1**.: Resultados: Génes dif exp muestra 1a.

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g172 | 1 | 1 | 1 | 1 |
| g361 | 1 | 1 | 1 | 1 |
| g232 | 1 | 1 | 1 | 1 |
| g314 | 1 | 1 | 1 | 1 |
| g490 | 1 | 1 | 1 | 1 |
| g58 | 1 | 1 | 1 | 1 |
| g414 | 1 | 1 | 1 | 1 |
| g111 | 1 | 1 | 1 | 1 |
| g182 | 1 | 1 | 1 | 1 |
| g140 | 1 | 1 | 1 | 1 |
| g206 | 1 | 1 | 1 | 1 |
| g337 | 1 | 1 | 1 | 1 |
| g190 | 1 | 1 | 1 | 1 |
| g464 | 1 | 1 | 1 | 1 |
| g313 | 1 | 1 | 1 | 1 |
| g336 | 1 | 1 | 1 | 1 |
| g10 | 1 | 1 | 1 | 1 |
| g296 | 1 | 1 | 1 | 1 |
| g330 | 1 | 1 | 1 | 1 |
| g136 | 1 | 1 | 1 | 1 |
| g30 | 1 | 1 | 1 | 1 |
| g353 | 1 | 1 | 1 | 1 |
| g203 | 1 | 1 | 1 | 1 |
| g48 | 1 | 1 | 1 | 1 |
| g77 | 1 | 1 | 1 | 1 |
| g92 | 1 | 1 | 1 | 1 |
| g107 | 1 | 1 | 1 | 1 |
| g363 | 1 | 1 | 1 | 1 |
| g72 | 1 | 1 | 1 | 1 |
| g124 | 1 | 1 | 1 | 1 |
| g98 | 1 | 1 | 1 | 1 |
| g331 | 1 | 1 | 1 | 1 |
| g493 | 1 | 1 | 1 | 1 |
| g189 | 1 | 1 | 1 | 1 |
| g116 | 1 | 1 | 1 | 1 |
| g342 | 1 | 1 | 1 | 1 |
| g96 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g139 | 1 | 1 | 1 | 1 |
| g297 | 1 | 1 | 1 | 1 |
| g300 | 1 | 1 | 1 | 1 |
| g303 | 1 | 1 | 1 | 1 |
| g230 | 1 | 1 | 1 | 1 |
| g14 | 1 | 1 | 1 | 1 |
| g34 | 1 | 1 | 1 | 1 |
| g250 | 1 | 1 | 1 | 1 |
| g2 | 1 | 1 | 1 | 1 |
| g25 | 1 | 1 | 1 | 1 |
| g163 | 1 | 1 | 1 | 1 |
| g462 | 1 | 1 | 1 | 1 |
| g389 | 1 | 1 | 1 | 1 |
| g237 | 1 | 1 | 1 | 1 |
| g444 | 1 | 1 | 1 | 1 |
| g88 | 1 | 1 | 1 | 1 |
| g199 | 1 | 1 | 1 | 1 |
| g408 | 1 | 1 | 1 | 1 |
| g51 | 1 | 1 | 1 | 1 |
| g12 | 1 | 1 | 1 | 1 |
| g299 | 1 | 1 | 1 | 1 |
| g161 | 1 | 1 | 1 | 1 |
| g62 | 1 | 1 | 1 | 1 |
| g443 | 1 | 1 | 1 | 1 |
| g450 | 1 | 1 | 1 | 1 |
| g357 | 1 | 1 | 1 | 1 |
| g281 | 1 | 1 | 1 | 1 |
| g221 | 1 | 1 | 1 | 1 |
| g320 | 1 | 1 | 1 | 1 |
| g268 | 1 | 1 | 1 | 1 |
| g430 | 1 | 1 | 1 | 1 |
| g106 | 1 | 1 | 1 | 1 |
| g388 | 1 | 1 | 1 | 1 |
| g478 | 1 | 1 | 1 | 1 |
| g197 | 1 | 1 | 1 | 1 |
| g205 | 1 | 1 | 1 | 1 |
| g249 | 1 | 1 | 1 | 1 |
| g8 | 1 | 1 | 1 | 1 |
| g184 | 1 | 1 | 1 | 1 |
| g79 | 1 | 1 | 1 | 1 |
| g215 | 1 | 1 | 1 | 1 |
| g407 | 1 | 1 | 1 | 1 |
| g192 | 1 | 1 | 1 | 1 |
| g168 | 1 | 1 | 1 | 1 |
| g39 | 1 | 1 | 1 | 1 |
| g126 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g233 | 1 | 1 | 1 | 1 |
| g406 | 1 | 1 | 1 | 1 |
| g272 | 1 | 1 | 1 | 1 |
| g22 | 1 | 1 | 1 | 1 |
| g129 | 1 | 1 | 1 | 1 |
| g348 | 1 | 1 | 1 | 1 |
| g321 | 1 | 1 | 1 | 1 |
| g243 | 1 | 1 | 1 | 1 |
| g70 | 1 | 1 | 1 | 1 |
| g75 | 1 | 1 | 1 | 1 |
| g198 | 1 | 1 | 1 | 1 |
| g305 | 1 | 1 | 1 | 1 |
| g108 | 1 | 1 | 1 | 1 |
| g328 | 1 | 1 | 1 | 1 |
| g26 | 1 | 1 | 1 | 1 |
| g57 | 1 | 1 | 1 | 1 |
| g127 | 1 | 1 | 1 | 1 |
| g148 | 1 | 1 | 1 | 1 |
| g488 | 1 | 1 | 1 | 1 |
| g11 | 1 | 1 | 1 | 1 |
| g261 | 1 | 1 | 1 | 1 |
| g437 | 1 | 1 | 1 | 1 |
| g465 | 1 | 1 | 1 | 1 |
| g426 | 1 | 1 | 1 | 1 |
| g149 | 1 | 1 | 1 | 1 |
| g480 | 1 | 1 | 1 | 1 |
| g403 | 1 | 1 | 1 | 1 |
| g220 | 1 | 1 | 1 | |
| g454 | 1 | 1 | 1 | 1 |
| g187 | 1 | 1 | 1 | 1 |
| g429 | 1 | 1 | 1 | |
| g147 | 1 | 1 | 1 | 1 |
| g291 | 1 | 1 | 1 | 1 |
| g339 | 1 | 1 | 1 | 1 |
| g214 | 1 | 1 | 1 | 1 |
| g260 | 1 | 1 | 1 | 1 |
| g183 | 1 | 1 | 1 | 1 |
| g177 | 1 | 1 | 1 | 1 |
| g52 | 1 | 1 | 1 | 1 |
| g308 | 1 | 1 | 1 | 1 |
| g391 | 1 | 1 | 1 | 1 |
| g119 | 1 | 1 | 1 | 1 |
| g132 | 1 | 1 | 1 | 1 |
| g346 | 1 | 1 | 1 | 1 |
| g145 | 1 | 1 | 1 | 1 |
| g170 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g441 | 1 | 1 | 1 | 1 |
| g151 | 1 | 1 | 1 | 1 |
| g142 | 1 | 1 | 1 | 1 |
| g239 | 1 | 1 | 1 | 1 |
| g356 | 1 | 1 | 1 | 1 |
| g289 | 1 | 1 | 1 | 1 |
| g242 | 1 | 1 | 1 | 1 |
| g470 | 1 | 1 | 1 | 1 |
| g240 | 1 | 1 | 1 | 1 |
| g253 | 1 | 1 | 1 | 1 |
| g41 | 1 | 1 | 1 | 1 |
| g495 | 1 | 1 | 1 | 1 |
| g312 | 1 | 1 | 1 | |
| g207 | 1 | 1 | 1 | 1 |
| g54 | 1 | 1 | 1 | 1 |
| g340 | 1 | 1 | 1 | 1 |
| g469 | 1 | 1 | 1 | 1 |
| g267 | 1 | 1 | 1 | 1 |
| g474 | 1 | 1 | 1 | 1 |
| g273 | 1 | 1 | 1 | 1 |
| g409 | 1 | 1 | 1 | 1 |
| g225 | 1 | 1 | 1 | 1 |
| g256 | 1 | 1 | 1 | 1 |
| g164 | 1 | 1 | 1 | 1 |
| g293 | 1 | 1 | 1 | 1 |
| g451 | 1 | 1 | 1 | 1 |
| g216 | 1 | 1 | 1 | 1 |
| g231 | 1 | 1 | 1 | 1 |
| g358 | 1 | 1 | 1 | 1 |
| g382 | 1 | 1 | 1 | 1 |
| g445 | 1 | 1 | 1 | 1 |
| g3 | 1 | 1 | 1 | 1 |
| g457 | 1 | 1 | 1 | 1 |
| g78 | 1 | 1 | 1 | 1 |
| g329 | 1 | 1 | 1 | 1 |
| g204 | 1 | 1 | 1 | 1 |
| g224 | 1 | 1 | 1 | 1 |
| g422 | 1 | 1 | 1 | |
| g301 | 1 | 1 | 1 | 1 |
| g81 | 1 | 1 | 1 | 1 |
| g285 | 1 | 1 | 1 | 1 |
| g109 | 1 | 1 | 1 | 1 |
| g384 | 1 | 1 | 1 | 1 |
| g266 | 1 | 1 | 1 | |
| g208 | 1 | 1 | 1 | 1 |
| g459 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g36 | 1 | 1 | 1 | 1 |
| g23 | 1 | 1 | 1 | 1 |
| g284 | 1 | 1 | 1 | 1 |
| g481 | 1 | 1 | 1 | 1 |
| g194 | 1 | 1 | 1 | 1 |
| g99 | 1 | 1 | 1 | 1 |
| g318 | 1 | 1 | 1 | |
| g264 | 1 | 1 | 1 | 1 |
| g42 | 1 | 1 | 1 | 1 |
| g370 | 1 | 1 | 1 | 1 |
| g368 | 1 | 1 | 1 | |
| g417 | 1 | 1 | 1 | 1 |
| g65 | 1 | 1 | 1 | 1 |
| g43 | 1 | 1 | 1 | 1 |
| g456 | 1 | 1 | 1 | 1 |
| g82 | 1 | 1 | 1 | 1 |
| g446 | 1 | 1 | 1 | |
| g463 | 1 | 1 | 1 | 1 |
| g473 | 1 | 1 | 1 | 1 |
| g68 | 1 | 1 | 1 | 1 |
| g453 | 1 | 1 | 1 | 1 |
| g319 | 1 | 1 | 1 | 1 |
| g367 | 1 | 1 | 1 | 1 |
| g125 | 1 | 1 | 1 | 1 |
| g461 | 1 | 1 | 1 | 1 |
| g46 | 1 | 1 | 1 | |
| g128 | 1 | 1 | 1 | 1 |
| g254 | 1 | 1 | 1 | 1 |
| g351 | 1 | 1 | 1 | |
| g174 | 1 | 1 | 1 | 1 |
| g248 | 1 | 1 | 1 | 1 |
| g93 | 1 | 1 | 1 | 1 |
| g263 | 1 | 1 | 1 | 1 |
| g334 | 1 | 1 | 1 | 1 |
| g173 | 1 | 1 | 1 | 1 |
| g227 | 1 | 1 | 1 | |
| g217 | 1 | 1 | 1 | 1 |
| g379 | 1 | 1 | 1 | 1 |
| g306 | 1 | 1 | 1 | 1 |
| g427 | 1 | 1 | 1 | |
| g335 | 1 | 1 | 1 | 1 |
| g428 | 1 | 1 | 1 | |
| g9 | 1 | 1 | 1 | 1 |
| g316 | 1 | 1 | 1 | 1 |
| g387 | 1 | 1 | 1 | 1 |
| g1 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g471 | 1 | 1 | 1 | 1 |
| g85 | 1 | 1 | 1 | |
| g135 | 1 | 1 | 1 | 1 |
| g307 | 1 | 1 | 1 | |
| g439 | 1 | 1 | 1 | 1 |
| g35 | 1 | 1 | 1 | 1 |
| g466 | 1 | 1 | 1 | 1 |
| g377 | 1 | 1 | 1 | 1 |
| g144 | 1 | 1 | 1 | 1 |
| g143 | 1 | 1 | 1 | |
| g398 | 1 | 1 | 1 | 1 |
| g113 | 1 | 1 | 1 | 1 |
| g252 | 1 | 1 | 1 | 1 |
| g486 | 1 | 1 | 1 | |
| g276 | 1 | 1 | 1 | 1 |
| g355 | 1 | 1 | 1 | 1 |
| g419 | 1 | 1 | 1 | 1 |
| g326 | 1 | 1 | 1 | 1 |
| g67 | 1 | 1 | 1 | 1 |
| g394 | 1 | 1 | 1 | 1 |
| g15 | 1 | 1 | 1 | 1 |
| g421 | 1 | 1 | 1 | |
| g375 | 1 | 1 | 1 | |
| g386 | 1 | 1 | 1 | |
| g476 | 1 | 1 | 1 | |
| g410 | 1 | 1 | 1 | |
| g117 | 1 | 1 | 1 | 1 |
| g175 | 1 | 1 | 1 | 1 |
| g153 | 1 | 1 | 1 | 1 |
| g292 | 1 | 1 | 1 | 1 |
| g16 | 1 | 1 | 1 | |
| g112 | 1 | 1 | 1 | 1 |
| g167 | 1 | 1 | 1 | |
| g489 | 1 | 1 | 1 | |
| g37 | 1 | 1 | 1 | 1 |
| g364 | 1 | 1 | 1 | 1 |
| g76 | 1 | 1 | 1 | 1 |
| g7 | 1 | 1 | 1 | |
| g255 | 1 | 1 | 1 | 1 |
| g114 | 1 | 1 | 1 | 1 |
| g365 | 1 | 1 | 1 | 1 |
| g201 | 1 | 1 | 1 | |
| g343 | 1 | 1 | 1 | |
| g327 | 1 | 1 | 1 | |
| g487 | 1 | 1 | 1 | 1 |
| g467 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g399 | 1 | 1 | 1 | |
| g21 | 1 | 1 | 1 | |
| g302 | 1 | 1 | 1 | 1 |
| g50 | 1 | 1 | 1 | 1 |
| g131 | 1 | 1 | 1 | |
| g31 | 1 | 1 | 1 | 1 |
| g468 | 1 | 1 | 1 | |
| g374 | 1 | 1 | 1 | |
| g73 | 1 | 1 | 1 | 1 |
| g49 | 1 | 1 | 1 | 1 |
| g258 | 1 | 1 | 1 | 1 |
| g202 | 1 | 1 | 1 | 1 |
| g222 | 1 | 1 | 1 | 1 |
| g155 | 1 | 1 | 1 | 1 |
| g80 | 1 | 1 | 1 | 1 |
| g485 | 1 | 1 | 1 | |
| g195 | 1 | 1 | 1 | 1 |
| g369 | 1 | 1 | 1 | 1 |
| g212 | 1 | 1 | 1 | 1 |
| g241 | 1 | 1 | 1 | |
| g154 | 1 | 1 | 1 | |
| g19 | 1 | 1 | 1 | 1 |
| g83 | 1 | 1 | 1 | 1 |
| g405 | 1 | 1 | 1 | 1 |
| g274 | 1 | 1 | 1 | |
| g449 | 1 | 1 | 1 | |
| g349 | 1 | 1 | 1 | |
| g338 | 1 | 1 | 1 | 1 |
| g185 | 1 | 1 | 1 | |
| g378 | 1 | 1 | 1 | 1 |
| g350 | 1 | 1 | 1 | |
| g94 | 1 | 1 | 1 | 1 |
| g400 | 1 | 1 | 1 | |
| g9241 | | 1 | 1 | |
| g100 | 1 | 1 | 1 | 1 |
| g433 | 1 | 1 | 1 | |
| g415 | 1 | 1 | 1 | |
| g211 | 1 | 1 | 1 | 1 |
| g56 | 1 | 1 | 1 | 1 |
| g257 | 1 | 1 | 1 | |
| g141 | 1 | 1 | 1 | 1 |
| g309 | 1 | 1 | 1 | 1 |
| g423 | 1 | 1 | 1 | |
| g6 | 1 | 1 | 1 | |
| g385 | 1 | 1 | 1 | |
| g223 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g247 | 1 | 1 | 1 | 1 |
| g120 | 1 | 1 | 1 | 1 |
| g74 | 1 | 1 | 1 | |
| g137 | 1 | 1 | 1 | 1 |
| g60 | 1 | 1 | 1 | |
| g123 | 1 | 1 | 1 | 1 |
| g294 | 1 | 1 | 1 | |
| g484 | 1 | 1 | 1 | |
| g179 | 1 | 1 | 1 | 1 |
| g71 | 1 | 1 | 1 | 1 |
| g280 | 1 | 1 | 1 | 1 |
| g283 | 1 | 1 | 1 | 1 |
| g4 | 1 | 1 | 1 | 1 |
| g317 | 1 | 1 | 1 | |
| g12781 | | 1 | 1 | |
| g259 | 1 | 1 | 1 | |
| g269 | 1 | 1 | 1 | 1 |
| g219 | 1 | 1 | 1 | 1 |
| g102 | 1 | 1 | 1 | 1 |
| g492 | 1 | 1 | 1 | |
| g63 | 1 | 1 | 1 | 1 |
| g104 | 1 | 1 | 1 | 1 |
| g278 | 1 | 1 | 1 | |
| g279 | 1 | 1 | 1 | |
| g61 | 1 | 1 | 1 | 1 |
| g91 | 1 | 1 | 1 | 1 |
| g156 | 1 | 1 | 1 | 1 |
| g90 | 1 | 1 | 1 | 1 |
| g97 | 1 | 1 | 1 | 1 |
| g110 | 1 | 1 | 1 | |
| g29 | 1 | 1 | 1 | |
| g146 | 1 | 1 | 1 | |
| g95 | 1 | 1 | 1 | |
| g13859 | | 1 | 1 | 1 |
| g29495 | | 1 | | |
| g7241 | | 1 | 1 | |
| g64 | 1 | 1 | 1 | |
| g4993 | | 1 | 1 | |
| g436 | 1 | 1 | 1 | 1 |
| g59 | 1 | 1 | 1 | 1 |
| g13762 | | 1 | 1 | 1 |
| g404 | 1 | 1 | 1 | 1 |
| g366 | 1 | 1 | 1 | |
| g447 | 1 | 1 | 1 | |
| g6955 | | 1 | 1 | |
| g16069 | | 1 | 1 | |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g352 | 1 | 1 | 1 | |
| g397 | 1 | 1 | 1 | |
| g66 | 1 | 1 | 1 | 1 |
| g218 | 1 | 1 | 1 | |
| g322 | 1 | 1 | 1 | |
| g229 | 1 | 1 | 1 | 1 |
| g181 | 1 | 1 | 1 | 1 |
| g354 | 1 | 1 | | |
| g371 | 1 | 1 | | |
| g359 | 1 | 1 | 1 | 1 |
| g482 | 1 | 1 | 1 | |
| g390 | 1 | 1 | 1 | |
| g178 | 1 | 1 | 1 | 1 |
| g158 | 1 | 1 | 1 | 1 |
| g418 | 1 | 1 | 1 | |
| g25617 | | 1 | 1 | |
| g392 | 1 | 1 | 1 | |
| g166 | 1 | 1 | 1 | 1 |
| g10103 | | 1 | | |
| g25850 | | 1 | | 1 |
| g18928 | | 1 | 1 | |
| g5 | 1 | 1 | 1 | 1 |
| g24291 | | 1 | 1 | |
| g157 | 1 | 1 | 1 | 1 |
| g22706 | | 1 | 1 | |
| g176 | 1 | 1 | 1 | 1 |
| g332 | 1 | 1 | 1 | 1 |
| g333 | 1 | 1 | 1 | 1 |
| g22637 | | 1 | 1 | 1 |
| g13147 | | 1 | 1 | 1 |
| g17806 | | 1 | | 1 |
| g23949 | | 1 | 1 | |
| g8177 | | 1 | 1 | 1 |
| g431 | 1 | 1 | | |
| g452 | 1 | 1 | 1 | |
| g402 | 1 | 1 | 1 | 1 |
| g14136 | | 1 | 1 | |
| g15117 | | 1 | | |
| g3737 | | 1 | | |
| g5521 | | 1 | | |
| g8176 | | 1 | 1 | 1 |
| g271 | 1 | 1 | 1 | |
| g45 | 1 | 1 | 1 | 1 |
| g5994 | | 1 | | 1 |
| g311 | 1 | 1 | | 1 |
| g47 | 1 | 1 | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g13 | 1 | | | 1 |
| g17 | 1 | | | 1 |
| g18 | 1 | | | 1 |
| g20 | 1 | | | 1 |
| g24 | 1 | | | 1 |
| g27 | 1 | | | 1 |
| g28 | 1 | | | 1 |
| g32 | 1 | | 1 | 1 |
| g33 | 1 | | | |
| g38 | 1 | | 1 | |
| g40 | 1 | | | |
| g44 | 1 | | 1 | |
| g53 | 1 | | | |
| g55 | 1 | | 1 | 1 |
| g69 | 1 | | | 1 |
| g84 | 1 | | 1 | 1 |
| g86 | 1 | | 1 | 1 |
| g87 | 1 | | | 1 |
| g89 | 1 | | 1 | 1 |
| g101 | 1 | | | 1 |
| g103 | 1 | | 1 | 1 |
| g105 | 1 | | 1 | 1 |
| g115 | 1 | | | 1 |
| g118 | 1 | | | 1 |
| g121 | 1 | | 1 | 1 |
| g122 | 1 | | 1 | 1 |
| g130 | 1 | | | 1 |
| g133 | 1 | | 1 | 1 |
| g134 | 1 | | 1 | 1 |
| g138 | 1 | | 1 | 1 |
| g150 | 1 | | | 1 |
| g152 | 1 | | 1 | 1 |
| g159 | 1 | | | 1 |
| g160 | 1 | | 1 | 1 |
| g162 | 1 | | 1 | 1 |
| g165 | 1 | | | 1 |
| g169 | 1 | | | 1 |
| g171 | 1 | | | 1 |
| g180 | 1 | | | 1 |
| g186 | 1 | | 1 | 1 |
| g188 | 1 | | 1 | 1 |
| g191 | 1 | | 1 | 1 |
| g193 | 1 | | | 1 |
| g196 | 1 | | | 1 |
| g200 | 1 | | | 1 |
| g209 | 1 | | 1 | 1 |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g210 | 1 | | | 1 |
| g213 | 1 | | | 1 |
| g226 | 1 | | | 1 |
| g228 | 1 | | 1 | 1 |
| g234 | 1 | | 1 | 1 |
| g235 | 1 | | | 1 |
| g236 | 1 | | | 1 |
| g238 | 1 | | | 1 |
| g244 | 1 | | | 1 |
| g245 | 1 | | 1 | 1 |
| g246 | 1 | | 1 | 1 |
| g251 | 1 | | | 1 |
| g262 | 1 | | | 1 |
| g265 | 1 | | | 1 |
| g270 | 1 | | | 1 |
| g275 | 1 | | | 1 |
| g277 | 1 | | | 1 |
| g282 | 1 | | | 1 |
| g286 | 1 | | | 1 |
| g287 | 1 | | | 1 |
| g288 | 1 | | | 1 |
| g290 | 1 | | | 1 |
| g295 | 1 | | | 1 |
| g298 | 1 | | | 1 |
| g304 | 1 | | | 1 |
| g310 | 1 | | | 1 |
| g315 | 1 | | | 1 |
| g323 | 1 | | | 1 |
| g324 | 1 | | | |
| g325 | 1 | | 1 | |
| g341 | 1 | | | |
| g344 | 1 | | | |
| g345 | 1 | | | |
| g347 | 1 | | 1 | |
| g360 | 1 | | | |
| g362 | 1 | | 1 | |
| g372 | 1 | | | |
| g373 | 1 | | 1 | 1 |
| g376 | 1 | | 1 | |
| g380 | 1 | | | 1 |
| g381 | 1 | | | |
| g383 | 1 | | | |
| g393 | 1 | | | |
| g395 | 1 | | | |
| g396 | 1 | | 1 | |
| g401 | 1 | | | |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g411 | 1 | | | |
| g412 | 1 | | | |
| g413 | 1 | | | |
| g416 | 1 | | | |
| g420 | 1 | | | |
| g424 | 1 | | | |
| g425 | 1 | | | |
| g432 | 1 | | 1 | |
| g434 | 1 | | | |
| g435 | 1 | | 1 | |
| g438 | 1 | | | |
| g440 | 1 | | | 1 |
| g442 | 1 | | 1 | 1 |
| g448 | 1 | | | |
| g455 | 1 | | | |
| g458 | 1 | | | |
| g460 | 1 | | | |
| g472 | 1 | | 1 | |
| g475 | 1 | | | |
| g477 | 1 | | 1 | |
| g479 | 1 | | | |
| g483 | 1 | | 1 | 1 |
| g491 | 1 | | | |
| g494 | 1 | | | |
| g10065 | | | 1 | |
| g5356 | | | 1 | |
| g3478 | | | 1 | |
| g1143 | | | 1 | |
| g7551 | | | 1 | |
| g2537 | | | 1 | 1 |
| g5880 | | | 1 | |
| g9091 | | | 1 | |
| g13906 | | | 1 | |
| g29065 | | | 1 | |
| g18537 | | | 1 | 1 |
| g20380 | | | 1 | |
| g12665 | | | 1 | |
| g5973 | | | 1 | |
| g13132 | | | 1 | |
| g9169 | | | 1 | |
| g12615 | | | 1 | |
| g5729 | | | 1 | |
| g13229 | | | 1 | |
| g28324 | | | 1 | |
| g29428 | | | 1 | |
| g27969 | | | 1 | |

**Tabla C-1 Resultados: Génes dif exp muestra 1a.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g21346 | | | 1 | |
| g28898 | | | 1 | 1 |
| g2777 | | | 1 | |
| g21022 | | | 1 | |
| g19072 | | | 1 | 1 |
| g25541 | | | 1 | |
| g4907 | | | 1 | |
| g25794 | | | 1 | 1 |
| g9735 | | | 1 | |
| g24759 | | | 1 | |
| g12186 | | | 1 | |
| g20410 | | | 1 | 1 |
| g27042 | | | 1 | |
| g20473 | | | 1 | |
| g25418 | | | 1 | |
| g7738 | | | 1 | |
| g20297 | | | 1 | |
| g29772 | | | 1 | |
| g17214 | | | 1 | |
| g3969 | | | 1 | |
| g19180 | | | 1 | |
| g4901 | | | 1 | |
| g21476 | | | | 1 |
| g12930 | | | | 1 |
| g29304 | | | | 1 |
| g25402 | | | | 1 |
| g10056 | | | | 1 |
| g22817 | | | | 1 |
| g29567 | | | | 1 |
| g27070 | | | | 1 |
| g7275 | | | | 1 |
| g26617 | | | | 1 |
| g9090 | | | | 1 |
| g25233 | | | | 1 |
| g787 | | | | 1 |
| g7568 | | | | 1 |
| g24547 | | | | 1 |
| g24454 | | | | 1 |
| g19506 | | | | 1 |
| g15674 | | | | 1 |
| g6054 | | | | 1 |
| g4243 | | | | 1 |
| g22238 | | | | 1 |
| g4518 | | | | 1 |
| g4429 | | | | 1 |
| g7619 | | | | 1 |

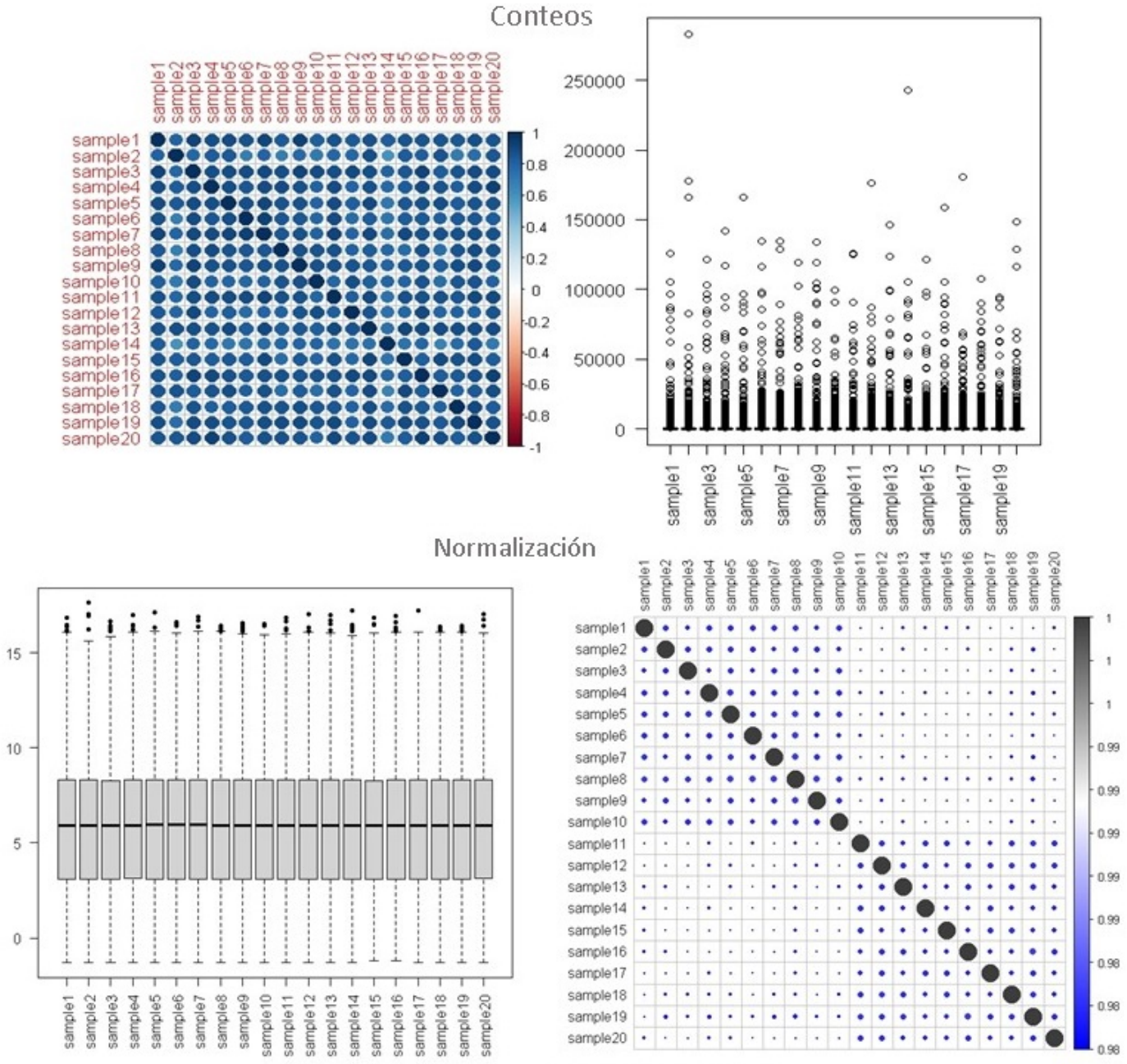



**Figura C-1**.: Gráficos de conteos y normalización de los conteos

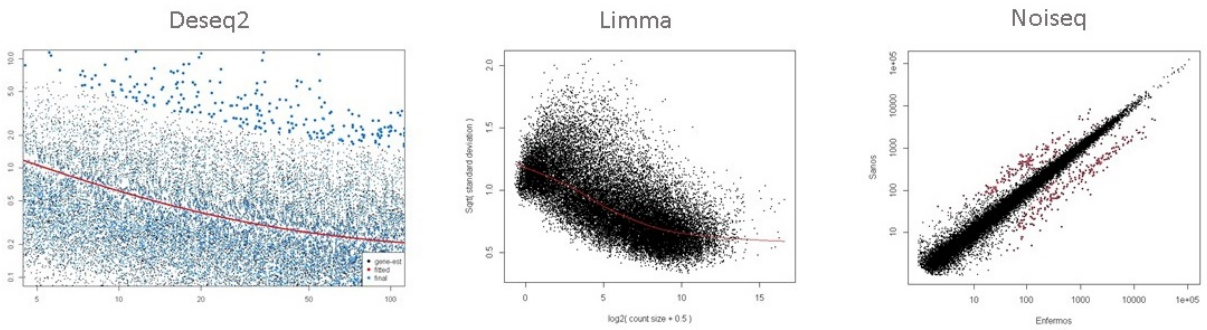



**Figura C-2**.: Resultado de los modelos

**Tabla C-2**.: Resultados: Génes dif exp muestra 1b.

| **Muestra 30000-20-500** | **Reales** | **Limma** | **Deseq2** | **Noiseq** |
|---|---|---|---|---|
| g56 | 1 | 1 | 1 | 1 |
| g132 | 1 | 1 | 1 | 1 |
| g174 | 1 | 1 | 1 | 1 |
| g467 | 1 | 1 | 1 | 1 |
| g440 | 1 | 1 | 1 | 1 |
| g352 | 1 | 1 | 1 | 1 |
| g136 | 1 | 1 | 1 | 1 |
| g464 | 1 | 1 | 1 | 1 |
| g33 | 1 | 1 | 1 | 1 |
| g17 | 1 | 1 | 1 | 1 |
| g149 | 1 | 1 | 1 | 1 |
| g172 | 1 | 1 | 1 | 1 |
| g323 | 1 | 1 | 1 | 1 |
| g359 | 1 | 1 | 1 | 1 |
| g224 | 1 | 1 | 1 | 1 |
| g300 | 1 | 1 | 1 | 1 |
| g111 | 1 | 1 | 1 | 1 |
| g72 | 1 | 1 | 1 | 1 |
| g73 | 1 | 1 | 1 | 1 |
| g255 | 1 | 1 | 1 | 1 |
| g377 | 1 | 1 | 1 | 1 |
| g97 | 1 | 1 | 1 | 1 |
| g216 | 1 | 1 | 1 | 1 |
| g497 | 1 | 1 | 1 | 1 |
| g478 | 1 | 1 | 1 | 1 |
| g494 | 1 | 1 | 1 | 1 |
| g159 | 1 | 1 | 1 | 1 |
| g79 | 1 | 1 | 1 | 1 |
| g424 | 1 | 1 | 1 | 1 |
| g313 | 1 | 1 | 1 | 1 |
| g26 | 1 | 1 | 1 | 1 |
| g292 | 1 | 1 | 1 | 1 |
| g239 | 1 | 1 | 1 | 1 |
| g365 | 1 | 1 | 1 | 1 |
| g408 | 1 | 1 | 1 | 1 |
| g385 | 1 | 1 | 1 | 1 |
| g175 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g20 | 1 | 1 | 1 | 1 |
| g47 | 1 | 1 | 1 | 1 |
| g407 | 1 | 1 | 1 | 1 |
| g393 | 1 | 1 | 1 | 1 |
| g427 | 1 | 1 | 1 | 1 |
| g287 | 1 | 1 | 1 | 1 |
| g225 | 1 | 1 | 1 | 1 |
| g332 | 1 | 1 | 1 | 1 |
| g127 | 1 | 1 | 1 | 1 |
| g177 | 1 | 1 | 1 | 1 |
| g406 | 1 | 1 | 1 | 1 |
| g230 | 1 | 1 | 1 | 1 |
| g124 | 1 | 1 | 1 | 1 |
| g397 | 1 | 1 | 1 | 1 |
| g144 | 1 | 1 | 1 | 1 |
| g31 | 1 | 1 | 1 | 1 |
| g325 | 1 | 1 | 1 | 1 |
| g34 | 1 | 1 | 1 | 1 |
| g25 | 1 | 1 | 1 | 1 |
| g205 | 1 | 1 | 1 | 1 |
| g462 | 1 | 1 | 1 | 1 |
| g293 | 1 | 1 | 1 | 1 |
| g246 | 1 | 1 | 1 | 1 |
| g78 | 1 | 1 | 1 | 1 |
| g199 | 1 | 1 | 1 | 1 |
| g466 | 1 | 1 | 1 | 1 |
| g243 | 1 | 1 | 1 | 1 |
| g277 | 1 | 1 | 1 | 1 |
| g156 | 1 | 1 | 1 | 1 |
| g275 | 1 | 1 | 1 | 1 |
| g398 | 1 | 1 | 1 | 1 |
| g231 | 1 | 1 | 1 | 1 |
| g104 | 1 | 1 | 1 | 1 |
| g384 | 1 | 1 | 1 | 1 |
| g401 | 1 | 1 | 1 | 1 |
| g141 | 1 | 1 | 1 | 1 |
| g416 | 1 | 1 | 1 | 1 |
| g85 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g129 | 1 | 1 | 1 | 1 |
| g315 | 1 | 1 | 1 | 1 |
| g254 | 1 | 1 | 1 | 1 |
| g309 | 1 | 1 | 1 | 1 |
| g13 | 1 | 1 | 1 | 1 |
| g283 | 1 | 1 | 1 | 1 |
| g162 | 1 | 1 | 1 | 1 |
| g449 | 1 | 1 | 1 | 1 |
| g64 | 1 | 1 | 1 | 1 |
| g260 | 1 | 1 | 1 | 1 |
| g305 | 1 | 1 | 1 | 1 |
| g43 | 1 | 1 | 1 | 1 |
| g110 | 1 | 1 | 1 | 1 |
| g354 | 1 | 1 | 1 | 1 |
| g82 | 1 | 1 | 1 | 1 |
| g452 | 1 | 1 | 1 | 1 |
| g235 | 1 | 1 | 1 | 1 |
| g448 | 1 | 1 | 1 | 1 |
| g457 | 1 | 1 | 1 | 1 |
| g65 | 1 | 1 | 1 | 1 |
| g22 | 1 | 1 | 1 | 1 |
| g8 | 1 | 1 | 1 | 1 |
| g316 | 1 | 1 | 1 | 1 |
| g433 | 1 | 1 | 1 | 1 |
| g44 | 1 | 1 | 1 | 1 |
| g395 | 1 | 1 | 1 | 1 |
| g483 | 1 | 1 | 1 | 1 |
| g317 | 1 | 1 | 1 | 1 |
| g195 | 1 | 1 | 1 | 1 |
| g495 | 1 | 1 | 1 | 1 |
| g458 | 1 | 1 | 1 | 1 |
| g382 | 1 | 1 | 1 | 1 |
| g386 | 1 | 1 | 1 | 1 |
| g148 | 1 | 1 | 1 | 1 |
| g403 | 1 | 1 | 1 | 1 |
| g420 | 1 | 1 | 1 | 1 |
| g90 | 1 | 1 | 1 | 1 |
| g10 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g142 | 1 | 1 | 1 | 1 |
| g185 | 1 | 1 | 1 | 1 |
| g12 | 1 | 1 | 1 | 1 |
| g182 | 1 | 1 | 1 | 1 |
| g485 | 1 | 1 | 1 | 1 |
| g154 | 1 | 1 | 1 | 1 |
| g30 | 1 | 1 | 1 | 1 |
| g15 | 1 | 1 | 1 | 1 |
| g210 | 1 | 1 | 1 | 1 |
| g84 | 1 | 1 | 1 | 1 |
| g425 | 1 | 1 | 1 | 1 |
| g200 | 1 | 1 | 1 | 1 |
| g443 | 1 | 1 | 1 | 1 |
| g54 | 1 | 1 | 1 | 1 |
| g101 | 1 | 1 | 1 | 1 |
| g261 | 1 | 1 | 1 | 1 |
| g202 | 1 | 1 | 1 | 1 |
| g423 | 1 | 1 | 1 | 1 |
| g341 | 1 | 1 | 1 | 1 |
| g344 | 1 | 1 | 1 | 1 |
| g204 | 1 | 1 | 1 | 1 |
| g158 | 1 | 1 | 1 | 1 |
| g32 | 1 | 1 | 1 | 1 |
| g459 | 1 | 1 | 1 | 1 |
| g191 | 1 | 1 | 1 | 1 |
| g203 | 1 | 1 | 1 | 1 |
| g48 | 1 | 1 | 1 | 1 |
| g128 | 1 | 1 | 1 | 1 |
| g183 | 1 | 1 | 1 | 1 |
| g334 | 1 | 1 | 1 | 1 |
| g242 | 1 | 1 | 1 | |
| g70 | 1 | 1 | 1 | 1 |
| g198 | 1 | 1 | 1 | 1 |
| g252 | 1 | 1 | 1 | 1 |
| g421 | 1 | 1 | 1 | 1 |
| g63 | 1 | 1 | 1 | 1 |
| g410 | 1 | 1 | 1 | 1 |
| g404 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g475 | 1 | 1 | 1 | 1 |
| g121 | 1 | 1 | 1 | 1 |
| g9 | 1 | 1 | 1 | 1 |
| g477 | 1 | 1 | 1 | 1 |
| g333 | 1 | 1 | 1 | 1 |
| g45 | 1 | 1 | 1 | 1 |
| g289 | 1 | 1 | 1 | 1 |
| g49 | 1 | 1 | 1 | 1 |
| g339 | 1 | 1 | 1 | 1 |
| g120 | 1 | 1 | 1 | 1 |
| g295 | 1 | 1 | 1 | 1 |
| g212 | 1 | 1 | 1 | 1 |
| g302 | 1 | 1 | 1 | 1 |
| g178 | 1 | 1 | 1 | 1 |
| g480 | 1 | 1 | 1 | 1 |
| g67 | 1 | 1 | 1 | 1 |
| g249 | 1 | 1 | 1 | 1 |
| g455 | 1 | 1 | 1 | 1 |
| g150 | 1 | 1 | 1 | 1 |
| g209 | 1 | 1 | 1 | 1 |
| g324 | 1 | 1 | 1 | 1 |
| g434 | 1 | 1 | 1 | 1 |
| g451 | 1 | 1 | 1 | |
| g259 | 1 | 1 | 1 | 1 |
| g294 | 1 | 1 | 1 | 1 |
| g201 | 1 | 1 | 1 | 1 |
| g188 | 1 | 1 | 1 | 1 |
| g390 | 1 | 1 | 1 | 1 |
| g409 | 1 | 1 | 1 | 1 |
| g19 | 1 | 1 | 1 | 1 |
| g388 | 1 | 1 | 1 | |
| g95 | 1 | 1 | 1 | 1 |
| g68 | 1 | 1 | 1 | 1 |
| g238 | 1 | 1 | 1 | 1 |
| g62 | 1 | 1 | 1 | 1 |
| g387 | 1 | 1 | 1 | 1 |
| g482 | 1 | 1 | 1 | 1 |
| g439 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g314 | 1 | 1 | 1 | |
| g147 | 1 | 1 | 1 | |
| g291 | 1 | 1 | 1 | 1 |
| g165 | 1 | 1 | 1 | 1 |
| g213 | 1 | 1 | 1 | 1 |
| g380 | 1 | 1 | 1 | 1 |
| g465 | 1 | 1 | 1 | 1 |
| g77 | 1 | 1 | 1 | 1 |
| g168 | 1 | 1 | 1 | 1 |
| g276 | 1 | 1 | 1 | 1 |
| g266 | 1 | 1 | 1 | |
| g414 | 1 | 1 | 1 | 1 |
| g161 | 1 | 1 | 1 | 1 |
| g52 | 1 | 1 | 1 | |
| g479 | 1 | 1 | 1 | |
| g481 | 1 | 1 | 1 | 1 |
| g468 | 1 | 1 | 1 | 1 |
| g93 | 1 | 1 | 1 | 1 |
| g81 | 1 | 1 | 1 | 1 |
| g139 | 1 | 1 | 1 | 1 |
| g229 | 1 | 1 | 1 | 1 |
| g400 | 1 | 1 | 1 | 1 |
| g23 | 1 | 1 | 1 | 1 |
| g383 | 1 | 1 | 1 | 1 |
| g166 | 1 | 1 | 1 | 1 |
| g18 | 1 | 1 | 1 | 1 |
| g488 | 1 | 1 | 1 | |
| g428 | 1 | 1 | 1 | 1 |
| g179 | 1 | 1 | 1 | |
| g214 | 1 | 1 | 1 | 1 |
| g83 | 1 | 1 | 1 | 1 |
| g335 | 1 | 1 | 1 | |
| g190 | 1 | 1 | 1 | |
| g336 | 1 | 1 | 1 | 1 |
| g498 | 1 | 1 | 1 | |
| g329 | 1 | 1 | 1 | 1 |
| g143 | 1 | 1 | 1 | 1 |
| g126 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g107 | 1 | 1 | 1 | 1 |
| g197 | 1 | 1 | 1 | 1 |
| g311 | 1 | 1 | 1 | 1 |
| g394 | 1 | 1 | 1 | |
| g102 | 1 | 1 | 1 | 1 |
| g286 | 1 | 1 | 1 | |
| g326 | 1 | 1 | 1 | 1 |
| g206 | 1 | 1 | 1 | 1 |
| g211 | 1 | 1 | 1 | 1 |
| g262 | 1 | 1 | 1 | 1 |
| g374 | 1 | 1 | 1 | 1 |
| g222 | 1 | 1 | 1 | 1 |
| g487 | 1 | 1 | 1 | 1 |
| g37 | 1 | 1 | 1 | 1 |
| g489 | 1 | 1 | 1 | |
| g441 | 1 | 1 | 1 | |
| g250 | 1 | 1 | 1 | 1 |
| g372 | 1 | 1 | 1 | 1 |
| g430 | 1 | 1 | 1 | 1 |
| g274 | 1 | 1 | 1 | 1 |
| g415 | 1 | 1 | 1 | |
| g470 | 1 | 1 | 1 | 1 |
| g99 | 1 | 1 | 1 | 1 |
| g140 | 1 | 1 | 1 | 1 |
| g492 | 1 | 1 | 1 | 1 |
| g471 | 1 | 1 | 1 | 1 |
| g91 | 1 | 1 | 1 | 1 |
| g112 | 1 | 1 | 1 | 1 |
| g419 | 1 | 1 | 1 | |
| g240 | 1 | 1 | 1 | 1 |
| g312 | 1 | 1 | 1 | |
| g173 | 1 | 1 | 1 | 1 |
| g375 | 1 | 1 | 1 | |
| g273 | 1 | 1 | 1 | 1 |
| g450 | 1 | 1 | 1 | 1 |
| g318 | 1 | 1 | 1 | |
| g74 | 1 | 1 | 1 | 1 |
| g131 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g476 | 1 | 1 | 1 | |
| g367 | 1 | 1 | 1 | 1 |
| g368 | 1 | 1 | 1 | |
| g96 | 1 | 1 | 1 | 1 |
| g361 | 1 | 1 | 1 | |
| g469 | 1 | 1 | 1 | |
| g431 | 1 | 1 | 1 | |
| g196 | 1 | 1 | 1 | |
| g137 | 1 | 1 | 1 | 1 |
| g232 | 1 | 1 | 1 | 1 |
| g164 | 1 | 1 | 1 | |
| g194 | 1 | 1 | 1 | 1 |
| g248 | 1 | 1 | 1 | |
| g296 | 1 | 1 | 1 | |
| g27 | 1 | 1 | 1 | |
| g345 | 1 | 1 | 1 | 1 |
| g351 | 1 | 1 | 1 | |
| g356 | 1 | 1 | 1 | 1 |
| g358 | 1 | 1 | 1 | |
| g114 | 1 | 1 | 1 | 1 |
| g113 | 1 | 1 | 1 | 1 |
| g357 | 1 | 1 | 1 | |
| g350 | 1 | 1 | 1 | |
| g446 | 1 | 1 | 1 | |
| g366 | 1 | 1 | 1 | 1 |
| g88 | 1 | 1 | 1 | 1 |
| g258 | 1 | 1 | 1 | |
| g348 | 1 | 1 | 1 | |
| g151 | 1 | 1 | 1 | 1 |
| g444 | 1 | 1 | 1 | |
| g87 | 1 | 1 | 1 | 1 |
| g100 | 1 | 1 | 1 | 1 |
| g426 | 1 | 1 | 1 | 1 |
| g241 | 1 | 1 | 1 | 1 |
| g220 | 1 | 1 | 1 | 1 |
| g413 | 1 | 1 | 1 | |
| g319 | 1 | 1 | 1 | |
| g328 | 1 | 1 | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g227 | 1 | 1 | 1 | 1 |
| g370 | 1 | 1 | 1 | |
| g445 | 1 | 1 | 1 | |
| g279 | 1 | 1 | 1 | |
| g453 | 1 | 1 | 1 | 1 |
| g36 | 1 | 1 | 1 | 1 |
| g442 | 1 | 1 | 1 | |
| g363 | 1 | 1 | 1 | |
| g19823 | | 1 | 1 | |
| g473 | 1 | 1 | 1 | 1 |
| g53 | 1 | 1 | 1 | |
| g167 | 1 | 1 | 1 | |
| g122 | 1 | 1 | 1 | |
| g69 | 1 | 1 | 1 | |
| g193 | 1 | 1 | 1 | |
| g355 | 1 | 1 | 1 | |
| g3 | 1 | 1 | 1 | |
| g484 | 1 | 1 | 1 | 1 |
| g396 | 1 | 1 | 1 | |
| g412 | 1 | 1 | 1 | |
| g51 | 1 | 1 | 1 | 1 |
| g76 | 1 | 1 | 1 | 1 |
| g320 | 1 | 1 | 1 | |
| g298 | 1 | 1 | 1 | |
| g265 | 1 | 1 | 1 | 1 |
| g322 | 1 | 1 | | |
| g66 | 1 | 1 | 1 | |
| g247 | 1 | 1 | 1 | 1 |
| g290 | 1 | 1 | 1 | |
| g146 | 1 | 1 | 1 | 1 |
| g16517 | | 1 | 1 | 1 |
| g237 | 1 | 1 | 1 | |
| g438 | 1 | 1 | 1 | |
| g362 | 1 | 1 | 1 | 1 |
| g369 | 1 | 1 | 1 | |
| g342 | 1 | 1 | 1 | |
| g268 | 1 | 1 | 1 | 1 |
| g24720 | | 1 | | |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g55 | 1 | 1 | 1 | |
| g221 | 1 | 1 | 1 | |
| g472 | 1 | 1 | 1 | |
| g11 | 1 | 1 | 1 | |
| g40 | 1 | 1 | 1 | 1 |
| g490 | 1 | 1 | 1 | |
| g1 | 1 | 1 | 1 | |
| g228 | 1 | 1 | 1 | 1 |
| g331 | 1 | 1 | 1 | 1 |
| g223 | 1 | 1 | 1 | |
| g135 | 1 | 1 | 1 | 1 |
| g264 | 1 | 1 | 1 | |
| g4451 | | 1 | 1 | |
| g303 | 1 | 1 | 1 | |
| g98 | 1 | 1 | | |
| g1268 | | 1 | 1 | |
| g60 | 1 | 1 | 1 | 1 |
| g46 | 1 | 1 | 1 | |
| g272 | 1 | 1 | 1 | |
| g337 | 1 | 1 | 1 | |
| g371 | 1 | 1 | 1 | |
| g461 | 1 | 1 | 1 | |
| g17215 | | 1 | 1 | |
| g2 | 1 | 1 | 1 | 1 |
| g376 | 1 | 1 | 1 | |
| g170 | 1 | 1 | 1 | 1 |
| g92 | 1 | 1 | 1 | |
| g117 | 1 | 1 | 1 | 1 |
| g435 | 1 | 1 | 1 | |
| g80 | 1 | 1 | 1 | |
| g119 | 1 | 1 | 1 | |
| g3401 | | 1 | 1 | |
| g125 | 1 | 1 | 1 | 1 |
| g360 | 1 | 1 | 1 | |
| g496 | 1 | 1 | | |
| g346 | 1 | 1 | 1 | 1 |
| g10290 | | 1 | 1 | |
| g11610 | | 1 | 1 | |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g192 | 1 | 1 | 1 | 1 |
| g282 | 1 | 1 | 1 | |
| g310 | 1 | 1 | 1 | 1 |
| g3231 | | 1 | 1 | 1 |
| g138 | 1 | 1 | 1 | 1 |
| g447 | 1 | 1 | 1 | |
| g94 | 1 | 1 | 1 | 1 |
| g106 | 1 | 1 | 1 | 1 |
| g12804 | | 1 | 1 | |
| g422 | 1 | 1 | 1 | |
| g24081 | | 1 | 1 | 1 |
| g379 | 1 | 1 | | |
| g25494 | | 1 | 1 | |
| g42 | 1 | 1 | 1 | 1 |
| g814 | | 1 | 1 | 1 |
| g6 | 1 | 1 | 1 | |
| g270 | 1 | 1 | 1 | 1 |
| g189 | 1 | 1 | 1 | 1 |
| g11259 | | 1 | | |
| g399 | 1 | 1 | 1 | |
| g4 | 1 | | 1 | 1 |
| g5 | 1 | | 1 | 1 |
| g7 | 1 | | | |
| g14 | 1 | | 1 | 1 |
| g16 | 1 | | | |
| g21 | 1 | | | |
| g24 | 1 | | 1 | |
| g28 | 1 | | | |
| g29 | 1 | | | 1 |
| g35 | 1 | | | 1 |
| g38 | 1 | | | 1 |
| g39 | 1 | | 1 | 1 |
| g41 | 1 | | 1 | |
| g50 | 1 | | 1 | |
| g57 | 1 | | 1 | |
| g58 | 1 | | 1 | |
| g59 | 1 | | | |
| g61 | 1 | | | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g71 | 1 | | | |
| g75 | 1 | | | |
| g86 | 1 | | | |
| g89 | 1 | | | |
| g103 | 1 | | | |
| g105 | 1 | | 1 | 1 |
| g108 | 1 | | | |
| g109 | 1 | | | 1 |
| g115 | 1 | | 1 | |
| g116 | 1 | | | |
| g118 | 1 | | | 1 |
| g123 | 1 | | 1 | |
| g130 | 1 | | | |
| g133 | 1 | | | 1 |
| g134 | 1 | | | |
| g145 | 1 | | | 1 |
| g152 | 1 | | 1 | 1 |
| g153 | 1 | | 1 | |
| g155 | 1 | | 1 | 1 |
| g157 | 1 | | 1 | |
| g160 | 1 | | 1 | 1 |
| g163 | 1 | | | |
| g169 | 1 | | 1 | |
| g171 | 1 | | | 1 |
| g176 | 1 | | 1 | 1 |
| g180 | 1 | | | |
| g181 | 1 | | | 1 |
| g184 | 1 | | 1 | |
| g186 | 1 | | 1 | |
| g187 | 1 | | | |
| g207 | 1 | | | |
| g208 | 1 | | 1 | 1 |
| g215 | 1 | | 1 | 1 |
| g217 | 1 | | | |
| g218 | 1 | | | |
| g219 | 1 | | 1 | |
| g226 | 1 | | 1 | |
| g233 | 1 | | | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g234 | 1 | | | |
| g236 | 1 | | | |
| g244 | 1 | | 1 | 1 |
| g245 | 1 | | | |
| g251 | 1 | | | |
| g253 | 1 | | 1 | |
| g256 | 1 | | | |
| g257 | 1 | | | |
| g263 | 1 | | 1 | 1 |
| g267 | 1 | | | |
| g269 | 1 | | 1 | |
| g271 | 1 | | 1 | 1 |
| g278 | 1 | | | |
| g280 | 1 | | | |
| g281 | 1 | | | |
| g284 | 1 | | | |
| g285 | 1 | | | |
| g288 | 1 | | | |
| g297 | 1 | | | |
| g299 | 1 | | | |
| g301 | 1 | | | |
| g304 | 1 | | | |
| g306 | 1 | | | |
| g307 | 1 | | 1 | |
| g308 | 1 | | | 1 |
| g321 | 1 | | | 1 |
| g327 | 1 | | 1 | 1 |
| g330 | 1 | | | 1 |
| g338 | 1 | | | 1 |
| g340 | 1 | | | 1 |
| g343 | 1 | | 1 | 1 |
| g347 | 1 | | 1 | 1 |
| g349 | 1 | | | 1 |
| g353 | 1 | | | 1 |
| g364 | 1 | | | 1 |
| g373 | 1 | | | 1 |
| g378 | 1 | | | 1 |
| g381 | 1 | | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g389 | 1 | | | 1 |
| g391 | 1 | | | 1 |
| g392 | 1 | | | 1 |
| g402 | 1 | | 1 | 1 |
| g405 | 1 | | | 1 |
| g411 | 1 | | | 1 |
| g417 | 1 | | | 1 |
| g418 | 1 | | 1 | 1 |
| g429 | 1 | | | 1 |
| g432 | 1 | | 1 | 1 |
| g436 | 1 | | | 1 |
| g437 | 1 | | | 1 |
| g454 | 1 | | 1 | 1 |
| g456 | 1 | | | 1 |
| g460 | 1 | | | 1 |
| g463 | 1 | | | 1 |
| g474 | 1 | | | 1 |
| g486 | 1 | | | 1 |
| g491 | 1 | | | 1 |
| g493 | 1 | | | 1 |
| g3982 | | | 1 | 1 |
| g25888 | | | 1 | 1 |
| g16655 | | | 1 | 1 |
| g25580 | | | 1 | 1 |
| g9427 | | | 1 | 1 |
| g5756 | | | 1 | 1 |
| g27087 | | | 1 | 1 |
| g23359 | | | 1 | 1 |
| g3414 | | | 1 | 1 |
| g19488 | | | 1 | 1 |
| g2712 | | | 1 | 1 |
| g20637 | | | 1 | 1 |
| g16473 | | | 1 | 1 |
| g3178 | | | 1 | 1 |
| g19756 | | | 1 | 1 |
| g3247 | | | 1 | 1 |
| g10932 | | | 1 | 1 |
| g4891 | | | 1 | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g6648 | | | 1 | 1 |
| g24283 | | | 1 | 1 |
| g18966 | | | 1 | 1 |
| g16859 | | | 1 | 1 |
| g10406 | | | 1 | |
| g9889 | | | 1 | |
| g630 | | | 1 | |
| g6847 | | | 1 | |
| g13869 | | | 1 | |
| g10353 | | | 1 | |
| g21058 | | | 1 | |
| g667 | | | 1 | 1 |
| g14654 | | | 1 | |
| g26881 | | | 1 | 1 |
| g20735 | | | 1 | |
| g27467 | | | 1 | 1 |
| g1487 | | | 1 | |
| g4727 | | | 1 | |
| g1805 | | | 1 | |
| g27790 | | | 1 | 1 |
| g7712 | | | 1 | |
| g10650 | | | 1 | |
| g3644 | | | 1 | |
| g13354 | | | 1 | |
| g17418 | | | 1 | |
| g11857 | | | 1 | |
| g13815 | | | 1 | |
| g1393 | | | 1 | |
| g9064 | | | 1 | |
| g21563 | | | 1 | 1 |
| g6243 | | | 1 | |
| g16781 | | | 1 | |
| g12184 | | | | 1 |
| g28971 | | | | 1 |
| g18960 | | | | 1 |
| g10544 | | | | 1 |
| g18028 | | | | 1 |
| g1227 | | | | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g13362 | | | | 1 |
| g27567 | | | | 1 |
| g28860 | | | | 1 |
| g3274 | | | | 1 |
| g29658 | | | | 1 |
| g22489 | | | | 1 |
| g8682 | | | | 1 |
| g21597 | | | | 1 |
| g28987 | | | | 1 |
| g26377 | | | | 1 |
| g28086 | | | | 1 |
| g10480 | | | | 1 |
| g2170 | | | | 1 |
| g13039 | | | | 1 |
| g13518 | | | | 1 |
| g19104 | | | | 1 |
| g28571 | | | | 1 |
| g13043 | | | | 1 |
| g5011 | | | | 1 |
| g6659 | | | | 1 |
| g21748 | | | | 1 |
| g14239 | | | | 1 |
| g25657 | | | | 1 |
| g2998 | | | | 1 |
| g15818 | | | | 1 |
| g13694 | | | | 1 |
| g11005 | | | | 1 |
| g2389 | | | | 1 |
| g9149 | | | | 1 |
| g21172 | | | | 1 |
| g5385 | | | | 1 |
| g2363 | | | | 1 |
| g1623 | | | | 1 |
| g23227 | | | | 1 |
| g7608 | | | | 1 |
| g8220 | | | | 1 |
| g22389 | | | | 1 |
| g16363 | | | | 1 |

**Tabla C-2 Resultados: Génes dif exp muestra 1b.**

| Muestra 30000-20-500 | Reales | Limma | Deseq2 | Noiseq |
|---|---|---|---|---|
| g14328 | | | | 1 |
| g22978 | | | | 1 |
| g6282 | | | | 1 |
| g2475 | | | | 1 |
| g3310 | | | | 1 |

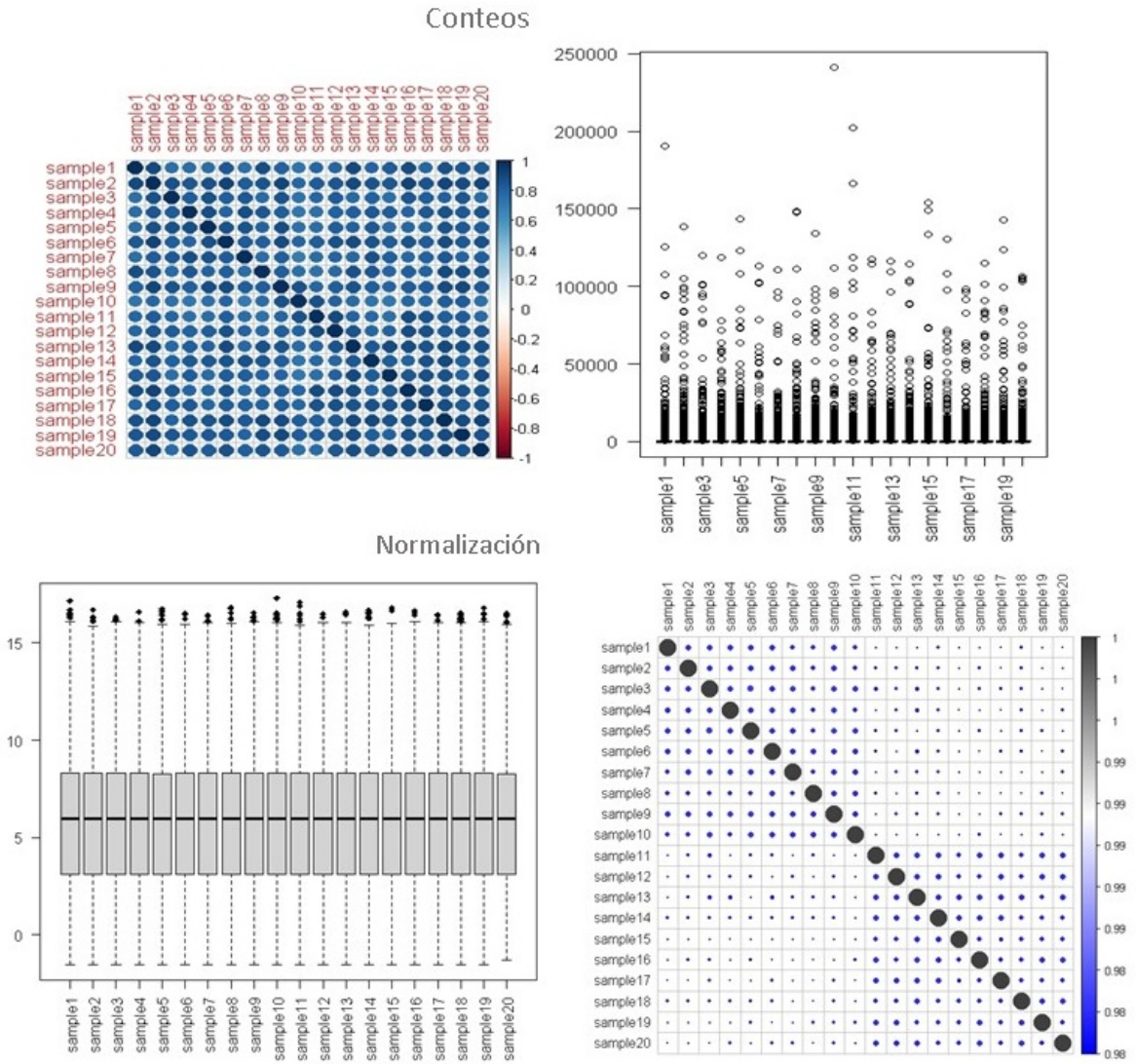



**Figura C-3**.: Gráficos de conteos y normalización de los conteos muestra1b

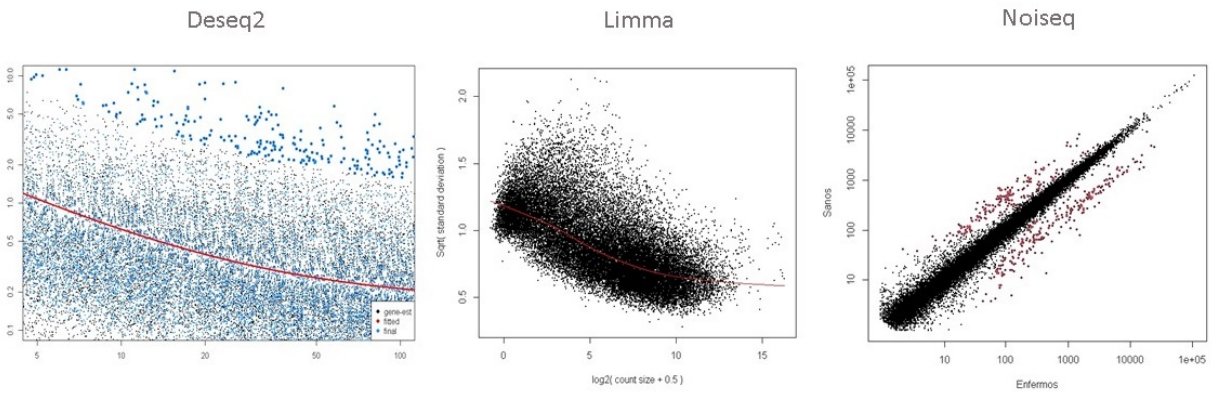



**Figura C-4**.: Resultado de los modelos muestra1b

# Bibliografía


[1] Aliferis, C. ; Hardin, D. ; Massion, P.: Machine learning models for lung cancer classification using array comparative genomic hybridization. En: *Annual Symposium AMIA.* (2002)

[2] Aurélien, Géron: *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.* Canada : OREILLY, 2019

[3] Bhattacharjee, A ; Richards, W G. ; Staunton, J ; Li, C ; et al.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. En: *Proc Natl Acad Sci U.S.A* (2001)

[4] Veiga de Cabo, J. ; De la fuente, E. ; Zimmermann, V. ; et al.: *Modelos de estudios en investigación aplicada: conceptos y criterios para el diseño.* Madrid, 2008

[5] Cadena, P. ; Rendón, R. ; et al.: *Quantitative methods, qualitative methods or combination of research: an approach in the social sciences.* México, 2017

[6] Chamorro, C. ; Merino, D.: *Análisis de datos de RNA-Seq empleando diferentes paquetes desarrollados dentro del proyecto Bioconductor para estudios de expresión génica diferencial.* Cataluya, 2019

[7] Chen, Y ; McCarthy, D ; Lun, A ; Zhou, X ; Robinson, M ; Smyth., G K.: edgeR Package Introduction. En: *Bioconductor* (2014)

[8] Creswell, J.: *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* 2008

[9] Fox, R. ; Dimmic, M. ; Trafford, A. ; Zhang, H. ; Kitmitto, A.: *A twosample Bayesian t-test for microarray data.* USA, 2006

[10] GK, Smyth: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. En: *Stat Appl Genet Mol Biol* 3 (2004), p. 1–25

[11] Han, H. ; Li, X.: *Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery.* Inchon, Korea, 2011

[12] Han, X.: *Nonnegative principal component analysis for cancer molecular pattern discovery.* China, 2010

[13] Jiménez, V. ; Vega, L.: *Flujo Bioinformático para el Análisis de Expresión Diferencial.* México, 2014

[14] Kim, B. ; et al.: *Clinical Validity of the Lung Cancer Biomarkers Identified by Bioinformatics Analysis of Public Expression Data.* Seoul,Korea, 2007

[15] L., Kline: *Introducción al análisis de datos de microarreglos: pre- procesamiento y manejo de datos de expresión en R.* Colombia, 2019

[16] Love, M.I. ; Huber, W. ; Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. En: *Ind. Eng. Chem. Res.* Genome Biol 15 (2014), p. 550

[17] Lu, W. ; Chen, L. ; Fu, D. ; Kong, X. ; et al.: *FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms.* China, 2020

[18] McDermaid, A ; Monier, B ; Zhao, J ; Liu, B ; Ma, Q.: *Interpretation of differential gene expression result of RNA-seq data: review and integration.* 2018

[19] PV, Nazarov ; A, Muller ; T, Kaoma ; et al., Nicot N.: RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. En: *BMC Genomics* (2017)

[20] Ritchie, Matthew E. ; Phipson, Belinda ; Wu, Di ; Hu, Yifang ; Law, Charity W. ; Shi, Wei ; Smyth, Gordon K. *limma powers differential expression analyses for RNA-sequencing and microarray studies.* 20 enero 2015

[21] Rodríguez, Cubillos A. ; Perlaza, Jimenez L. ; Bernal Giraldo, AJ.: *Analizando datos de RNA-Seq en procariotas: una revisión para no expertos.* Colombia, 2014

[22] S., Tarazona ; F., García ; A., Ferrer ; J., Dopazo ; A., Conesa. *NOIseq a RNA-seq differential expression method robust for sequencing depth biases.* 2012

[23] Sanchez, S.: *Análisis de datos de RNA-seq comparación de métodos para el estudio de expresión génica diferencial.* España, 2015

[24] Siegel, R. ; Ma, J. ; Zou, Z. ; Jemal, A.: *Cancer Statistics.* USA, 2014

[25] Smyth, Gordon K. ; Ritchie, Matthew ; Thorne, Natalie ; Wettenhall, James ; Shi, Wei ; Division, Yifang Hu B. ; Walter, The ; of Medical Research, Eliza Hall I. ; Melbourne ; Australia. *Linear Models for Microarray and RNA-Seq Data Userś Guide.* Noviembre 14, 2021

[26] Society, American C. *https://www.cancer.org/es/cancer/cancer-de-pulmon/.* Recuperado el 01 de octubre 2020

[27] Soneson ; Delorenz: generateSyntheticData: Generate synthetic count data sets. En: *R Documentation* (2013)

[28] Soneson, Charlotte. *Package compcodeR.* Enero 20,2022

[29] Tarazona, S. ; GarcÃa, F. ; Dopazo, J. ; Ferrer, A. ; Conesa, A.: *Differential expression in RNA-seq: A matter of depth.* Valencia, España, 2011

[30] Vanegas, L.: *Modelos Lineales Generalizados.* Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia., 2016

[31] Zararsis, G. ; et al.: *A comprehensive simulation study on classification of RNA-seq data.* 2017