UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE BIOTECNOLOGÍA

MODELAMIENTO MATEMÁTICO SOBRE LA EXPRESIÓN GENÉTICA PARA EL DIAGNÓSTICO DE SIBILANCIAS RECURRENTES ASOCIADAS A AEROALÉRGENOS

Trabajo de titulación previo a la obtención
del título de Ingeniero/a Biotecnólogo/a

AUTORES: ACCEL RAFAEL REYES CHÁVEZ
LESLIE PAULETTE SANTANDER RAMÓN
TUTOR: ING. EDMOND GÉRAUD, M.Sc.

Cuenca - Ecuador
2025

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Accel Rafael Reyes Chávez con documento de identificación N° 1104342892 y Leslie Paulette Santander Ramón con documento de identificación N° 0707065538; manifestamos que:

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

Cuenca, 21 de julio del 2025

Atentamente,

Accel Rafael Reyes Chávez
1104342892

Leslie Paulette Santander Ramón
0707065538

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

Nosotros, Accel Rafael Reyes Chávez con documento de identificación N° 1104342892 y Leslie Paulette Santander Ramón con documento de identificación N° 0707065538 expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Trabajo experimental: "Modelamiento matemático sobre la expresión genética para el diagnóstico de sibilancias recurrentes asociadas a aeroalérgenos", el cual ha sido desarrollado para optar por el título de: Ingeniero/a Biotecnólogo/a, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente

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

Cuenca, 21 de julio del 2025

Atentamente,

Accel Rafael Reyes Chávez
1104342892

Leslie Paulette Santander Ramón
0707065538

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Edmond Géraud con documento de identificación N° 0152387734, docente de la Universidad Politécnica Salesiana, declaró que bajo mi tutoría fue desarrollado el trabajo de titulación: MODELAMIENTO MATEMÁTICO SOBRE LA EXPRESIÓN GENÉTICA PARA EL DIAGNÓSTICO DE SIBILANCIAS RECURRENTES ASOCIADAS A AEROALÉRGENOS, realizado por Accel Rafael Reyes Chávez con documento de identificación N° 1104342892 y por Leslie Paulette Santander Ramón con documento de identificación N° 0707065538, obteniendo como resultado final el trabajo de titulación bajo la opción Trabajo experimental que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 21 de julio del 2025

Atentamente,

Ing. Edmond Géraud, M.Sc.
0152387734

DEDICATORIA

Dedico este trabajo a Dios y a su infinita sabiduría que han sabido colocarme en los lugares y momentos correctos, por presentarme las personas más importantes en mi vida, que han sido fuente de sabiduría y apoyo en los momentos más complicados y en los más chingones celebrar conmigo.

A mi mamá por su apoyo incondicional, por su gran sacrificio ya que es difícil ser una mujer soltera sacando adelante a sus hijos, por brindarme su amor y su paciencia que es infinita.

A mi hermano por apoyarme en todo y ser un gran pilar en el que me puedo apoyar a mi abuelita y a mi tío que han estado apoyándome y dándome consejos desde el día 1.

Accel

Dedico este trabajo a quienes han sido pilares en mi vida y acompañaron este camino de aprendizajes, desafíos y crecimiento personal.

A mis padres, por su amor incondicional, su esfuerzo incansable y sus valores que me sostienen día a día.

A mis hermanas, que, aunque la distancia nos separe, siempre han estado cerca con su cariño, sus mensajes, y su fuerza silenciosa.

A mi tía Patricia, por su apoyo constante, ya que ha sido un pilar fundamental en los días difíciles.

A mi gatita Skadi, que con su ternura y compañía me enseñó que incluso en el silencio y la rutina puede habitar la paz. Cada uno de sus ronroneos fue un consuelo en los momentos más tensos y solitarios. Aunque ya no esté conmigo, su recuerdo sigue siendo abrigo, y su amor silencioso permanece en cada rincón de mi memoria.

Leslie

AGRADECIMIENTO

Me gustaría agradecer a mi familia por a mi mamá Mirian por siempre apoyarme y no rendirse conmigo, a mi hermano Diego Andrés, por ser la persona que me aconseja y no deja que haga cosas fuera de lugar, mi abuelita Gloria, por siempre estar con nosotros dándonos ánimos a toda la familia, a mi tío Franklin por ser como un padre para mí y a todos los demás por darme el apoyo que me han brindado incondicionalmente durante todos estos años.

Quiero expresar mi más profundo agradecimiento al Ing. Edmond Géraud, mi tutor de tesis, por su valiosa guía, y sobre todo paciencia y compromiso a lo largo de todo este proceso académico.

A mis amigos de la universidad: Gaby, Leslie, Gustavo, Carlos, Mateo quienes no solo compartieron largas jornadas de trabajo, sino también risas, desvelos y motivación constante y eso que soy bien fastidiosos muchas gracias, a mis amigos de Loja que a este punto de la vida los considero mis hermanos que, aunque están lejos siguen dándome ánimos y ánimos de seguir adelante.

Accel

Quiero expresar mi más profundo agradecimiento al Ing. Edmond Géraud, mi tutor de tesis, por su valiosa guía, paciencia y compromiso a lo largo de todo este proceso académico.

Agradezco al universo y a la vida, por regalarme esta oportunidad de aprender, crecer y culminar una etapa tan importante. Por los caminos inesperados, por los encuentros valiosos, por los silencios fértiles y por todo lo vivido.

A mi familia, por su apoyo inquebrantable, por ser mi refugio en los momentos difíciles y mi motor para seguir adelante.

A mis amigos de la universidad: Gaby, Accel, Gustavo y Carlos, quienes no solo compartieron largas jornadas de trabajo, sino también risas, desvelos y motivación constante.

A mis amigos fuera de la universidad, Joseph y Marlon, por sus palabras de aliento y amistad genuina.

A mis amigos del Discord y del Dota, porque incluso entre partidas y risas digitales, supieron ser un respiro necesario y una conexión honesta en tiempos de carga y exigencia.

Gracias a todos los que, de una u otra forma, me tendieron la mano, creyeron en mí y me recordaron que no estaba sola en este camino.

Leslie

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

En el siguiente trabajo de titulación, se estudió la expresión genética de todo el transcriptoma de niños de edad preescolar con y sin sibilancias, con el objetivo de obtener un modelo matemático que permita la predicción de genes involucrados en distintas vías metabólicas y de distintas enfermedades los cuales están implicados en las sibilancias. Con tal de obtener dicho modelo, se realizó un análisis de expresión diferencial dada la condición de tener sibilancias y no tenerlas, con el fin de poder realizar un filtrado inicial de los genes más importantes, para luego poder realizar un modelo el cual permite abordar el problema supervisado de clasificación del cual se trata el estudio en cuestión. Este modelamiento matemático permitiría a un laboratorio, tener su propio panel de genes para diagnosticar sibilancias en niños preescolares.

El modelo fue evaluado en grupos de datos independientes, logrando un desempeño predictivo ideal. Finalmente, se llevó a cabo un estudio de enriquecimiento funcional (GO, KEGG, GSEA y DOSE) con el objetivo de detectar rutas metabólicas significativas asociadas a la fisiopatología de las sibilancias.

Palabras clave: RNA-seq, sibilancias recurrentes, aeroalérgenos, análisis de expresión diferencial, regresión lasso, validación cruzada, enriquecimiento funcional.

ABSTRACT

In the present degree project, the gene expression of the entire transcriptome of preschool children with and without wheezing was studied, with the objective of developing a mathematical model capable of predicting genes involved in various metabolic pathways and diseases associated with wheezing. To achieve this, a differential expression analysis was performed based on the condition of presenting or not presenting wheezing, in order to conduct an initial filtering of the most relevant genes. These selected genes were then used to build a supervised classification model, which is the central focus of this study.

This mathematical modeling approach could enable laboratories to develop their own gene panels for diagnosing wheezing in preschool children. The model was evaluated on independent datasets, achieving excellent predictive performance. Finally, a functional enrichment analysis (GO, KEGG, GSEA, and DOSE) was conducted to identify significant metabolic pathways associated with the pathophysiology of wheezing

Keywords: RNA-seq, recurrent wheezing, aeroallergens, differential expression analysis, lasso regression, cross-validation, functional enrichment.

CAPÍTULO I

INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

Las afecciones respiratorias son una de las causas más comunes de morbilidad infantil en todo el mundo, siendo las sibilancias y el asma patologías muy comunes y que afectan significativamente la calidad de vida de los niños. De acuerdo con la Organización Mundial de la Salud (OMS), existen más de 339 millones de individuos con asma, y un porcentaje significativo de los síntomas comienzan durante la infancia (OMS, 2024).

Las sibilancias, que se distinguen por un sonido silbante al respirar, son uno de los signos clínicos más relevantes de posibles afecciones respiratorias crónicas como la enfermedad ya mencionada. En numerosas situaciones, las sibilancias en la infancia son subdiagnosticadas o tratadas como infecciones respiratorias habituales, lo que demora un tratamiento apropiado.

En la ciudad de Cuenca, Ecuador, estudios locales han revelado cifras alarmantes sobre la frecuencia de este síntoma respiratorio. En una investigación realizada en niños de 2 a 5 años, se encontró que el 92,4% fueron positivos al API. Entre los criterios más frecuentes se encontraron las sibilancias no asociadas a resfriados (93,3%) y el diagnóstico de rinitis alérgica (85,7%) (Sempértegui & Bautista, 2020). Estos datos respaldan la necesidad de herramientas diagnósticas más específicas y moleculares que complementen los enfoques clínicos tradicionales y permitan un diagnóstico más preciso.

Estudios recientes han evidenciado el potencial del análisis de la expresión genética y la bioinformática en la identificación de biomarcadores en diversas enfermedades respiratorias alérgicas, tales como la rinitis alérgica y la neumonía, lo que sugiere la viabilidad de aplicar este enfoque al estudio de las sibilancias recurrentes (Han et al., 2015)

En este contexto, el análisis del transcriptoma mediante tecnologías de secuenciación de nueva generación (NGS) ofrece una alternativa innovadora para identificar patrones de expresión genética relacionados con las sibilancias recurrentes. Sin embargo, la gran cantidad de datos

generados por estas tecnologías requiere el desarrollo de herramientas de modelamiento matemático que permitan interpretar los resultados de manera eficiente. Por lo tanto, se plantea como problema central la necesidad de desarrollar un modelo matemático que, a partir de datos transcriptómicos, permita mejorar el diagnóstico de sibilancias recurrentes en niños sensibilizados a aeroalérgenos.

1.2 PREGUNTA DE INVESTIGACIÓN

¿A partir de la expresión genética se puede mejorar la precisión en el diagnóstico de sibilancias recurrentes asociadas a aeroalérgenos?

1.3 JUSTIFICACIÓN

Las sibilancias recurrentes en la infancia representan un problema de salud pública relevante, especialmente por su posible asociación con el desarrollo de asma en etapas posteriores. Según la Organización Mundial de la Salud, más de 262 millones de personas viven con asma, y una proporción importante presenta síntomas desde la infancia, aunque estos suelen ser sub diagnosticados en edades tempranas. Sin embargo, el diagnóstico en niños menores de seis años continúa siendo un desafío clínico debido a la limitada especificidad de las herramientas convencionales (OMS, 2024).

Investigaciones genéticas relacionadas con sibilancias mostraron una amplia gama de genes potenciales vinculados con el sistema inmunológico que podrían estar implicados en la patogénesis de la enfermedad. Yang et al (2007) evidenciaron que los niveles de IgE en la sangre del cordón umbilical, la edad masculina, el humo del cigarrillo de terceros y los antecedentes familiares de atopia son factores predictivos de sibilancias recurrentes.

Según Savenije et al. (2014), hallaron una relación con el polimorfismo CC10 G+38A de la proteína celular Clara y niveles inferiores de CC10 en niños que sufren esta condición. Otros marcadores genéticos fueron estudiados, como la vía IL33-IL1RL1, que ha sido vinculada con sibilancias intermedias, de inicio tardío y persistencia

. El aumento en el acceso a la tecnología NGS y los progresos en bioinformática y modelado matemático convierten este método en una estrategia actual y pertinente para tratar este complicado problema diagnóstico en la infancia (Sempértegui & Bautista, 2020).

En cuanto a sus aplicaciones prácticas, los hallazgos podrían contribuir al diseño de estrategias clínicas más eficaces y personalizadas para la detección y manejo temprano de niños con riesgo de desarrollar sibilancias, disminuyendo así las complicaciones a largo plazo y reduciendo la carga sobre los sistemas de salud (GINA, 2023).

Esta investigación se plantea con el propósito de contribuir al entendimiento de los mecanismos moleculares subyacentes a las sibilancias recurrentes, mediante el análisis del transcriptoma utilizando, en combinación con herramientas de la bioinformática.

1.5 LIMITACIONES

La búsqueda de métodos de diagnóstico más exactos para los episodios repetidos de sibilancias en niños de primera infancia, motiva la investigación de marcadores biológicos a nivel molecular a través del estudio de la expresión génica utilizando la secuenciación de nueva generación. No obstante, la puesta en práctica de esta línea de investigación se topa con determinadas restricciones propias del procedimiento y de los medios al alcance.

- El tiempo requerido para el procesamiento y análisis de los datos transcriptómicos disponibles es considerable, debido a la complejidad inherente a la alta dimensionalidad de este tipo de datos. Esto incluye la implementación de recursos computacionales avanzados para la normalización, análisis de expresión génica diferencial y el desarrollo, validación y evaluación de modelos matemáticos predictivos orientados al diagnóstico.
- **Distribución desigual de los datos:** El archivo analizado, proveniente de Fitzpatrick et al. (2024), contiene un mayor número de niños sin sensibilización a aeroalérgenos (n=36) en comparación con aquellos que sí presentan dicha condición (n=16). Esta disparidad introduce un sesgo potencial en los modelos de clasificación, al favorecer la predicción del grupo mayoritario, lo que puede dificultar la correcta identificación de los casos

clínicamente relevantes, es decir, niños con sibilancias recurrentes asociadas a sensibilización por aeroalérgenos.

Para mitigar este sesgo, se aplicó técnicas de balanceo de datos y métricas robustas frente al desbalance, como el F1-score. Sin embargo, se reconoce que el desbalance de clases podría conducir a problemas de sobreajuste o infraajuste del modelo, lo que representa una limitación metodológica que debe ser considerada al interpretar los hallazgos del estudio.

1.4 OBJETIVOS

1.4.1 OBJETIVO GENERAL

Desarrollar un modelo matemático basado en el análisis del transcriptoma obtenido para mejorar la precisión en el diagnóstico de sibilancias recurrentes asociadas a aeroalérgenos.

1.4.2 OBJETIVOS ESPECÍFICOS

- Validar la calidad y estructura de los datos transcriptómicos mediante un análisis de componentes principales (PCA) con el fin de asegurar la coherencia del conjunto de datos.
- Identificar genes diferencialmente expresados asociados a la presencia de sibilancias recurrentes mediadas por aeroalérgenos comprendiendo mejor la biología de la enfermedad.
- Construir un modelo matemático predictivo utilizando los genes diferencialmente expresados identificados en individuos con sibilancias recurrentes.
- Evaluar la capacidad del modelo matemático para diferenciar entre individuos con y sin sibilancias recurrentes asociadas a sensibilización por aeroalérgenos, con el fin de determinar su utilidad como herramienta diagnóstica.

1.6 HIPÓTESIS

El modelamiento matemático basado en el análisis del transcriptoma permite mejorar la precisión en el diagnóstico de sibilancias recurrentes asociadas a aeroalérgenos, identificando patrones diferenciales de expresión genética.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1 ESTADO DEL ARTE

Diversas investigaciones han abordado la complejidad de las sibilancias recurrentes en la infancia y su relación con el desarrollo de asma en etapas posteriores, utilizando enfoques clínicos, inmunológicos y moleculares. Estos antecedentes constituyen la base científica sobre la cual se estructura la presente investigación.

Zhang et al. (2020) realizaron un estudio de RNA-seq en 150 niños con sibilancias recurrentes, identificando 127 genes diferencialmente expresados, destacando la sobreexpresión de ORMDL3 y la subexpresión de IFN-γ como marcadores clave. Su principal aporte fue la identificación de una firma genética predictiva de progresión al asma con 82% de precisión. En el contexto de modelamiento matemático, Lee et al. (2021) desarrollaron un algoritmo de machine learning que integraba datos de expresión génica y variables ambientales para predecir exacerbaciones de sibilancias.

En Ecuador, un estudio realizado por Sempértegui & Bautista (2020) en el Hospital Monte Sinaí y el Hospital Militar de Cuenca analizó a 105 niños de entre 2 y 5 años con sibilancias recurrentes. Los autores encontraron que el 92.4% fueron API positivos, siendo las sibilancias no asociadas a resfriados el criterio más frecuente (93.3%), seguido de rinitis alérgica (85.7%). Este estudio demostró la alta prevalencia del riesgo de asma en preescolares, reforzando la necesidad de métodos diagnósticos tempranos y precisos.

Por otro lado, investigaciones recientes han comenzado a explorar el rol de la expresión génica y el transcriptoma en enfermedades respiratorias infantiles. Altman et al. (2022) identificaron que niños preescolares con sibilancias y sensibilización a aeroalérgenos presentan una respuesta disfuncional del interferón tipo I, asociada con una vía inmunológica mediada por IL-4, lo que sugiere un fenotipo inflamatorio particular.

En el ámbito bioinformático, Zhou et al. (2021) aplicaron análisis transcriptómicos y aprendizaje automático para identificar patrones moleculares en niños asmáticos, encontrando

biomarcadores genéticos específicos que podrían ser utilizados en la creación de algoritmos predictivos. Este tipo de enfoque multidisciplinario resalta la importancia del modelamiento matemático y la integración de datos ómicos para una medicina más personalizada.

2.3 MARCO TEÓRICO

2.3.1 Sibilancias recurrentes y sensibilización a aeroalérgenos

Las sibilancias, un sonido respiratorio agudo y musical, común en la obstrucción de las vías respiratorias, han sido investigadas para su relevancia diagnóstica en el asma en niños. Estudios como el realizado por Castro-Rodríguez et al. (2010) en el Journal of Allergy and Clinical Immunology indican que, aunque esta condición es muy frecuente, su capacidad para predecir con certeza qué niños desarrollarán broncoespasmos es limitada. Por otro lado, Paleari et al. (2016) en Respiratory Medicine proponen que los patrones de sibilancias podrían señalar distintas respuestas al tratamiento, subrayando la importancia de una evaluación clínica completa, tal como destacan las directrices GINA (2023) y GOLD (2023). Además, el desarrollo de herramientas de análisis acústico, como se explora en el estudio de Tapia et al. (2018) en BMC Pulmonary Medicine, busca mejorar la objetividad en el diagnóstico y seguimiento de este síntoma respiratorio.

La patogénesis compleja de las sibilancias en niños, involucra la interacción entre genética, ambiente y desarrollo inmunológico, según Bacharier et al. (2012) en The Lancet, subraya la necesidad de comprender los mecanismos subyacentes para mejorar la prevención y el tratamiento, cuyo enfoque actual se centra en el control de síntomas y la prevención de exacerbaciones mediante farmacoterapia según Busse et al., 2018, New England Journal of Medicine y el manejo de factores ambientales, según las directrices de GINA (2023).

En este contexto, los aeroalérgenos, que son antígenos transportados por el aire, están presentes en ambientes interiores como exteriores y pueden inducir una respuesta inmunitaria en individuos susceptibles. Esta respuesta se caracteriza por la producción de inmunoglobulina E (IgE), un tipo de anticuerpo codificado por el gen *IGHε,* localizado en el brazo largo del cromosoma 14 (14q32.33). Su ubicuidad en ambientes interiores y exteriores, incluyendo pólenes, ácaros, epitelio animal y esporas de hongos, se asocia fuertemente con la sensibilización alérgica y el riesgo de asma, como señala el estudio de Platts-Mills et al. (1997) en el New England Journal of

Medicine. La exposición continua a estos alérgenos puede exacerbar la inflamación de las vías aéreas y la persistencia de los síntomas, como revisan Bousquet et al. (2001) en el Journal of Allergy and Clinical Immunology.

La interacción de los aeroalérgenos con el sistema inmunitario, activando células presentadoras de antígenos y promoviendo la diferenciación de linfocitos Th2 y la producción de citocinas proinflamatorias, que son proteínas pequeñas generadas por diversas células, especialmente del sistema inmune, y actúan como transmisores químicos que regulan la respuesta inflamatoria e inmunológica. (Akdis, 2006, Nature Reviews Immunology), las cuales son clave para el diagnóstico molecular y la inmunoterapia.

El entendimiento de la patobiología de las alergias provocadas por aeroalérgenos se amplía con la investigación de la expresión genética, en la que la exposición a dichos alérgenos provoca alteraciones importantes en las células inmunológicas y epiteliales de las vías respiratorias. Esto se evidencia en el estudio de perfiles de expresión genética en células epiteliales nasales de pacientes con rinitis alérgica expuestos a ácaros en polvo según Lee et al., 2006, Journal of Allergy and Clinical Immunology.

Estos cambios en la expresión génica dan lugar a biomarcadores moleculares con potencial diagnóstico y terapéutico. La identificación de firmas de expresión génica específicas, como se explora en el estudio de "inmunogramas" para predecir la respuesta a la inmunoterapia (Banchereau et al., 2011, Immunity) y la búsqueda de biomarcadores no invasivos en muestras de las vías aéreas (Djukanović et al., 2002, American Journal of Respiratory and Critical Care Medicine), prometen mejorar el manejo de las alergias.

Finalmente, la Secuenciación de Nueva Generación (NGS) ha revolucionado la identificación de estos biomarcadores moleculares al permitir el análisis de genomas, transcriptomas y epigenomas con alta resolución. En el contexto de las alergias, la NGS ha facilitado la identificación de variaciones genéticas asociadas a la susceptibilidad (Moffatt et al., 2010, Nature) y la caracterización detallada de la expresión génica en respuesta a la exposición a alérgenos mediante RNA-seq (Wang et al., 2009, Nature Reviews Genetic). La capacidad de generar grandes cantidades de datos moleculares ha impulsado el descubrimiento de biomarcadores

diagnósticos y predictivos, incluyendo el análisis de repertorios de receptores inmunitarios y modificaciones epigenéticas.

2.3.2 Bioinformática y análisis diferencial

El análisis de expresión diferencial es fundamental en transcriptómica, ya que permite identificar genes que muestran variaciones significativas en sus niveles de expresión entre distintos grupos o condiciones experimentales (Love et al., 2014). Estas diferencias pueden revelar rutas biológicas alteradas, identificar biomarcadores y mejorar la comprensión de mecanismos moleculares asociados a las sibilancias y posteriormente a enfermedades como el asma.

DESeq2 es un paquete que se encuentra en R, el cual es usado para el estudio de la expresión genética diferencial, diseñado para datos de RNA-Seq, el cual utiliza un modelo de regresión binomial negativa, método estadístico empleado para modelar datos de conteo (como el número de casos, expresiones genéticas, etc.) en situaciones de sobre dispersión, o sea, cuando la varianza supera la media. Esta circunstancia infringe una premisa fundamental de la regresión de Poisson, en la que se presupone que la media y la varianza son equivalentes. Por lo tanto, se considera que la binomial negativa es una ampliación del modelo de Poisson, creado para gestionar esta sobre dispersión (Love et al., 2014). Esta función ofrece técnicas estadísticas robustas para gestionar tamaños de muestra reducidos y variabilidad en la información, proporcionando resultados fiables. El resultado presenta listados de genes que han experimentado alteraciones importantes en la expresión, junto con valores estadísticos relacionados como la variación del pliegue y los valores p.

Los genes expresados diferencialmente (DEG) hacen referencia a genes cuyos niveles de expresión se incrementan o reducen considerablemente entre distintas condiciones o grupos. Varios softwares se han implementado para identificar genes que se expresan de manera diferencial, algunos ejemplos ampliamente reconocidos son edgeR, DESeq2 y limma. (BxINFO, 2024)

El fold change o también conocido como razón de cambio es un indicador de la magnitud de la variación en la expresión de un gen entre dos condiciones.

$$Fold\ Change\ (FC)\ = \frac{Expresión\ en\ condición\ A}{Expresión\ en\ condición\ B}$$

Donde:

- Condición A puede ser la experimental (por ejemplo, enfermo, tratado, etc.) Condición B puede ser la de referencia o control (ejemplo, sano, no tratado, etc.)

A partir de la razón de cambio y del error estándar de ésta, se puede obtener el pvalor asociado, el cual nos permite conocer la significancia estadística, aparte de conocer el efecto o significancia biológica y con él, la razón de cambio. Dado que se computa un modelo por cada gen, el número total de pruebas estadísticas realizadas es elevado, lo que incrementa drásticamente la tasa de falsos positivos.

El análisis de expresión diferencial calcula la variación en la expresión genética a través de un modelo de regresión binomial negativa, en el que el Fold Change (FC) entre dos condiciones se establece como el cociente entre sus medias de expresión normalizadas.

$$FC\ = \frac{\mu 1}{\mu 2}\ \ ;\ \ log_2\ (FC) = \ log_2(\frac{\mu 1}{\mu 2})$$

Estas medias son ajustadas mediante factores de normalización y modeladas en escala log mediante:

$$log\ (\mu ij)\ = log\ (sj) + xj \cdot \beta i$$

Donde:

- sj es el factor de normalización para la muestra j
- xj representa la condición experimental (por ejemplo, control o tratamiento)
- βi es el coeficiente que estima el logaritmo en base 2 del Fold Change para el gen i

Este modelo calcula un valor p para cada gen, determinando si el coeficiente βi difiere significativamente de cero. Por esta razón, es necesario aplicar una corrección a los valores *p*, con el fin de controlar los descubrimientos erróneos. Uno de los métodos más aceptados para este ajuste

es la corrección por tasa de falsos descubrimientos (FDR, por sus siglas en inglés), también conocida como el método de Benjamini y Hochberg (BioDatev, 2023)

El análisis de componentes principales (PCA), disminuye la cantidad de dimensiones en grandes volúmenes de datos a unas nuevas variables producto de la combinación lineal de las variables originales proyectándolas a un espacio de dimensión menor, pero conservando la mayoría de la información original (varianza) principales que mantienen la mayoría de la información original. Para conseguir esto, convierte las variables que podrían estar correlacionadas en un grupo más reducido de variables, conocidos como componentes principales.

El PCA se emplea frecuentemente en el preprocesamiento de datos para su utilización en algoritmos de aprendizaje automático (IBM, 2023). Esto disminuye la complejidad del modelo, dado que la incorporación de cada nueva característica impacta de manera negativa en el desempeño del modelo, lo cual también se denomina frecuentemente como la "maldición de la dimensionalidad".

En este tipo de análisis es muy común observar outliers (en español valor atípico), son esos valores rebeldes que se alejan significativamente del comportamiento global de tus datos, básicamente se caracterizan por apartarse considerablemente del resto, desmantelar los patrones generales que corresponden con tus datos y ser potencialmente legítimos o fallos y por ende, reconocerlos es un reto (Lead Up Collective, 2017).

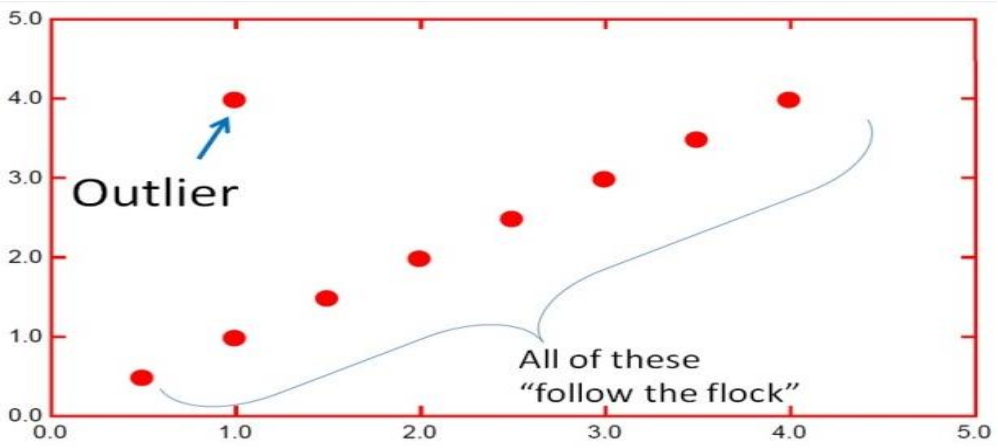


Figura 1. Statistics & High Performers: Studying the Outliers**.** ***Fuente:*** *Lead Up Collective, (2017).*

Para estos valores atípicos, existe la distancia de mahalanobis el cual nos permite eliminar estos datos, básicamente determina cuánto está aislado un punto del núcleo de una distribución multivariada mediante una matriz de covarianza. En R, puede asistir en la identificación de valores inusuales al comparar la distancia al cuadrado de cada punto con un límite de chi cuadrado, reconociendo aquellos puntos que se encuentran más lejos de los datos principales (Cansiz, 2023).

$$D^2 = \left(X_{p_1} - X_{p_2}\right)^T . C^{-1} . \left(X_{p_1} - X_{p_2}\right)$$

Figura 2. Fórmula de la distancia de Mahalanobis. ***Fuente:*** Sergen Cansiz (*2023).*

La distancia de Mahalanobis permite medir la distancia entre un punto y un conjunto de datos, teniendo en cuenta la correlación entre las variables a través de la matriz de covarianza. Como se puede apreciar en la figura 2, Mahalanobis emplea la inversa de la matriz de covarianza, o matriz de precisión de los componentes principales a estudiar, que suelen ser dos. El uso de la matriz de precisión nos permite tener en cuenta la estructura interna de los datos y nos permite identificar los valores atípicos en datos multivariantes. Esta función se encuentra presente en el paquete de R,

2.3.3 Modelado estadístico y métodos de regularización

La Regresión logística binomial se emplea para modelar una variable de respuesta binaria a partir de variables predictoras, y puede ser aplicado a problemas de múltiples clases, cuando la función de enlace es multinomial. No obstante, hay un pero, la suposición principal de la regresión logística es que los datos deben de ser independientes, en biología, esto nunca pasa, por lo que se debe de obtener técnicas más robustas frente a la colinealidad. Por lo tanto, se procede a insertar métodos de regularización o penalización. Esta metodología incluye un concepto de penalización (regularización) lo que posibilita regular la cantidad de coeficientes, prevenir el sobreajuste y elegir las variables más pertinentes, especialmente beneficioso en grupos de datos con numerosas variables o relaciones complejas (Akalin, 2020).

Existen tres tipos de regularización, la primera se basa en contraer los coeficientes colineales a cero, pero sin llegar a este valor. Por otro lado, otro tipo de penalización se conoce como Lasso, o L1, la cual directamente las variables colineales se hacen cero. Finalmente tenemos elastic net, el cual es una mezcla de ambos.

La regresión de Ridge, o también conocido como L2, es un método de normalización en el ámbito estadístico, el cual rectifica el exceso de ajustes en los datos de entrenamiento en los modelos de machine learning. Este método resulta beneficioso cuando se desarrollan modelos que contienen una gran cantidad de parámetros, especialmente si estos poseen pesos significativos (Murel & Kavlakoglu, 2025).

La regresión Lasso (Operador de Disminución Absoluta y Selección), también llamada regularización L1, es un método estadístico que mejora la exactitud de los modelos y evita el sobreajuste al incluir una penalización en la adición de los valores absolutos de los coeficientes. Este concepto de penalización, regulado por un parámetro λ (lambda), posibilita la disminución de ciertos coeficientes a cero, simplificando la elección automática de variables, lo cual resulta particularmente beneficioso en situaciones de alta dimensión, dado que contribuye a simplificar los modelos al mantener solo las variables más pertinentes (IBM, 2025).

Cabe recalcar que Lasso y Ridge aumentan la complejidad del modelo, aunque utilizando métodos distintos. La regresión Lasso disminuye la cantidad de variables autónomas que influyen en la salida, en cambio Ridge disminuye el peso de cada variable independiente en la salida.

Con tal de obtener los mejores parámetros de regularización, la validación cruzada es crucial en el aprendizaje automático, empleado para valorar el rendimiento de un modelo. El propósito principal es asegurar que el modelo sobreajuste a los datos de entrenamiento y que generalice correctamente con datos no vistos por el modelo. La validación cruzada conlleva la separación del conjunto de datos en varios subconjuntos, la formación del modelo en ciertos subconjuntos y su prueba en los subconjuntos restantes (Geeksforgeeks, 2025).

2.3.4 Enriquecimiento funcional y biología de sistemas

Para interpretar los datos de expresión génica de una manera más completa es necesario utilizar el análisis de enriquecimiento con el fin de extraer su información biológica, para así comprender mejor significancia. Para ello tenemos algunas metodologías de enriquecimiento funcional como: Singular Enrichment Analysis (SEA), Gene Set Enrichment Analysis (GSEA) y Modular Enrichment Analysis (MEA) (Garcia-Moreno et al., 2022), en este estudio solo nos vamos a concentrar en los dos primeros.

Antes de especificar en los análisis de enriquecimiento es necesario hablar de los lugares de donde se accede a la información para llevar a cabo estos, los cuales hacen posible determinar las vías biológicas y los genes que la componen llamados Bases de datos (García, 2023), existen varias, pero nos concentraremos en Gene Ontology y KEGG como las más importantes.

Gene Ontology (GO), es una base de datos creada por Gene Ontology Consortium, estos desarrollaron "un conjunto conocido como ontologías, para describir dominios clave de la biología molecular," (Harris et al., 2004), se subdivide en 3 ontologías:

- Función Molecular (MF): detalla las funciones de los genes a nivel molecular.
- Proceso biológico (BP): indica que actividades celulares son llevadas por un gen o grupo de genes.
- Componente celular (CC): muestra la ubicación del gen o grupo de genes ejercen la función.

GO también tiene una parte importante las cuales se llaman (Anotaciones GO), son esenciales para el análisis de enriquecimiento, ya que con ellas se puede comparar un gen con las diferentes ontologías genéticas, siendo respaldada por la literatura científica subida a esta base de datos. (García, 2023)

KEGG (Kyoto Encyclopedia of Genes and Genomes), base de datos que facilita las vías en su mayoría son rutas metabólicas, también participa en el análisis de procesos biológicos que se relacionan con funciones sistemáticas más complejas dentro del organismo, tiene dos repositorios

fundamentales que son ’Pathway’ y ’Genes’.

’Pathway’ contiene una amplia base de diagramas de vías bioquímicas y redes de señalización en formato interactivo y ’Genes’ que cuenta con información de varios genomas y un sistema de identificación llamado Ortología KEGG o (KO KEGG Orthology) por sus siglas en inglés, que asigna un número tras analizar genes o proteínas que los vinculan a una ruta de vías o redes de señalización (Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, & Mari Ishiguro-Watanabe, 2025).

REACTOME, contiene información de vías biológicas humanas, su enfoque abarca el análisis de proteínas, moléculas pequeñas, macromoléculas y genes que permiten identificar vías enriquecidas, por esto la convierte en una herramienta de identificación de gran precisión, se debe a que expertos hacen una selección de información a mano y de manera selectiva (Croft et al., 2011).

A continuación, se mencionan las metodologías de enriquecimiento:

Simple Enrichment Analysis (SEA), también conocido como ORA (Over Representation Analysis), es la metodología de enriquecimiento más antigua, se caracteriza por que analiza un gen a la vez de una lista proporcionada con anterioridad los cuales tiene que discretizar, después se compara con una base de datos como la de Gene Ontology y sus subontologías, utilizando el test de fisher, método eficiente y fácil de interpretar pero también tiene puntos débiles como que al tener un gran número de genes o un bajo número de ellos puede generar un sesgo y los resultados pueden alejarse de la significación biológica, otro punto es que, al discretizar los datos se pierde información, ya que se está asignando una condición en específico para clasificarlos. (Tipney & Hunter, 2010)

GSEA utiliza una medida cuantitativa para ordenar los genes en donde utiliza el Fold Change de la lista de genes diferencialmente expresados para ordenar los genes en un ranking de manera aleatoria, después se utiliza el test de Kolmogorov Smirnov para identificar los que están significativamente enriquecidos, una de sus limitaciones es que no toma en cuenta las relaciones que existen entre vías biológicas. (García, 2023)

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 NIVEL DE INVESTIGACIÓN

En la presente investigación, se adopta un enfoque que abarca tanto el nivel descriptivo, explicativo, inferencial y predictivo. Inicialmente, se realizará un análisis descriptivo e inferencial detallado de los datos transcriptómicos obtenidos mediante NGS en los grupos de individuos con y sin sibilancias recurrentes asociadas a aeroalérgenos.

Esta fase descriptiva e inferencial se centrará en caracterizar y comparar los perfiles de expresión genética entre ambos grupos, con la meta de identificar y contrastar las particularidades de la expresión génica y señalar los genes que exhiben niveles de actividad distintos. Posteriormente, la investigación avanzará hacia un nivel explicativo y predictivo. En esta etapa, se investigará si estos patrones genéticos, analizados mediante modelamiento matemático, permiten explicar y predecir con mayor precisión el diagnóstico de la condición. El estudio se enfocará sistemáticamente en identificar biomarcadores diagnósticos con valor predictivo a partir del análisis transcriptómico.

3.2 DISEÑO DE INVESTIGACIÓN

Este trabajo de titulación implementa un diseño observacional y analítico, caracterizado por el análisis de datos del transcriptoma en su forma existente, sin manipulación activa de variables, a diferencia de un estudio experimental. El enfoque principal es observar y analizar las diferencias naturales en la expresión genética entre individuos con y sin sibilancias recurrentes asociadas a aeroalérgenos, examinando la relación expresión génica-fenotipo en un contexto biológico real y evitando sesgos artificiales. Aunque los datos se almacenan electrónicamente, este diseño va más allá de la investigación documental, centrándose en el análisis profundo de los datos transcriptómicos mediante herramientas bioinformáticas y modelos matemáticos para identificar patrones, correlaciones y relaciones predictivas, y generar nuevos conocimientos sobre las bases moleculares de la enfermedad.

3.3 POBLACIÓN Y MUESTRA

La población en estudio está conformada por niños en edad preescolar, entre 12 y 59 meses, que presentan sibilancias recurrentes, con o sin sensibilización a aeroalérgenos, atendidos en el *Children's Healthcare of Atlanta*. Se excluyeron aquellos con trastornos comórbidos relacionados con las sibilancias, retraso significativo del desarrollo o fallo de medro.

La muestra analizada corresponde a un subconjunto de dicha población, compuesto por 52 niños cuyos datos transcriptómicos fueron obtenidos mediante secuenciación de nueva generación (NGS) y analizados en el estudio de Fitzpatrick et al. (2024). De estos, 36 niños no presentaban sensibilización a aeroalérgenos, mientras que 16 sí la tenían.

3.4 VARIABLES

Variables	Dimensiones	Unidades
Fenotipo de sibilancias recurrentes	44 (No o 0) 90 (Si 0 1) x 1	Presencia/Ausencia: 0 o No: Ausencia de sibilancias recurrentes asociadas a aeroalérgenos. 1 o Sí: Presencia de sibilancias recurrentes asociadas a aeroalérgenos.
Transcriptoma	134 x los transcritos que hay (61905)	Recuentos de lecturas (Reads): La unidad fundamental generada por la secuenciación de ARN (RNA-Seq) es el número de lecturas que se alinean a cada gen o transcrito. Número de reads por gen/transcrito.

Tabla 1. Variables. ***Fuente:*** *Propia autoría*

3.5 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

Esta investigación emplea una metodología que combina principalmente el análisis documental con un enfoque bioinformático. Se llevará a cabo una revisión bibliográfica para recopilar información de estudios previos sobre los mecanismos moleculares subyacentes a las

sibilancias recurrentes, la metodología de secuenciación de ARN (RNA-seq), el análisis bioinformático de datos transcriptómicos y la aplicación del modelado matemático en el diagnóstico. Este proceso implica la búsqueda en bases de datos académicas destacadas, incluyendo repositorios universitarios, Scopus, Scielo, ProQuest y Latindex y la gestión de las referencias bibliográficas se realizará mediante programas como Zotero y Mendeley.

3.6 TÉCNICAS DE PROCESAMIENTO Y ANÁLISIS DE DATOS

Los datos de RNA-seq fueron procesados mediante un flujo de trabajo bioinformático estandarizado, que incluye el control de calidad de las lecturas con formato tabular, los datos ya se encontraban alineados y se disponían de conteos, por lo que se procedió a normalizarlos para realizar el DGE. Los conteos se normalizaron a través del cálculo de factores de escala para cada muestra, que rectifican las variaciones en la profundidad de secuenciación y facilitan la comparación entre condiciones experimentales. Estos factores se derivan de la media de las relaciones entre los conteos de cada gen y una pseudo-referencia generada a partir del total de las muestras. El análisis exploratorio de datos se realizó con PCA, lo cual permitió realizar un control de calidad de los datos, seguido del análisis de expresión diferencial con DESeq2, corrigiendo para múltiples comparaciones. Las vías biológicas relevantes se identificaron mediante análisis de enriquecimiento. Para el modelado matemático del diagnóstico, se empleó la selección de características para reducir la dimensionalidad, modelos de clasificación y validación cruzada. Todo el análisis se llevó a cabo en el software R como el repositorio Bioconductor para el análisis de RNA-seq.

3.7 PROTOCOLO PARA IMPLEMENTAR

Los datos transcriptómicos empleados en este estudio fueron obtenidos del repositorio Gene Expression Omnibus (GEO) con número de acceso GSE261070 (Fitzpatrick et al., 2024). Se llevó a cabo la cuantificación de la expresión génica para obtener una matriz de conteos de lecturas por transcritos, que fueron agrupados por genes. Para la selección de variables se probaron varios métodos,entre ellos,por la contribución de las variables del PCA, que indican la importancia y la variabilidad explicada de cada transcrito/gen. De igual manera, se utilizó regresión por mínimos cuadrados parciales discriminantes, que a diferencia del PCA, no maximiza la varianza, sino la

covarianza. También se estudiaron las cargas o "loadings" de las variables latentes del PLSDA, para la misma selección de variables a distintos umbrales de la importancia de los genes respecto a su capacidad discriminante de los sujetos con sibilancias y sin sibilancias. Estos procedimientos constituyeron una base fundamental para el posterior desarrollo del modelo matemático predictivo, con tal de evitar genes redundantes y colinealidades, el cual representa el objetivo principal de esta tesis.

Se construyeron modelos de clasificación utilizando los algoritmos de regresión logística regularizada. Es decir, a la regresión logística binomial, se le agregan ciertos hiperparámetros los cuales son aprendidos por los mismos datos y optimizados mediante validación cruzada k-fold (k=5) en el conjunto de entrenamiento. El rendimiento se evaluó en el conjunto de prueba utilizando las métricas de precisión, exhaustividad, F1-score.

Con tal de evitar el sobreajuste, como se ha mencionado anteriormente, los hiperparametros fueron aprendidos por validación cruzada. Por la naturaleza de los datos, al ser genes, es normal suponer que, aunque se evite la colinealidad, siempre habrá, por lo que se implementó la regresión logística regularizada por lasso, con varias repeticiones cambiando la semilla de aleatorización, con el fin de encontrar los biomarcadores más representativos de la selección de variables final, los cuales nos dirán las vías metabólicas finales.

3.7.1 Preparación y preprocesamiento de datos

Como ya se mencionó anteriormente, se utilizó el archivo tabulado que contenía la matriz de conteos originales del conjunto de datos GSE261070, de la base de datos Geo Data Sets del repositorio NCBI. Esta matriz tiene los identificadores de genes como filas y los nombres de las muestras como columnas, con valores que indican el número de lecturas asignadas a cada gen por muestra. Se eliminaron los valores ausentes y se reestructuró el conjunto de datos para que los identificadores de los genes se convirtieran en los nombres de las filas.

Luego de eso, se establecieron dos grupos experimentales: **IC**: niños con infecciones virales (Infection-Control) y **PW**: niños sensibilizados a aeroalérgenos (Pre-Wheeze). A partir de esto, se empleó el conjunto de datos limpio para generar la matriz de conteos (genes × muestras). Cabe recalcar, que el buen preprocesamiento de datos asegura que los análisis posteriores puedan ser

válidos y reproducibles.

3.7.2 Análisis de expresión diferencial

Se aplicó un diseño sin intercepto (~ 0 + condición) que facilitó la comparación directa entre los dos grupos de investigación (IC, PW). Esto simplifica la comprensión de los resultados al identificar efectos particulares para cada situación y diferencias establecidas entre ambas. Previo al estudio estadístico, se realizó un pre filtrado de genes con el propósito de eliminar aquellos genes con menos variabilidad, de los cuales se realiza la suposición que éstos genes no tan variables entre muestras no regularán las condiciones de intereses que podrían generar irregularidades en los resultados. Los genes que mostraban al menos 5 lecturas normalizadas en un mínimo de 3 muestras fueron excluidos. Este criterio garantiza la inclusión de genes con suficiente variabilidad para realizar un adecuado análisis de expresión diferencial genética.

El análisis de expresión diferencial se realizó utilizando el flujo de trabajo del paquete DESeq2, con el cual se obtienen los estadísticos de la expresión diferencial (DE). Entre los principales resultados generados por este pipeline se encuentran: la expresión media por condición para cada gen, la razón de cambio (log2FoldChange o logFC) y su error estándar. A partir de estos valores, se calcula el p-valor, que permite evaluar la significancia estadística de la diferencia observada. La razón de cambio también se emplea para determinar la significancia biológica, permitiendo clasificar a los genes como sobreexpresados (UP) o infraexpresados (DOWN), según su comportamiento entre condiciones. Dado que se realizan tantas pruebas como se analizan, existe un riesgo elevado de obtener falsos descubrimientos. Por esta razón, los p-valores fueron ajustados mediante el método de Benjamini-Hochberg (BH) para controlar la tasa de descubrimiento falso (FDR).A parte de pre filtrados anteriormente comentados por PCA y PLSDA, también se filtraron aquellos genes que tenían una diferencia de razones biológicamente y estadísticamente diferentes entre ambas condiciones. Se calcularon parámetros como log2FoldChange y el p-valor ajustado (padj). Los valores por los cuales se asignaron la significancia biológica y estadística fueron los siguientes:

- padj < 0.01 para asegurar una confianza estadística del 99.9 %
- log2FoldChange| > 2 para considerar solo aquellos genes con un cambio de expresión

biológicamente relevante, tanto genes sobre regulados como infra regulados.

3.7.3 Detección y remoción de outliers

La primera acción antes de tratar los datos fue la realización del control de calidad con PCA. Esta técnica, permitió identificar dos valores atípicos visualmente hablando. En relación al análisis transcriptómico, se sabe que un outlier es una muestra cuyo perfil de expresión genética se diferencia significativamente del resto del conjunto de datos, lo que podría ser resultado de fallos en la recopilación de los datos o porque en realidad el dato que se observa es un valor extremadamente alto. La existencia de estas observaciones atípicas puede distorsionar los resultados del análisis diferencial, por lo que es de suma importancia identificarlos y suprimirlos antes de efectuar deducciones estadísticas.

Por ende, para poder identificar estas muestra atípicas, con total fiabilidad, a partir de los scores de las dos primeras componentes principales del PCA,s (PC1 y PC2) sobre los datos normalizados log-transformados, escalados a varianza unidad y centrados en media cero, se procedió a computar distancia de Mahalanobis donde se definió un límite del 99.9% como valor de corte basado en la distribución chi cuadrada con dos grados de libertad (PC1 y PC2), se pudieron identificar cuantitativamente los valores atípicos observados visualmente.

3.7.4 Balanceo de condiciones experimentales

El equilibrio de condiciones experimentales es un paso crucial en el procesamiento de datos cuando hay una disparidad en la cantidad de muestras entre los grupos. En esta investigación, después de eliminar los outliers correspondientes, se detectó un desequilibrio entre las muestras pertenecientes a los grupos "PW" e "IC”, lo que podría generar sesgos estadísticos en los modelos de clasificación que se pretenden aplicar.

Por tal motivo, para evitar el sesgo y garantizar una evaluación justa del desempeño del modelo, fue necesario igualar el número de muestras por condición, donde se reconoció al grupo "PW" como el predominante. Para disminuir su tamaño al nivel del grupo "IC", se escogieron de manera aleatoria 42 muestras de "PW" a través submuestreo aleatorio, pero reproducible,

garantizando una distribución balanceada. Tras el submuestreo aleatorio de la clase minoritaria, se realizó el pre-filtrado de la selección de variables, con el mismo fin de eliminar genes de baja expresión, garantizando así una base firme para el análisis de expresión diferencial y el modelado matemático posterior.

3.7.5 Selección de genes por PCA

Cómo se ha mencionado, en este estudio ha sido crucial el Análisis de Componentes Principales (PCA), como control de calidad y pre filtrado de variables menos importantes. El PCA, es una de las técnicas de análisis multivariado ideal para aminorar el número de dimensiones. El principio matemático de esta técnica, es encontrar un nuevo sistema de coordenadas, en base a la combinación lineal de las variables originales, tal que se maximice la varianza. De manera que este nuevo sistema de coordenadas, denominado componentes principales, permite reducir la dimensionalidad original, y comprimirla en unas nuevas variables compuestas por cargas y puntajes, tal que podemos representar los datos en gráficos de dos a 3 dimensiones. Normalmente el PCA va a acompañado del gráfico denominado Scree Plot, donde se pueden observar como la información de la dispersión se ve organizada en estas nuevas variables, escogiendo principalmente aquellos componentes que visualmente, clasifiquen o distingan las condiciones de interés. Esto nos permite escoger aquellas variables que mejor expliquen las condiciones de interés.

3.7.6 Construcción de matriz para modelado

Antes de realizar el modelo matemático se necesita realizar una serie de pasos fundamentales para la construcción y evaluación del mismo. Se utilizaron los datos normalizados y balanceados (42 muestras IC y 42 muestras PW), de los genes más importantes ya seleccionados en una matriz.

$X = n \times p$

Donde:

- **n**: son las filas que en este estudio serían las muestras (84).
- **p**: son las columnas que en este caso son el número de muestras (3981)

También se colocó una semilla para generar números de forma aleatoria que ayuda a la reproducibilidad del modelo matemático, en este caso mediante la regularización de la regresión logística LASSO.

Para el entrenamiento del modelo se utiliza el 80% de los datos y para la evaluación del mismo se utiliza el 20%. Dentro de los datos de entrenamiento, es decir del 80%, se aprendieron los mejores hiperparametros mediante validación cruzada y validando el modelo con el 20%. (Figura 1)

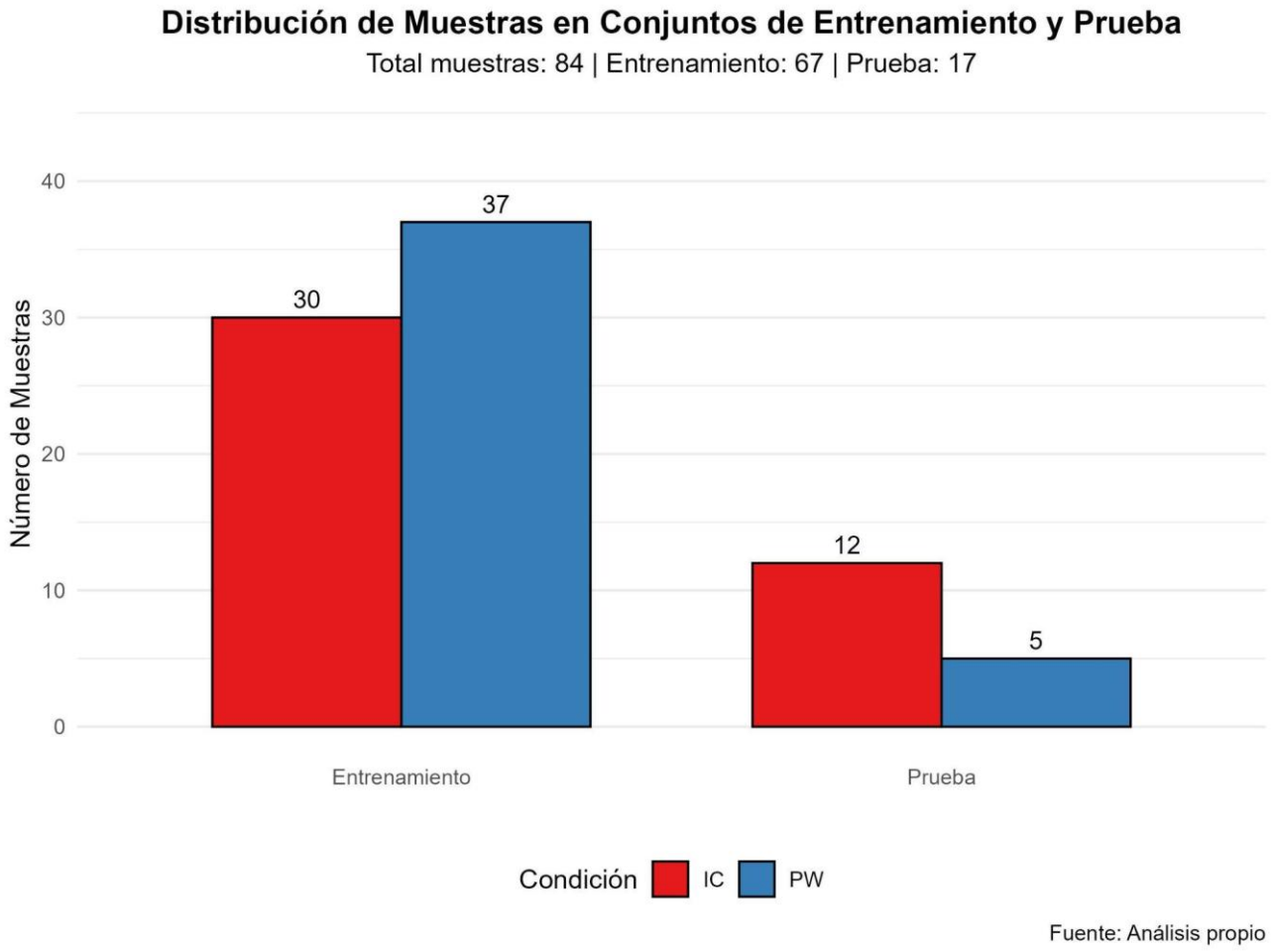


Figura 3. Train-test. ***Fuente:*** *Propia autoría*

3.7.7 Modelado predictivo

Para implementar el modelo de regresión Lasso, conocido como regularización L1 que es una técnica muy útil porque agrega una penalización basado en los valores absolutos de los coeficientes, descritos en la siguiente ecuación:

$$L\ 1 = \lambda \cdot (|\beta\ 1\ | + |beta\ 2\ | + \cdots + |\beta\ p|)$$

donde:

λ: Lambda controla el hiperparámetro ajustando su cantidad de regularización.

β 1 |+|beta 2 |+⋯+|β : suma los valores de los coeficientes del modelo

Este método de penalización Lambda (λ) a escala logarítmica se multiplica por la suma residual de cuadrados, equilibrando la compensación entre sesgo y varianza en los coeficientes resultantes. Se utilizó el paquete glmnet en R.

En este trabajo, para la construcción del modelo Lasso, se implementó un procedimiento iterativo que se repitió 500 veces. En cada repetición, el conjunto de datos fue dividido aleatoriamente en un 80% para train y un 20% para test. Para asegurar la reproducibilidad y controlar la aleatoriedad en cada iteración, se fijó una semilla o PRNG seed específica. Aunque este procedimiento no corresponde a la validación cruzada tradicional (k-fold), funciona como una estrategia de repetición aleatoria con reentrenamiento, útil para evaluar la estabilidad de la selección de variables.

Durante cada iteración, el modelo Lasso seleccionó un subconjunto de genes que consideró relevantes para la predicción. Para cuantificar la importancia de cada gen a lo largo de las repeticiones, se construyó un vector de frecuencias, registrando cuántas veces cada gen fue seleccionado. Posteriormente, estas frecuencias se normalizaron (por ejemplo, dividiendo por el total de repeticiones), obteniendo así un promedio de aparición que permitió identificar los genes más robustos o consistentemente seleccionados. Finalmente, el desempeño del modelo fue evaluado en cada repetición utilizando métricas como la exactitud (accuracy) y el puntaje F1 (F1-score), las cuales proporcionan una medida del equilibrio entre sensibilidad y precisión en la clasificación.

3.7.8 Análisis funcional y enriquecimiento

Los genes finales dados por el modelo Lasso son clasificados en dos grupos: genes sobre regulados (UP) y genes infra regulados (DOWN), comparando el Log2 Fold Change y el p-valor ajustado con bases de datos ya establecidas, este paso es importante ya que en la interpretación biológica se podría tener una mejor noción de qué vías metabólicas o procesos biológicos e inmunológicos están asociados con las sibilancias, si se encuentran sobreexpresados o infraexpresados.

Dentro del análisis de enriquecimiento de procesos biológicos (GO:BP), el término funcional Cytokine receptor binding, agrupó al gen SOCS3 como sobreexpresado en PW. SOCS3 (Supresor de la Señalización de Citocinas 3) es un regulador negativo de la vía JAK/STAT, actúa como un obstáculo ante la señalización de citocinas agresivas. En situaciones de inflamación crónica, como el asma o las sibilancias frecuentes, puede haber alteraciones en la regulación de este gen, lo que puede conducir a una respuesta inmunológica ineficaz o desbalanceada (Carow & Rottenberg, 2014).

Este mismo gen (SOCS3), identificado como sobreexpresado, también se encuentra implicado en la vía de señalización de IL-10 según los análisis de enriquecimiento Reactome y GSEA, indicando una posible regulación negativa de la inflamación. No obstante, esta sobreexpresión puede reflejar una respuesta compensatoria insuficiente para controlar la inflamación crónica de las vías respiratorias (Carow & Rottenberg, 2014).

Los resultados del Gene Ontology (GO) revelaron una representación considerable de funciones moleculares, que tienen relación directa con la señalización del sistema inmunológico. Estas funciones están vinculadas a genes reconocidos en el modelo predictivo, lo cual indica su participación en la regulación de respuestas inflamatorias en las vías respiratorias. Estos descubrimientos respaldan la importancia de los genes identificados, tales como SOCS3, CCL22 y PTGS2, al evidenciar una activación o regulación de vías inmunológicas esenciales en la aparición de inflamación crónica en las vías respiratorias (Barnes, 2008).

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 RESULTADOS

4.1.1 Resultados del Análisis de Expresión Génica

Partiendo de los datos crudos del estudio, se implementó el siguiente pipeline de preprocesamiento de datos iniciales:

- Construcción de la matriz de conteo: matriz inicial de 61,905 genes × 134 muestras
- Vector de condiciones: 44 IC y 90 PW.

Se llevó a cabo un análisis de expresión diferencial (sin balancear) utilizando un modelo basado en regresión binomial negativa. Para ello, se construyó un conjunto de datos que incluye los conteos de expresión y la información de las condiciones experimentales. Previamente, se aplicó un filtro para eliminar genes con niveles de expresión muy bajos, dando un total de 19690 genes, donde se seleccionaron los que tienen mayor significancia estadística. Entre estos, los 5 principales (ver tabla 2) destacan por su relevancia biológica.

ID del gen	baseMean	log2FC	lfcSE	stat	pvalue	padj
STX6	2325.981558	-3.454137	0.08179918	-42.22704	0.000000e+00	0.000000e+00
SKIL	3394.779711	-4.230783	0.07673146	-55.13753	0.000000e+00	0.000000e+00
LINC01033	662.706769	-8.242326	0.15510182	-53.14139	0.000000e+00	0.000000e+00
SQSTM1	14690.262851	-2.679271	0.06914017	-38.75129	0.000000e+00	0.000000e+00
IER3	750.873566	-4.646319	0.12088716	-38.43518	0.000000e+00	0.000000e+00

Tabla 2. Top 5 genes diferencialmente expresados. ***Fuente:*** *Propia autoría*

Luego de esto, se calculó la distancia de Mahalanobis basado en los dos primeros componentes principales (PC1 y PC2). Este análisis, representado gráficamente en la Figura 4, permitió detectar muestras atípicas que podrían sesgar los resultados posteriores.

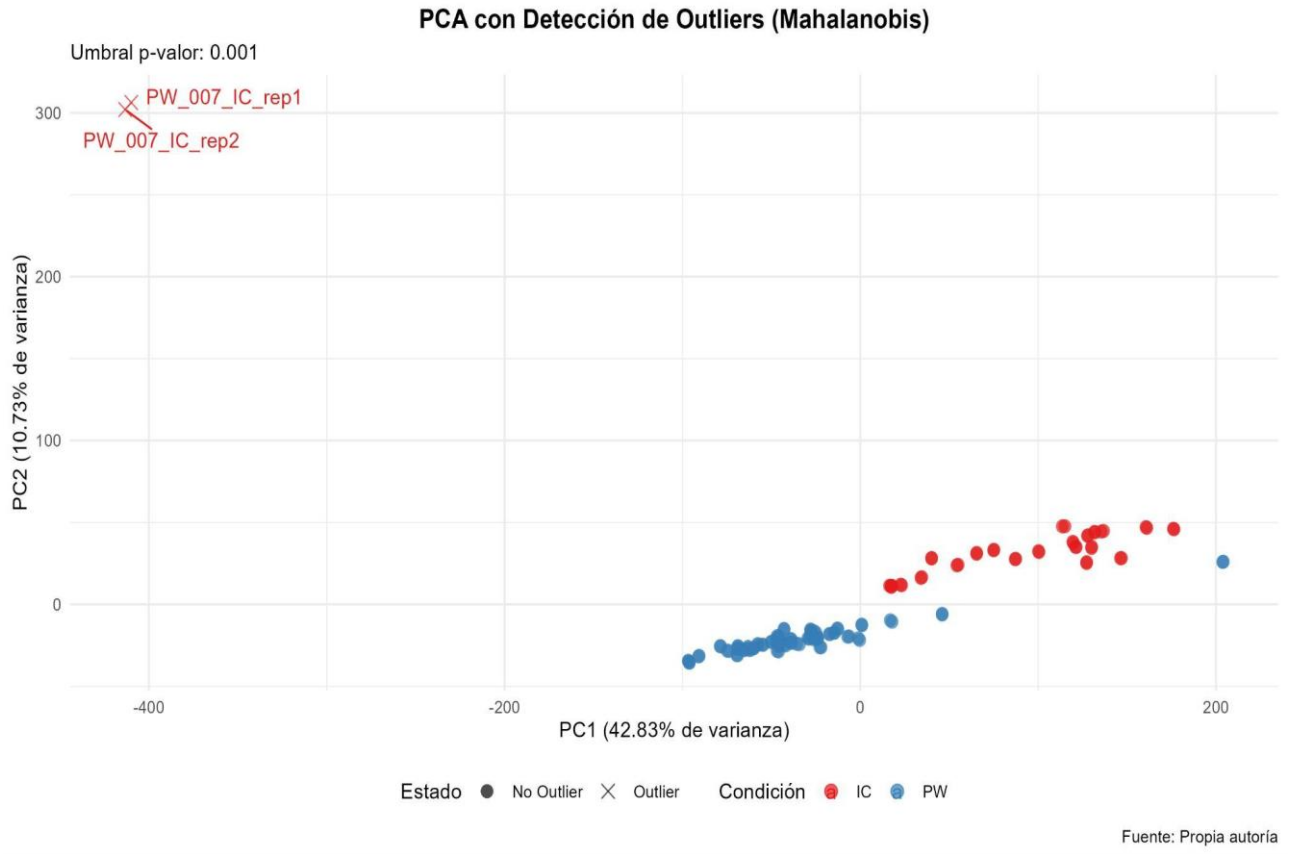


Figura 4. PCA y Mahalanobis con outliers y varianza del PC1 de 42.83% Y PC2 de 10.73%.
Fuente: *Propia autoría.*

Se fijó un umbral fundamentado en una distribución chi cuadrado con dos grados de libertad y un valor de corte de $p < 0.001$, lo que facilitó la identificación y eliminación de outliers que excedían el rango (Figura 5).

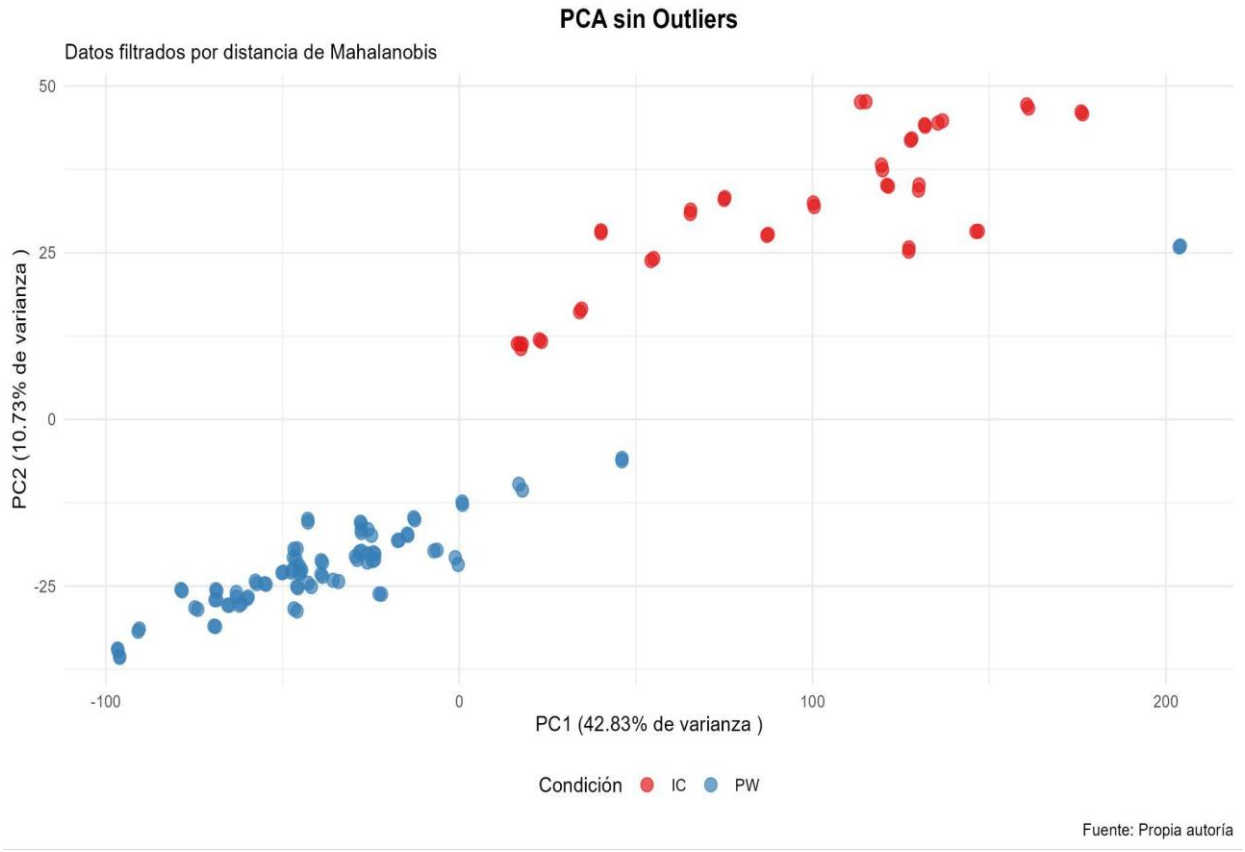


Figura 5. PCA sin outliers. ***Fuente:*** *Propia autoría.*

Matriz de Conteos Final:

- Muestras totales iniciales: 134 (42 IC + 92 PW)
- Muestras identificadas como outliers: 2 ("PW_007_IC_rep1", "PW_007_IC_rep2")
- Dimensiones: 61,905 genes × 132 muestras
- Genes analizados: 61,905
- Muestras válidas: 132 (42 IC + 90 PW)

Para garantizar la robustez del análisis de expresión diferencial, se implementó un pre-filtrado estricto de genes, bajo criterios estrictos:

- Significancia estadística: padj < 0.01
- Relevancia biológica: |log2FoldChange| ≥ 2

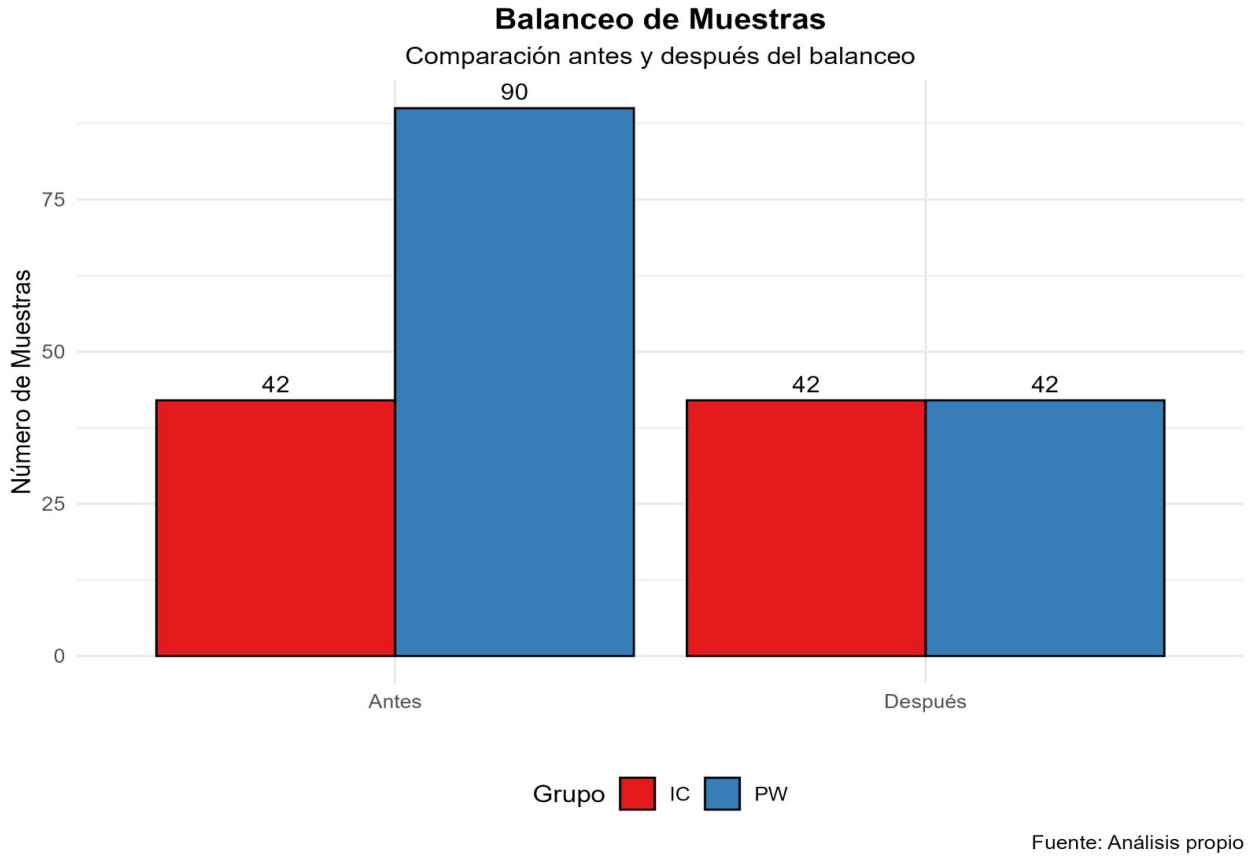


Figura 6. Balanceo de muestras. ***Fuente:*** *Propia autoría.*

Debido al desbalance entre las condiciones experimentales, se implementó un procedimiento de submuestreo para equilibrar el número de muestras entre los grupos comparados. En particular, se seleccionó aleatoriamente un subconjunto de muestras del grupo mayoritario, de modo que ambas condiciones quedarán representadas por un número equivalente de observaciones.

Una vez realizado el balanceo, se reconstruyó la matriz de conteos utilizando únicamente las muestras seleccionadas, y se actualizaron los metadatos correspondientes para asegurar la coherencia entre las matrices de expresión y las condiciones experimentales. Esta matriz balanceada se utilizó en los análisis posteriores, permitiendo una comparación más justa entre los grupos y evitando sesgos estadísticos derivados de un desequilibrio en el número de muestras. (Figura 6)

Tras el balance de las condiciones experimentales, se repitió el análisis de expresión diferencial empleando únicamente las muestras balanceadas. Se construyó un nuevo conjunto de datos con conteos normalizados y se aplicó nuevamente el modelo estadístico para detectar genes diferencialmente expresados entre las condiciones.

4.1.2 Selección de genes basada en PCA balanceado

Se aplicó un prefiltrado adicional para eliminar genes con baja representación, conservando únicamente aquellos expresados en al menos cinco muestras. Para identificar genes con mayor relevancia, se empleó un método exploratorio fundamentado en el Análisis de Componentes Principales (PCA), aplicado en la matriz de expresión normalizada previamente equilibrada entre las condiciones "IC" y "PW". El propósito era disminuir la dimensionalidad y resaltar los genes que explican una mayor variabilidad en la información, mejorando de esta manera la elección para el modelado.

El análisis PCA reveló que el primer componente (PC1) explicó el 31.3% de la varianza, mientras que el segundo componente (PC2) explicó el 11.1%, que incluyó elipses de confianza del 95% (ver Figura 7). Estos dos elementos concentraron la mayoría de la variabilidad biológica entre los grupos clínicos "PW" e "IC". Este análisis permitió visualizar la variabilidad de las muestras y evaluar si la separación entre las condiciones se mantenía tras el balance. Esto mostró una agrupación clara entre los grupos, lo que indica que la expresión genética de las muestras estudiadas muestra patrones diferentes dependiendo de su estado clínico.

Para construir una lista sólida de genes aspirantes, se emplearon dos métodos fusionados:

- Contribución al PC1 (genes con una carga factorial superior).
- Correlación > 0.5 entre PC1 y PC2.

La unificación de estos grupos posibilitó la creación de una lista final de genes empleados en el modelo predictivo a través de la regresión Lasso, garantizando tanto pertinencia estadística como representación estructural en el espacio multivariado.

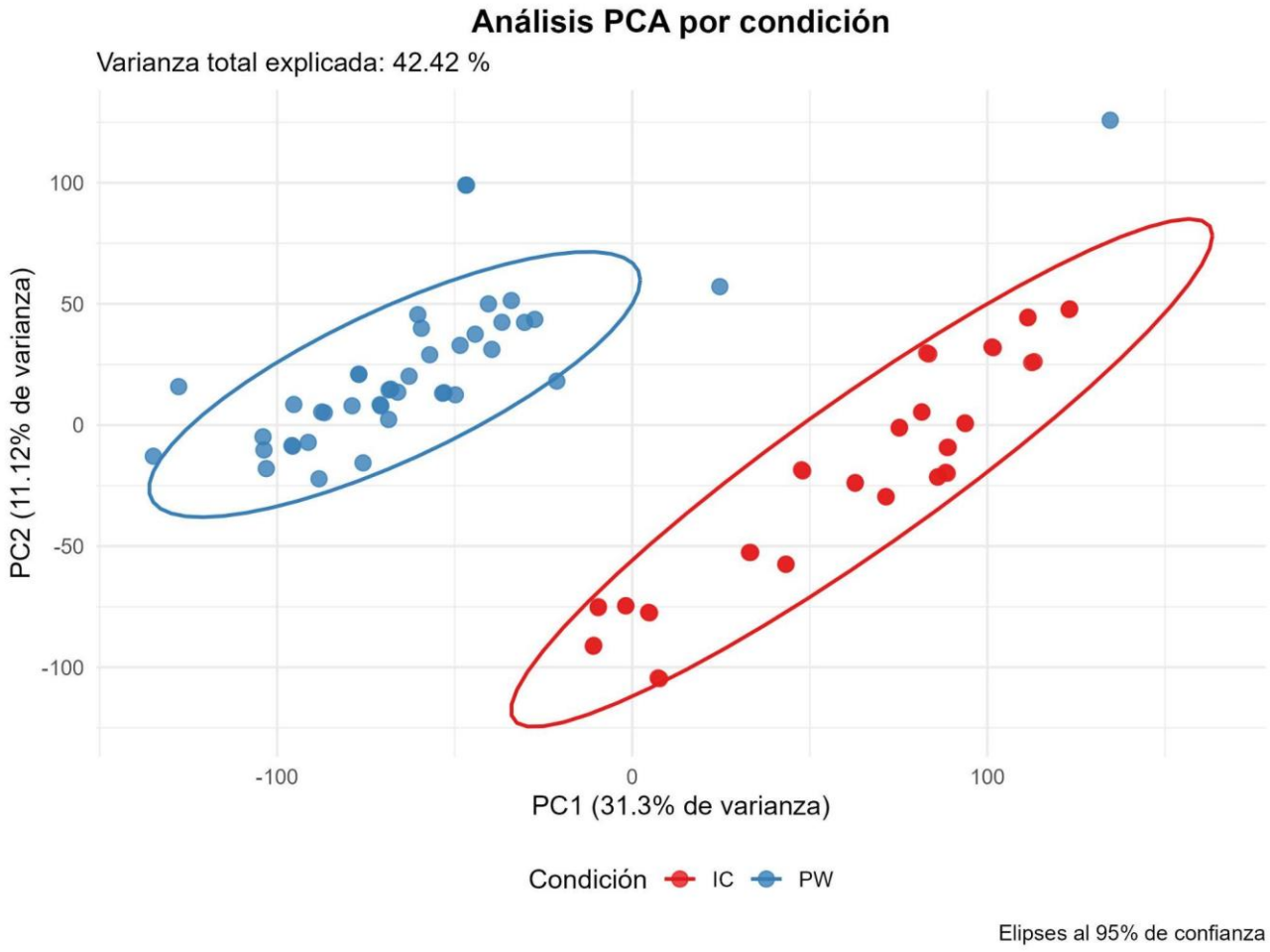


Figura 7. Visualización PCA de los datos transformados. ***Fuente:*** *Propia autoría*

Además, se construyó un gráfico de "Scree Plot" para determinar el porcentaje de varianza explicada por cada componente principal, y un gráfico bidimensional que representa las muestras en función de las dos primeras componentes, coloreadas según su condición experimental. (Figura 8)

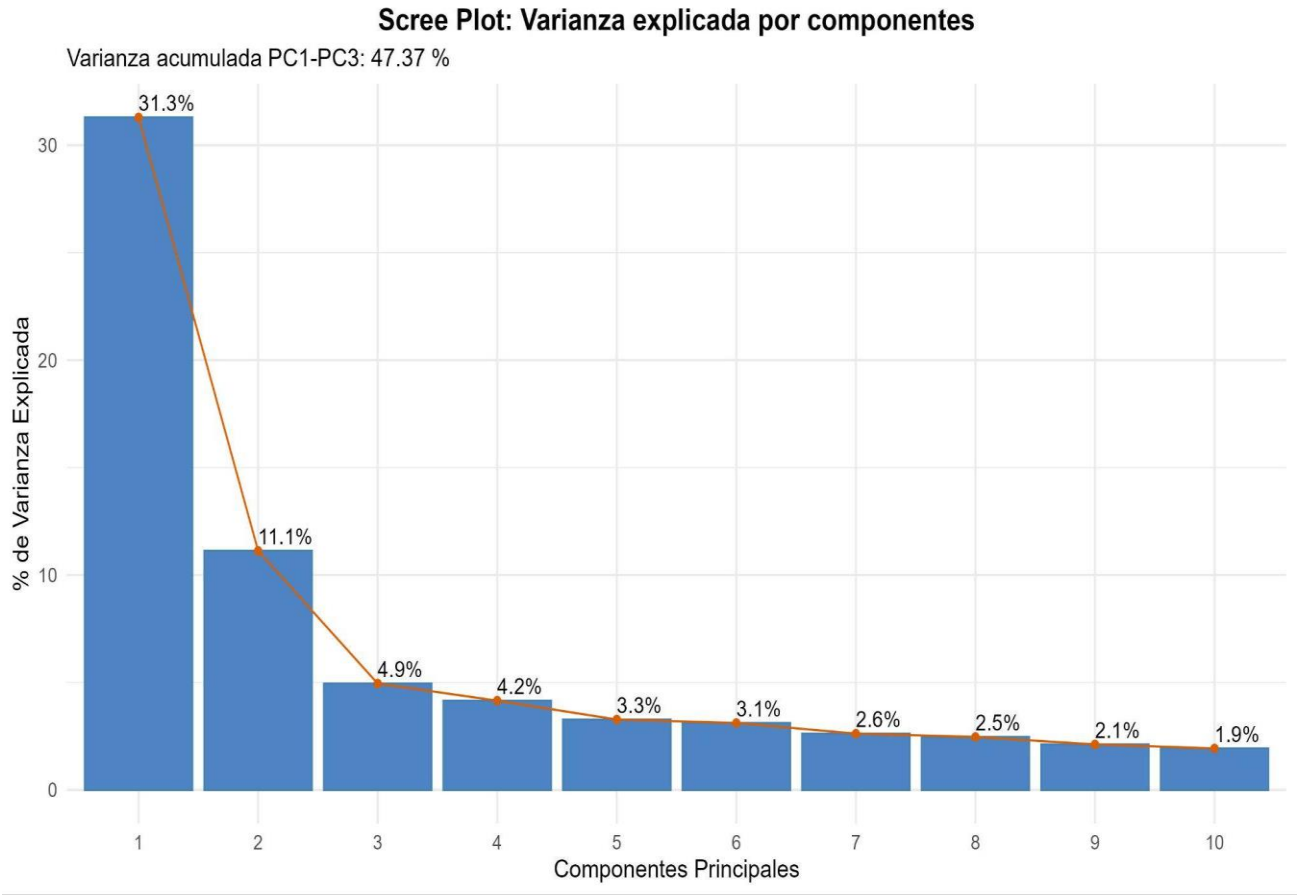


Figura 8. Scree Plot de PCA balanceado. ***Fuente:*** *Propia autoría*

4.1.3 Construcción y Evaluación del Modelo Lasso

Una vez obtenidos los datos normalizados y balanceados, y tras aplicar el análisis de componentes principales (PCA), se seleccionaron 3,981 genes expresados en un total de 84 muestras. Se procedió a construir el modelo Lasso y al finalizar el bucle, identificó un conjunto de 32 genes con coeficientes distintos de cero, es decir, aquellos considerados más relevantes para la predicción (Tabla 3).

Resultados del Modelo LASSO (500 repeticiones)			
ID	**Gen**	**Frecuencia de Selección**	**Coef. Promedio lasso (β)**
1	CDC20B	500	-1.32680154
2	SOCS3	492	0.68518900
3	CCL22	435	-0.27428518
4	SMAD7	434	-0.31658799
5	RIPOR2	432	0.31050516
6	PRADX	357	-0.21790350
7	CTSD	303	-0.12577007
8	NCCRP1	286	-0.14766689
9	AC112255.1	234	-0.09465140
10	CSKMT	200	-0.13621787
11	LINC01355	185	-0.08806704
12	LOC105374981	185	0.07239836
13	TRIM13	161	-0.12509545
14	CD83	159	-0.07077942
15	AL390066.2	145	-0.08341145
19	PPIAP29	128	0.14653577
17	KLF2	116	0.11279411
18	NUDT5	104	0.16160780
19	ERGIC1	93	0.46895857
20	FMNL3	90	-0.07882164
21	TAGLN2	88	0.40424072
22	AL162377.3	86	0.27002380
23	ECE1	85	0.09265589

24	LGALS3	84	-0.06223298
25	TUBA1A	81	0.24729685
26	PTGS2	80	0.49277009
27	DUSP1	76	0.41241687
28	CSF2RBP1	76	0.30307595
29	FAH	74	-0.16705807
30	CXCL2	70	-0.11091952
31	G0S2	67	0.16514471
32	RABEPK	54	0.04002456

Tabla 3. Top 32 genes según el modelo LASSO. ***Fuente:*** *Propia autoría*

Estos genes fueron seleccionados en función de su frecuencia de aparición, destacándose aquellos que se repitieron en más de 50 de las iteraciones, lo que refuerza su estabilidad y posible valor como biomarcadores diagnósticos. Como se muestra en la Tabla 4, estos genes se asocian en dos procesos principales:

- Genes asociados a vías metabólicas: Presentes en las columnas con sus respectivas vías.
- Genes de procesos inmunológicos: Indicados en la tabla, incluyen reguladores clave de la respuesta inflamatoria.

Resultados del Modelo LASSO					
ID	**Gen**	**baseMean**	**log2FC**	**p-value**	**Vías metabólicas**
1	CDC20B	56.88772	-7.988037	3.850308e-109	Genes relacionados a procesos inmunológicos
2	SOCS3	2873.97	5.817258	9.504844e-253	
3	CCL22	678.4371	-8.337372	4.276389e-149	
4	SMAD7	628.6176	-4.017983	7.235459e-213	
5	RIPOR2	28703.93	4.314179	5.142528e-266	
6	PRADX	7.745691	-5.708594	2.258013e-53	
7	NCCRP1	37.6801	-6.457915	1.089912e-66	
8	AC112255.1	58.52154	-5.606607	1.143991e-74	
9	CSKMT	441.2369	-6.617948	5.883814e-184	
10	LINC01355	18.15856	-6.212639	8.689331e-57	

11	LOC105374981	101.9829	4.969397	5.40496e-125	
12	TRIM13	460.4419	-1.750341	8.75112e-102	
13	CD83	3154.17	-4.437616	2.019287e-119	
14	AL390066.2	66.64604	-3.635443	3.365808e-105	
15	PPIAP29	27.69346	5.466388	1.654871e-89	
16	KLF2	4250.757	3.471042	9.298419e-145	
17	ERGIC1	1072.031	2.133311	1.934448e-139	Genes relacionados a procesos inmunológicos
18	FMNL3	845.172	-4.608274	7.014749e-163	
19	TAGLN2	33363.17	3.580748	3.847059e-196	
20	AL162377.3	25.3446	7.437511	3.03208e-79	
21	ECE1	5693.041	3.783666	2.865529e-136	
22	LGALS3	2529.265	-4.018577	2.463098e-108	

23	TUBA1A	17234.27	4.464649	1.181273e-243	
24	DUSP1	54942.26	4.521531	3.547967e-140	
25	CSF2RBP1	46.43564	6.827523	1.580956e-92	
26	CXCL2	216.5022	-6.710364	6.448055e-152	
27	RABEPK	261.3037	1.713711	1.898764e-87	
28	CTSD	12223.41	-2.953815	9.322849e-77	Metabolismo de Angiotensinógeno a Angiotensinas Metabolismo de Hormonas Peptídicas
29	NUDT5	1058.257	2.347245	3.358803e-169	Metabolismo de Nucleótidos
30	PTGS2	9726.525	6.391141	1.370716e-193	Metabolismo de Nicotinato Metabolismo de Ácido Araquidónico
31	FAH	145.5428	-3.736985	8.070251e-129	Metabolismo de Fenilalanina y Tirosina Metabolismo de vitaminas hidrosolubles y cofactores Metabolismo de los ácidos grasos Metabolismo de los aminoácidos y derivados

32	G0S2	10839.99	4.534748	1.844723e-73	Regulación del metabolismo lipídico por PPARalfa

Tabla 4. Genes asociados a vías metabólicas y procesos inmunológicos. ***Fuente:*** *Propia autoría*

Para evaluar la viabilidad y confiabilidad del modelo, se aplicaron dos métricas. La primera fue la precisión (accuracy), que mide la proporción de predicciones correctas sobre el total de predicciones realizadas. En este caso, el modelo alcanzó una precisión promedio de 0.989, con una desviación estándar de 0.0232, lo que indica una alta capacidad de clasificación. (Figura 9)

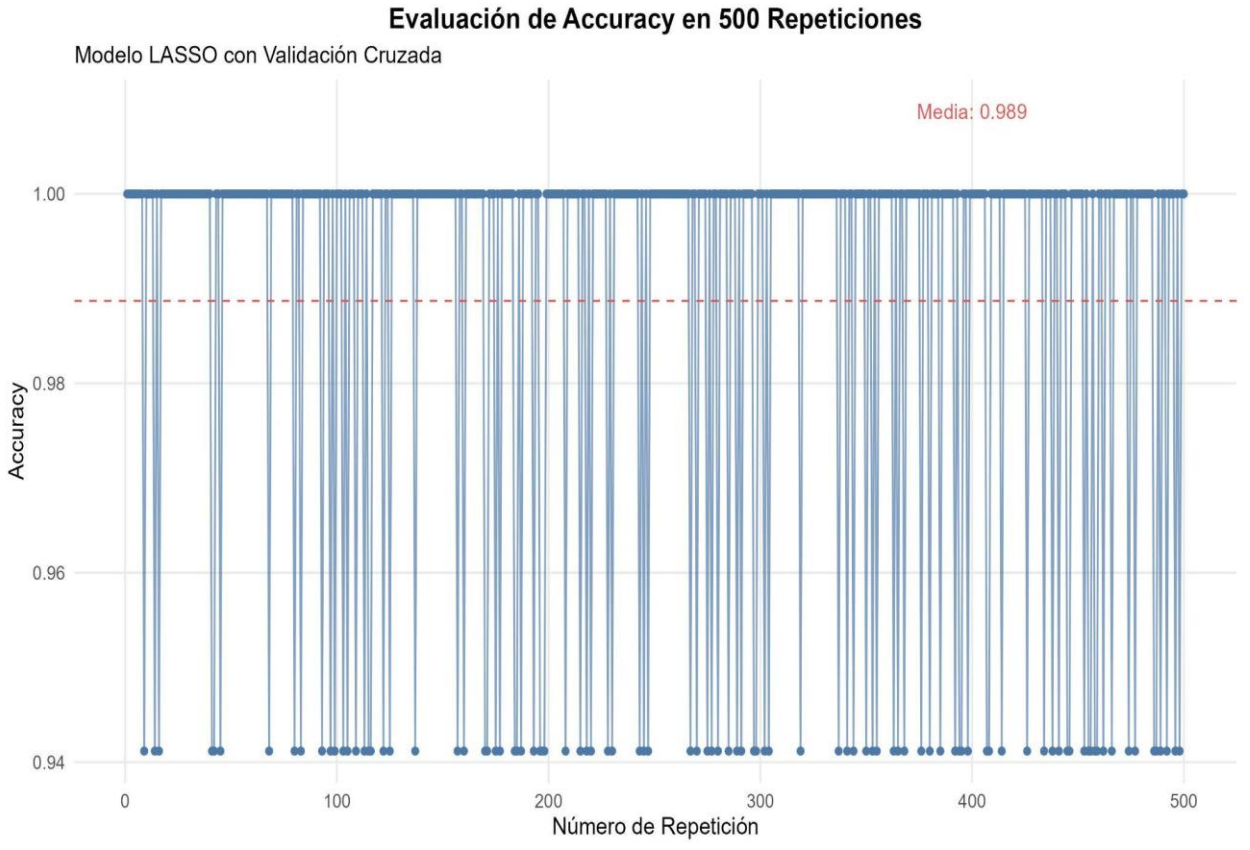


Figura 9. Evaluación de exactitud o accuracy. ***Fuente:*** *Propia autoría*

La segunda métrica utilizada fue el F1 Score, que representa el balance entre precisión y sensibilidad. Este indicador resultó en un valor promedio de 0.9879, con una desviación estándar de 0.0256, lo que refuerza la robustez del modelo incluso frente a desequilibrios entre clases. (Figura 10)

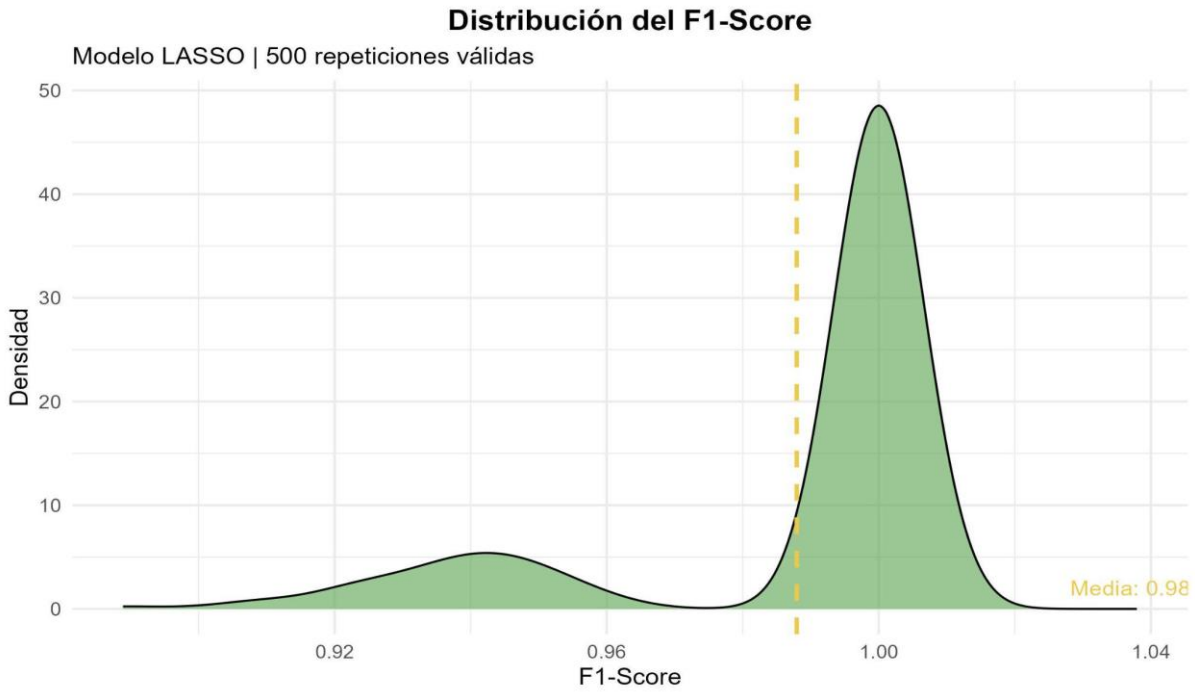


Figura 10. DotPlot de Distribución F1 score del modelo LASSO. ***Fuente:*** *Propia autoría*

El modelo LASSO demostró un rendimiento estable en las 500 repeticiones (ver figura 11) validando su confiabilidad para predecir sibilancias recurrentes. La pequeña diferencia entre Accuracy y *F1-Score* sugiere que el manejo de clases desbalanceadas (si existieran) fue efectivo.

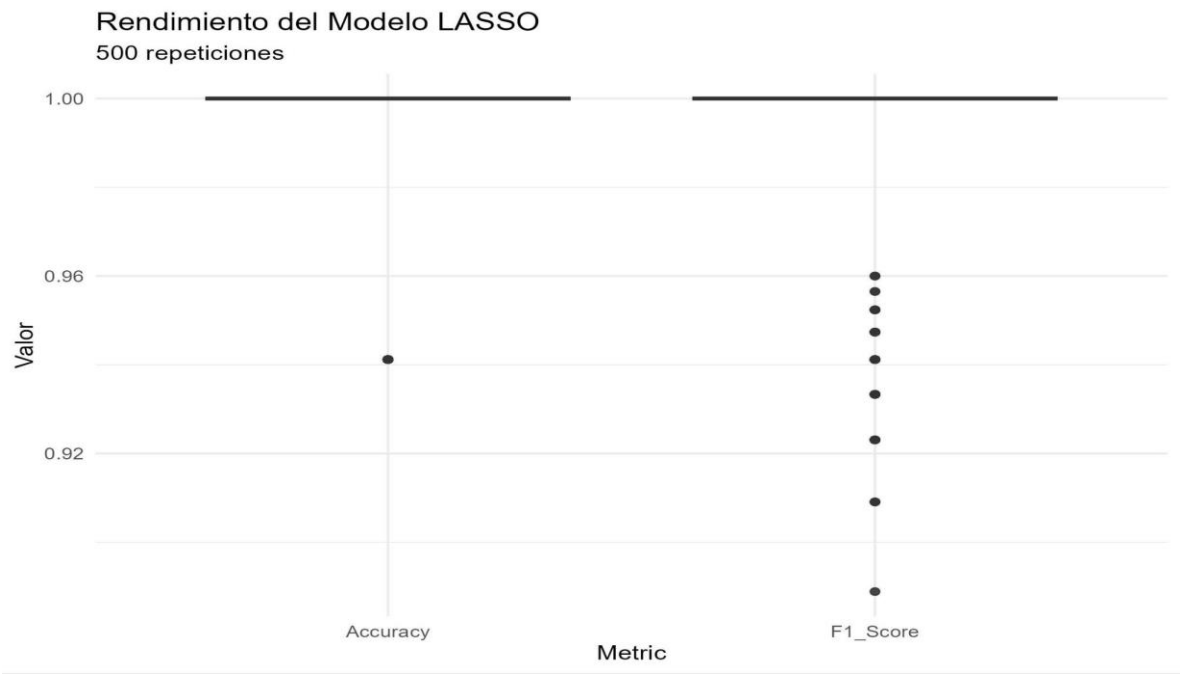


Figura 11. BoxPlot de la comparativa de la precisión (accuracy) y el F1 score LASSO.

Fuente: *Propia autoría*

4.1.4 Análisis de enriquecimiento funcional

Comprender cómo se relacionan los genes previamente seleccionados con las vías metabólicas, funciones moleculares, procesos biológicos y enfermedades, en lo que concierne con las sibilancias, es un paso muy importante, a fin de detectar biomarcadores potenciales, aquí es donde el análisis de enriquecimiento cobra mucha importancia.

Se realizaron dos tipos de análisis: el Análisis de Sobre-Representación (Over-Representation Analysis, ORA) y el Análisis de Enriquecimiento de Conjunto de Genes (Gene Set Enrichment Analysis, GSEA). En adelante, se utilizarán las siglas ORA para referirse al primer análisis y GSEA para el segundo, con el objetivo de describir los resultados del enriquecimiento eficazmente; se utilizaron gráficos de burbujas, en donde cada una representa una enfermedad, vía metabólica, funciones moleculares o procesos biológicos, cuyos nombres se leen en la parte derecha del gráfico, mientras que las burbujas más grandes significan la cantidad de genes

enriquecidos involucrados. Cabe recalcar que, pueden existir enfermedades o procesos con una cantidad importante de genes que no están relacionados con las sibilancias, a esos no se los tomó en cuenta, debido a que no son relevantes en este estudio.

Como ya se mencionó anteriormente, los genes detectados en esta investigación se escogieron a través del modelo LASSO, por su capacidad para anticipar la condición de sibilancias (PW) en comparación con controles sanos (IC). Su coeficiente LASSO señala la orientación e intensidad de la relación con ambas condiciones, en tanto que el log2FC (expresión diferencial) muestra su comportamiento biológico auténtico.

Interpretación del coeficiente y cambio en Log2FC

- Log2FC aumenta (+) y coeficiente positivo ($\beta > 0$):
  Un aumento en Log2FC incrementa la probabilidad de pertenecer a la clase 1.

- Log2FC aumenta (+) y coeficiente negativo ($\beta < 0$):
  Un aumento en Log2FC reduce la probabilidad de pertenecer a la clase 1.

- Log2FC disminuye (−) y coeficiente positivo ($\beta > 0$):
  Una disminución en Log2FC reduce la probabilidad de pertenecer a la clase 1.

- Log2FC disminuye (−) y coeficiente negativo ($\beta < 0$):
  Una disminución en Log2FC aumenta la probabilidad de pertenecer a la clase 1.

En la figura 12 se muestra la ontología GO:MF (Gene Ontology – Molecular Function), donde se identificaron dos términos funcionales enriquecidos: "cytokine receptor binding" y "chemokine activity".

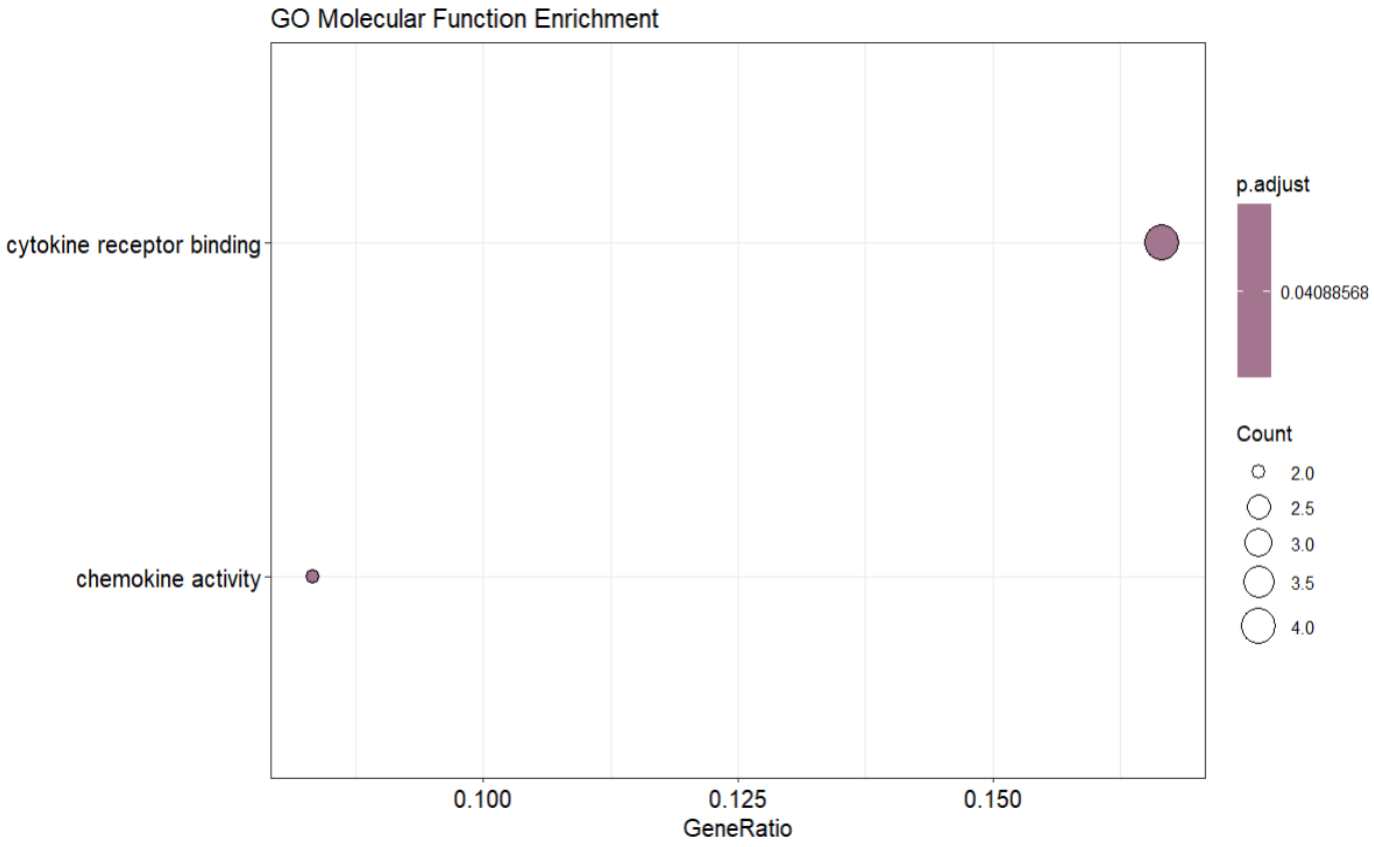


Figura 12. Visualización de funciones moleculares de la base de datos GO.
Fuente: *Propia autoría*

Estos términos están significativamente enriquecidos en los genes seleccionados por el modelo LASSO, los cuales muestran cambios marcados en su expresión génica entre niños con y sin sibilancias. Los resultados de este estudio revelaron un conjunto de genes altamente predictivos para la condición de sibilancias, los cuales se irán detallando a lo largo de esta sección.

La tabla 5 presenta los genes más relevantes correspondientes a las funciones moleculares de la base de datos GO, donde se integra los resultados para explicar su relevancia biológica:

Gen	Log2FC / Lasso	Dirección en PW	Ubicación celular (GO:CC)	Interpretación biológica	¿Concuerda con el resultado?
SOCS3	+5.82 / +0.685	Sobre expresado	Citoplasma/ Citosol	**Función molecular (GO:MF)** Cytokine receptor binding; inhibidor de la vía JAK/STAT. Regulación negativa de señalización por citoquinas; inflamación crónica tipo Th2.	**Sí**. Sobreexpresado en PW como respuesta a hiperinflamación, pero no es suficiente.
CCL22	− 8.34/ − 0.274	Infrarregulado	Espacio extracelular/ secreción	**Función molecular (GO:MF)** Chemokine activity, unión a receptor. Atracción de células T reguladoras (Tregs), regulación de la inflamación, señalización Th2	**Sí**. Reprimido en PW, lo que sugiere menor reclutamiento de Tregs e inflamación persistente.

Tabla 5. Interpretación biológica SOCS3, CCL22. ***Fuente:*** *Propia autoría*

A pesar de que SOCS3 es un regulador negativo, su expresión excesiva indica un esfuerzo del sistema inmune para curar la inflamación. La información GO coincide con el rol previsto de SOCS3 en la inflamación de tipo Th2. En cambio, CCL22 muestra una represión intensa en el grupo PW (niños con sibilancias), presentando un coeficiente Lasso negativo y log2FC bajo, lo que sugiere una expresión reducida, por tal motivo, si se produce menos CCL22, habrá menos Tregs y,

por ende, menos regulación de la inflamación, lo que podría provocar la persistencia del estado inflamatorio en PW.

Los resultados del análisis ORA indicaron que los procesos biológicos más relevantes fueron la respuesta a la citosina, quimiotaxis leucocitaria y la respuesta celular a la quimiocina (figura 13) involucrados en dar señalización de vías y procesos inflamatorios; estas se encuentran estrechamente relacionadas con las sibilancias.

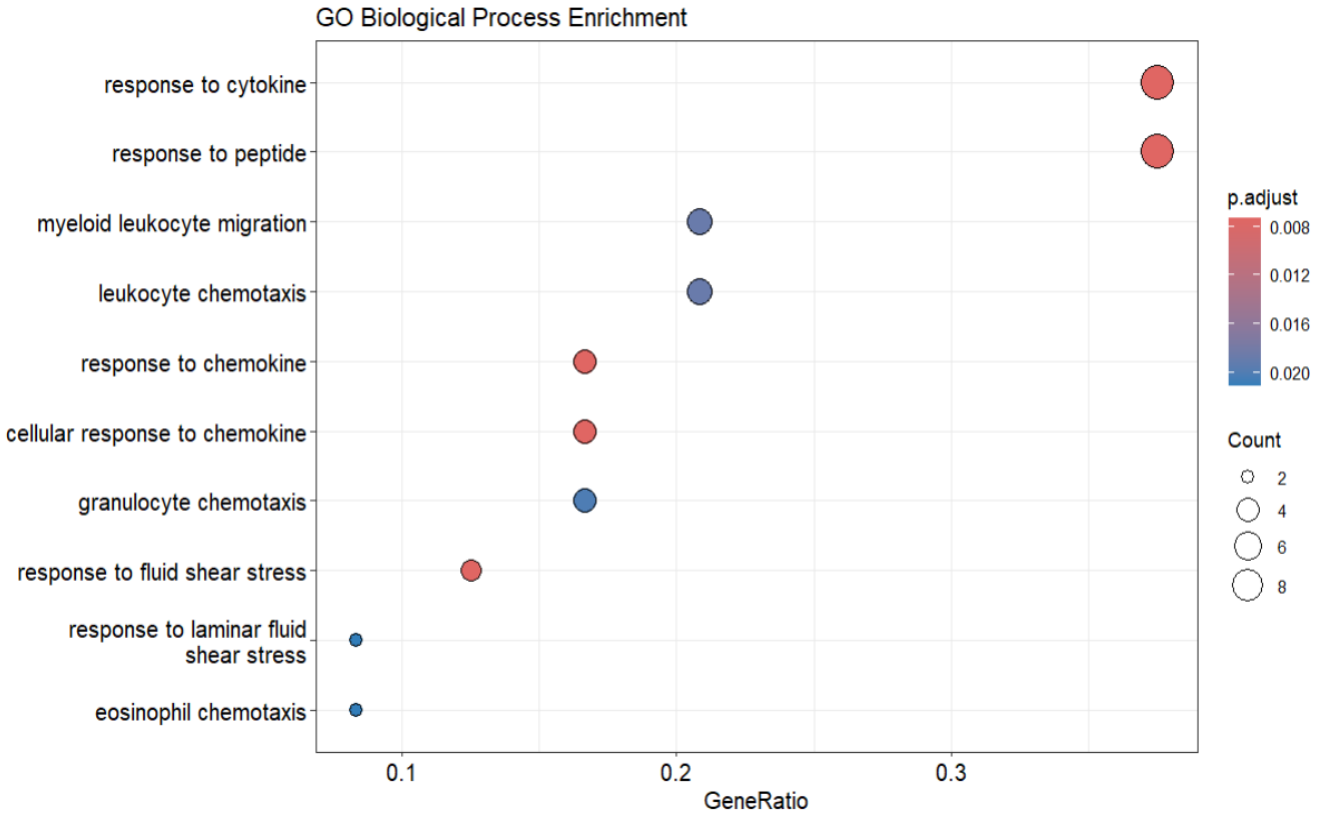


Figura 13. Procesos Biológicos por medio de la base de datos GO ***Fuente:*** *Propia autoría*

En la base de datos GO: BP, los genes que más destacaron fueron: LGALS3 (gen que codifica la galectina-3), se encuentra suprimido en PW, lo que podría indicar un desbalance en los procesos habituales de fibrosis y reestructuración epitelial en las vías respiratorias inflamadas y PTGS2 se encuentra sobreexpresado en PW, lo cual coincide con su función en la inflamación y la hiperreactividad bronquial, factores característicos de las sibilancias crónicas. (Tabla 6)

Gen	Log2FC / Lasso	Dirección en PW	Ubicación celular (GO:CC)	Interpretación biológica	¿Concuerda con el resultado?
LGALS3	− 4.02 / − 0.062	Infrarregulado	Extracelular, citoplasma, núcleo	**Proceso biológico (GO:BP / KEGG / Reactome)** Promueve fibrosis en vías aéreas, inflamación tisular.	**Sí**. Su represión puede reflejar pérdida del control del remodelado en vías aéreas inflamadas
PTGS2	+6.39 / +0.493	Sobre expresado	Retículo endoplasmático, citosol	**Proceso biológico (GO:BP/KEGG/ Reactome)** Inflamación aguda, síntesis de prostaglandinas, hiperreactividad bronquial	**Sí**. Su sobreexpresión se alinea con la inflamación activa y la hiperreactividad en PW.

Tabla 6. Interpretación biológica LGALS3, PTGS2. ***Fuente:*** *Propia autoría*

A continuación, se observa una única vía enriquecida del análisis ORA comparada con la base de datos Reactome: la señalización de la interleucina 10 (IL-10). Esta citoquina tiene características antiinflamatorias que pueden ayudar a disminuir los procesos inflamatorios vinculados a las sibilancias, disminuyendo de esta manera la posibilidad de que surjan los síntomas (Figura 14). Cabe recalcar que, SOCS3 desempeña un papel regulatorio clave en la vía de señalización de IL-10, ya que la disfunción de este se vincula con el asma severo, por tal motivo, fue el gen que más destacó en este análisis de enriquecimiento.

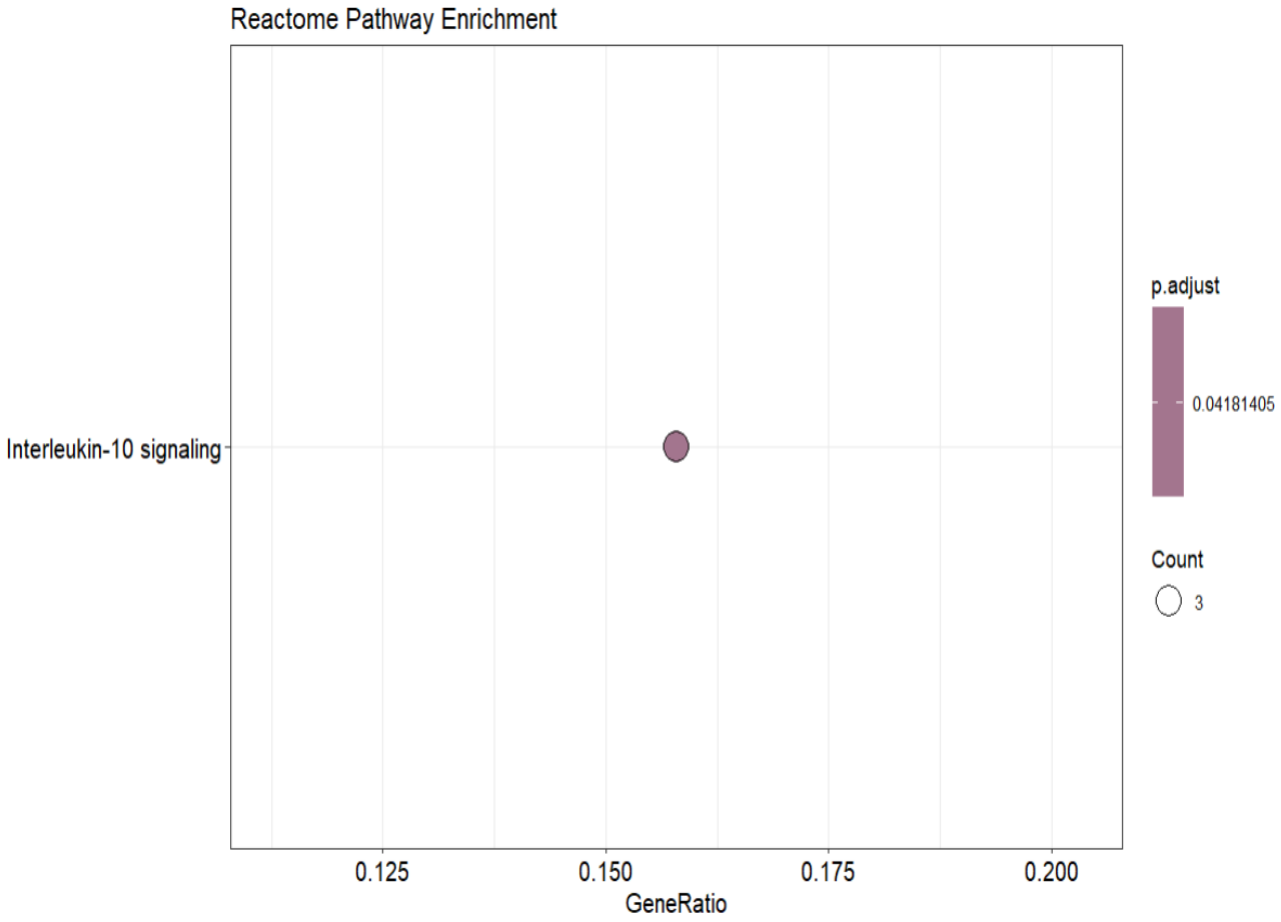


Figura 14. ORA con Base de datos de Reactome. ***Fuente:*** *Propia autoría.*

El análisis de enfermedades enriquecidas, basado en la base de datos Disease Ontology (DO), reveló diversas patologías asociadas (Figura 15). Entre las más destacadas se encuentran:

Las enfermedades pulmonares (Lung diseases), como el asma y la EPOC, que comparten con las sibilancias el mismo mecanismo como la inflamación crónica, la infiltración de macrófagos y la hiperreactividad bronquial.

En cuanto a la aterosclerosis, aunque se trata de una enfermedad vascular, está cada vez más reconocida como una condición inflamatoria sistémica que puede agravar la disfunción pulmonar, especialmente cuando coexiste con enfermedades respiratorias crónicas (Thompson et al., 2019).

Además, también se observó una posible asociación con diabetes, que, aunque no es una

enfermedad respiratoria directa, puede influir de manera indirecta, ya que las personas con esta enfermedad presentan una función inmune baja, lo que incrementa el desarrollo de episodios de sibilancias, especialmente en niños (Carey et al., 2018). Por tanto, los hallazgos de DO refuerzan el papel de la inflamación sistémica e inmune en el desarrollo de sibilancias, destacando la relevancia de los genes LGALS3 y PTGS2.

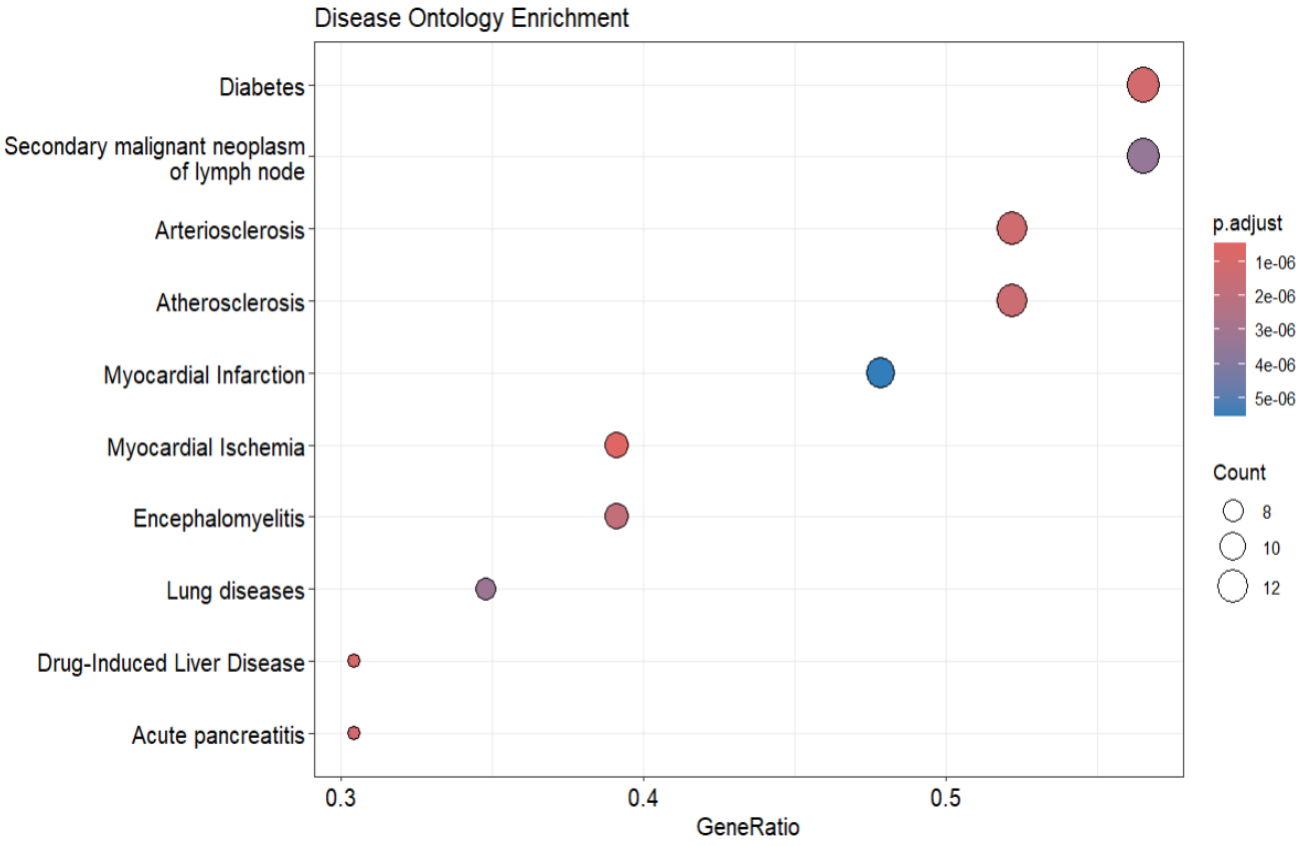


Figura 15. Cantidad de enfermedades enriquecidas de la base de datos Disease Ontology (DO).

Fuente: *Propia autoría*

En esta base de datos se destacaron dos genes asociados a las sibilancias (tabla 7). Tenemos al gen LGALS3 previamente mencionado, asociado con actividad de los macrófagos los cuales son de suma importancia en la respuesta inmune. Su downregulation podría favorecer a una obstrucción bronquial o una respuesta inmune agresiva y descontrolada. Por otro lado, PTGS2, codificante de

la ciclooxigenasa-2 (COX-2), participa en la síntesis de la hormona prostaglandina y cuando hay una sobrerregulación aumenta la producción de esta, y puede causar edema bronquial o hiperactividad de las vías aéreas, entre otros.

Gen	Log2FC / Lasso	Dirección en PW	Ubicación celular (GO:CC)	Interpretación biológica	¿Concuerda con el resultado?
LGALS3	− 4.02 / − 0.062	Subregulada	citoplasma, membrana y exoma extracelular	**Desease Ontology (DO)** Vinculada a procesos de inflamación y fibrosis con las enfermedades vinculadas	Si, se da una subregulación en PW, que bloquean los procesos proinflamatorios.
PTGS2	+6.39 / +0.493	Hiperregulada	peroxisoma, citoplasma y membrana nuclear,	**Desease Ontology (DO)** Es importante en procesos de inflamación y multiplicación celular y participa en enfermedades vasculares, artritis y cancer	Si,se da una hiperregulación en PW y esto lleva a activar las vías inflamatorias.

Tabla 7. Interpretación biológica LGALS, PTGS2. Fuente: Propia autoría

En el caso del análisis GSEA, los resultados se compararon con la base de datos KEGG, donde se identificaron varias vías biológicas relevantes relacionadas con las sibilancias. Entre ellas destacan la vía de señalización del TNF, vía de señalización de la interleucina 17 (IL-17) y la vía de señalización del NF-kappa B. (Figura 16)

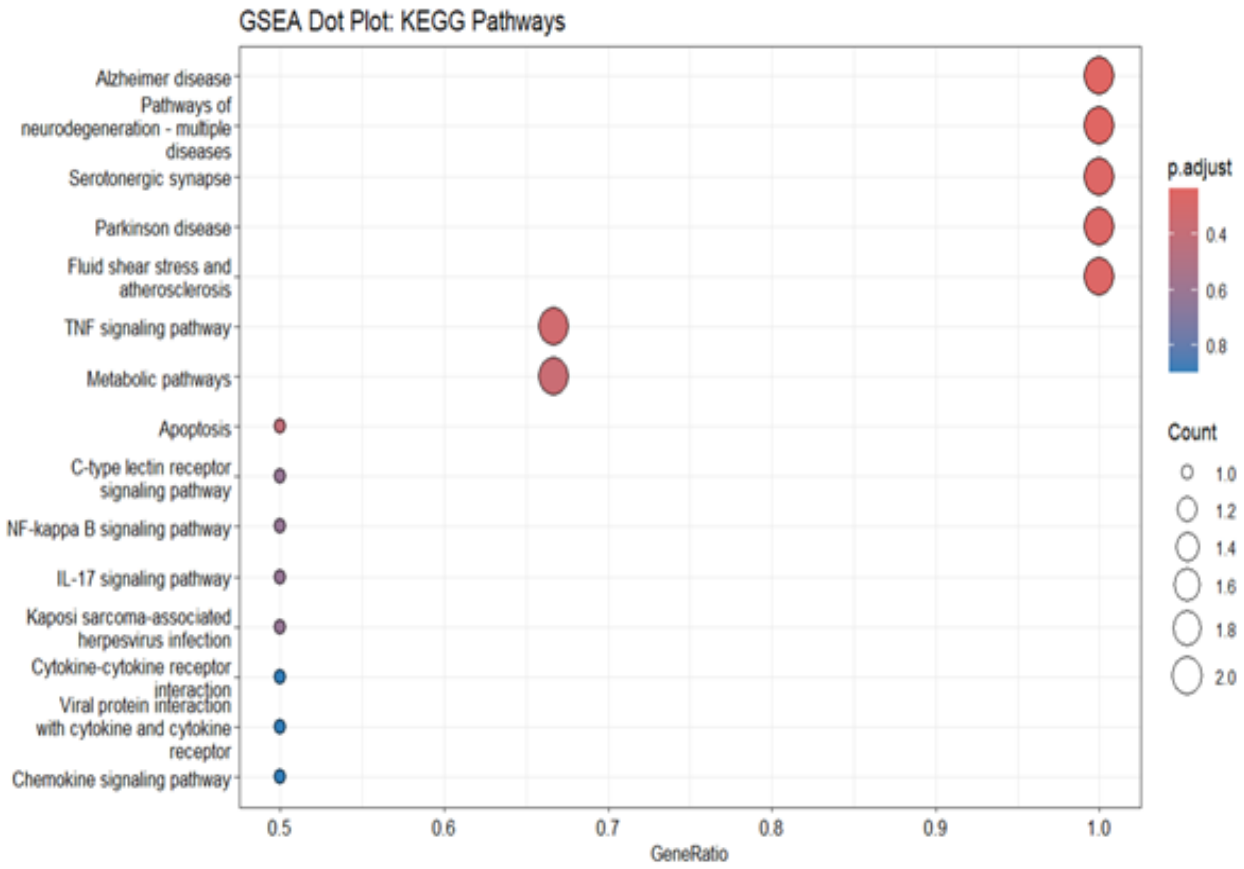


Figura 16. GSEA comparado con la base de datos KEGG. ***Fuente:*** *Propia autoría*

En este análisis destacan varios genes (ver tabla 8): El gen PTGS2, que activa procesos de inflamación, mientras que CXCL2 es relevante ante las infecciones respiratorias y SOCS3 manifiesta una retroalimentación negativa que reduce el impacto de hiperinflamación, se puede decir que en resumen todos estos causan estados de inflamación.

Gen	Log2FC / Lasso	Dirección en PW	Ubicación celular (GO:CC)	Interpretación biológica	¿Concuerda con el resultado?
PTGS2 (COX-2)	− 4.02 / − 0.062	Infrarregulada	Membrana nuclear, citoplasma	**Proceso biológico (GO:BP / KEGG / Reactome)** Indispensable en la inflamación.	Si, indica un método infrarregulado en PW que da una respuesta inflamatoria
CXCL2	-6.710 / -0.110	Infrarregulada	Matriz extracelular y secreciones	**Proceso biológico (GO:BP/KEGG/ Reactome)** Recluta neutrofilos es una quimiocina proinflamatoria.	Si, su infrarregulacion en PW muestra una disminución en el reclutamiento de neutrófilos.
SOCS3	+5.817/ +0.685	Sobrerrregulada	Complejo de señalización y citoplasma	**Proceso biológico (GO:BP/KEGG/ Reactome)** Inhibe la señalización de citoquinas como: IL-6 y vía JAK/STAT.	Si, su sobreregulación en PW indica una respuesta negativa al tratar de controlar la inflamación.

Tabla 8. Interpretación biológica PTGS2 (COX-2), CXCL2, SOCS3. Fuente: Propia autoría

En la figura 17 se muestra el análisis GSEA de los genes enriquecidos comparados con la base de datos Reactome. Se identificó una vía asociada a la señalización de citocinas en el sistema inmune. Además, la señalización general de interleucinas se destacó por su papel central en la activación del sistema inmunológico, el cual es fundamental en la defensa contra diversas enfermedades, incluidas las respiratorias.

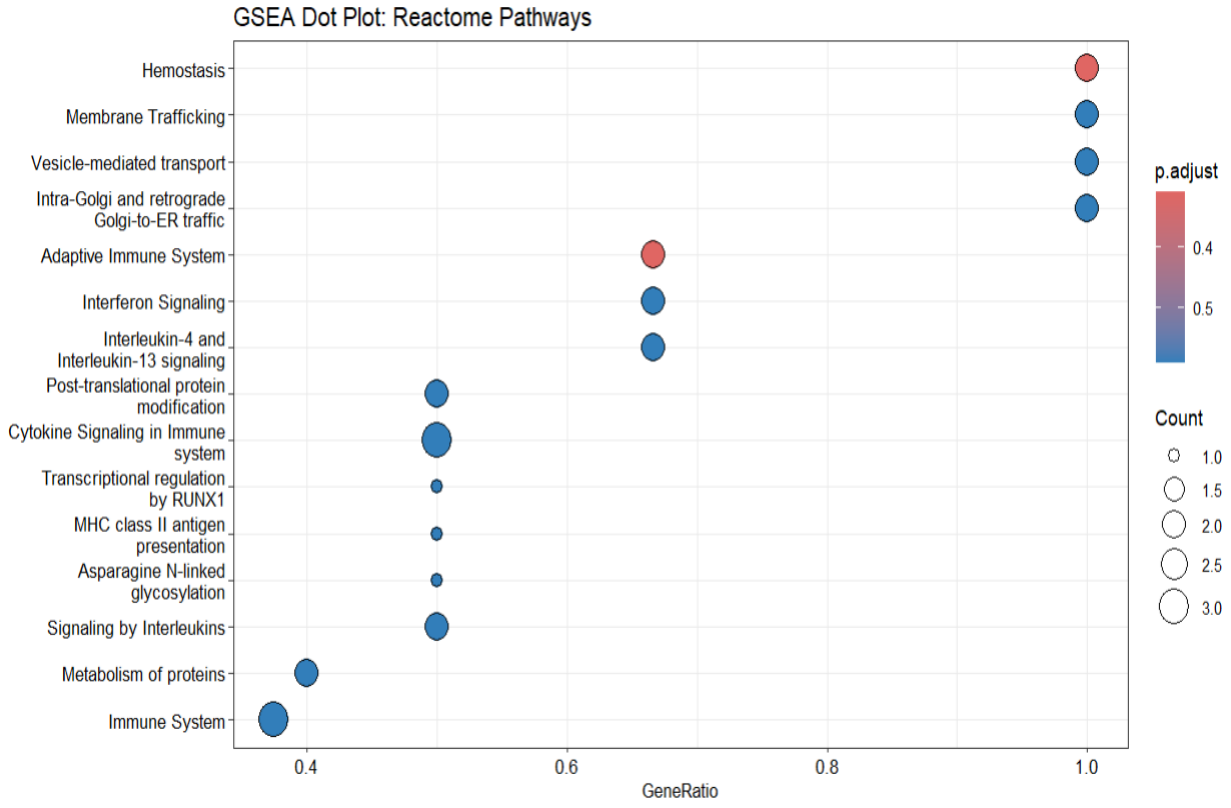


Figura 17. GSEA de genes enriquecidos comparados con la base de datos de Reactome.

Fuente: *Propia autoría.*

Vía de señalización TNF tiene a las TNF que son citocinas proinflamatorias, participa en diversos procesos como la diferenciación, modulación de respuestas inmunes y la apoptosis, tiene dos receptores: TNFR1: es el principal y puede causar cascadas de señalización o activar otras vías como la NF-kappa. TNFR2: media la respuesta biológica y modula la respuesta inmune y la supervivencia celular.

Vía de señalización NF-kappa B o (nuclear factor kappa-light-chain-enhancer of activated B cells) regula una gran parte de los genes involucrados en la respuesta inmune, al igual que con la vía TNF también es proinflamatoria, se activa por dos vías: la clásica que se activa por estímulos proinflamatorios como la de las quinasas y la vía alternativa que se activa por un conjunto de receptores de TNF, su desregularización da como resultado varias enfermedades inflamatorias y autoinmunes. La vía de señalización IL-17 al igual que las anteriores tienen citocinas proinflamatorias llamadas interleucinas generadas por los linfocitos CD4+ conocidas también como células THelper (Th17), ayudan a la respuesta inmune adaptativa activando varias cascadas de señalización y su desregularización puede llevar a varias enfermedades autoinmunes e inflamatorias.

Gen	Log2FC / Lasso	Dirección en PW	Ubicación celular (GO:CC)	Interpretación biológica	¿Concuerda con el resultado?
SOCS3	+5.817 / +0.685	Sobrerregulado	Citoplasma, complejo de señalización JAK-STAT	**Proceso biológico (GO:BP / KEGG / Reactome)** Inhibe las citocinas y las vías son Regulación negativa de la vía JAK-STAT (GO:0046426), KEGG: JAK-STAT signaling pathway (map04630).	Si, se muestra sobreregulado en PW dando un mecanismo de respuesta negativo al querer controlar la inflamación, es muy común en infecciones autoinmunes.
DUSP1	+4.521 / +0.412	Sobrerregulado	Citoplasma, Núcleo.	**Proceso biológico (GO:BP/KEGG/ Reactome)** Es una fosfatasa que desactiva las MAP quinasas, modulando la respuesta al estrés e inflamación. tiene vías como TNF signaling pathway (map04668). Reactome: MAPK family signaling cascades (R-HSA-5683057).	Si, se muestra sobreregulado en PW esto hace que se suprima la inflamación y también se asocia con la resistencia al estrés oxidativo en enfermedades pulmonares.

Tabla 8. Interpretación biológica SOCS3, DUSP1. Fuente: Propia autoría.

4.2 DISCUSIÓN

Los hallazgos de esta investigación evidencian que el modelo de regresión logística LASSO, utilizado en datos transcriptómicos filtrados por expresión diferencial y aportación a componentes principales, logró un desempeño sobresaliente en la categorización de las muestras (Accuracy = 98.87%, F1-Score = 98.79%).

Estos descubrimientos concuerdan con la bibliografía actual acerca de la elección de características en datos ómicos de alta dimensión, en la que técnicas de regularización como LASSO han probado ser especialmente eficaces para reconocer firmas genéticas sólidas a la vez que gestionan el sobreajuste (Tibshirani et al., 2008).

El elevado valor de F1-Score indica que el modelo conserva un balance ideal entre precisión y sensibilidad, aspecto crucial en usos biomédicos donde los falsos negativos pueden acarrear serias repercusiones (Saito & Rehmsmeier, 2015). La reducida desviación estándar detectada en las métricas después de 500 repeticiones señala una estabilidad significativa, corroborando que los genes elegidos recogen señales biológicas coherentes y no artefactos técnicos (Phipson et al., 2016).

Es importante resaltar que la estrategia de filtrado previo basado en la contribución al PCA mejoró la elección de genes informativos, coincidiendo con lo reportado por Abdi & Williams (2010) en estudios de expresión genética. Este método híbrido (filtrado + regularización) sobrepasa las restricciones de técnicas que emplean únicamente un criterio, tal como indican Boulesteix et al. (2017) en su estudio sobre la integración de datos ómicos.

Pese al buen desempeño del modelo, se detectaron elementos críticos. Si la relación entre clases (IC/PW) no fuera equitativa o estuviera balanceada, métricas globales como la precisión podrían incrementarse de manera artificial. Investigaciones simuladas indican que en grupos con ratios superiores a 4:1, el F1-Score resulta ser un indicador más fiable (Haixiang et al., 2017).

Se eligió ORA y GSEA como métodos para los análisis de enriquecimiento, estas son útiles para datos de alta dimensionalidad; según Harris et al., 2004 entre los dos existe una

complementariedad ya que pueden mostrar varios enfoques diferentes, lo que da como resultados datos más robustos de vías metabólicas, datos biológicos y enfermedades.

La sobreexpresión de SOCS3 refleja un esfuerzo del sistema inmunológico por regular la hiperinflamación Th2 (la cual es propia de las sibilancias), a través de de la vía JAK/STAT. No obstante, la señal proinflamatoria es tan potente que este mecanismo regulador es inútil, perpetuando el daño en los tejidos. Esta desregulación está vinculada con la resistencia a glucocorticoides, dado que altos niveles de SOCS3 ponen en riesgo la efectividad del tratamiento con esteroides en pacientes con asma severa (Yoshimura et al., 2007).

Además de los genes previamente conocidos y explicados en la sección del enriquecimiento, este estudio descubrió la presencia de dos genes anotados como novel transcripts (ENSG00000286276 - AL390066.2; ENSG00000285444 - AL162377.3), es decir, genes que no cuentan con una anotación previa en las bases de datos genómicas de referencia. Según Gamazon et al., 2018, estos genes previamente no caracterizados desempeñan roles importantes en diversos contextos fisiopatológicos, incluyendo respuestas inmunes, inflamación crónica y enfermedades pulmonares. Por ende, podríamos decir que, aunque estos genes no dispongan de una caracterización funcional definida, su detección en el contexto de esta condición clínica (sibilancias), podría abrir nuevas líneas de investigación para comprender su posible papel en la respuesta inmune, inflamación de las vías respiratorias o regulación génica diferencial.

Desde el punto de vista estadístico, la investigación de Fitzpatrick et al. (2024) identifica el desequilibrio de muestras entre los grupos estudiados como una restricción en su estudio. En cambio, esta tesis aplica un método de balanceo de datos para rectificar dicho sesgo, además de emplear validación cruzada y métricas de evaluación sólidas, lo que incrementa significativamente la confiabilidad de los resultados logrados y reduce el peligro de sobreajuste. Respecto al modelado matemático, el artículo no plantea herramientas con aplicabilidad diagnóstica, simplemente se limita a detallar los procesos inmunológicos observados. En cambio, este estudio sugiere un modelo predictivo fundamentado en regresión logística regularizada a través de Lasso, el cual muestra un elevado grado de exactitud y puede resultar beneficioso para profesionales clínicos al identificar sibilancias recurrentes vinculadas a sensibilización por aeroalérgenos.

Además, en el ámbito del análisis funcional, el trabajo de Fitzpatrick et al, se basa únicamente en la base de datos Reactome para interpretar los genes de interés. En cambio, esta investigación lleva a cabo un análisis de enriquecimiento funcional mucho más amplio, donde se incorporan diversas fuentes de información biológica, como Gene Ontology (GO), KEGG, Disease Ontology y Reactome, lo que facilita una comprensión más detallada e integral de los procesos biológicos, rutas metabólicas y enfermedades vinculadas a los genes expresados de manera diferencial. Por ejemplo, el estudio de estos mismos autores, observaron que los niños preescolares con sibilancia recurrente y sensibilización a aeroalérgenos presentan una disfunción en las respuestas de interferón tipo 1 (IFN-α/β) en neutrófilos, caracterizada por una regulación negativa basal de genes antivirales (como STAT2, MX1 y vías de señalización de IFN), pero una sobrerregulación tras la estimulación con poly (I:C). Esta respuesta alterada parece estar modulada por la IL-4, una citocina clave en la inflamación alérgica tipo 2, que suprime la expresión de genes de IFN tipo 1, pero que, ante un estímulo viral, induce una respuesta exagerada de estas mismas vías, junto con la activación de componentes de la señalización JAK/STAT (STAT1, STAT3, TYK2). Además, los neutrófilos de estos niños, al ser expuestos a poly (I:C), generan mediadores que inducen una respuesta proinflamatoria en células epiteliales de las vías respiratorias, con aumento en la expresión de genes como CCL2, CXCL10, IL-6, TIMP1 y VEGFA.

Ahora asociando nuestros resultados con el estudio de referencia, acotamos que, se destacan procesos como quimiotaxis leucocitaria (CXCL2, LGALS3), señalización de IL-10, IL-17 y TNF, y actividad de NF-κB, las cuales complementan y amplían los hallazgos del estudio original sobre la desregulación de IFN tipo 1 y JAK/STAT en neutrófilos de niños con sibilancia y alergia. La presencia de los genes SOCS3 y PTGS2 sugiere un intento fallido del sistema inmunológico por controlar la hiperinflamación impulsada por la IL-4 y las respuestas antivirales alteradas, lo que coincide con la mayor gravedad clínica observada en el estudio. Además, genes como DUSP1 y LGALS3 están vinculados a enfermedades crónicas (pulmonares, diabetes, aterosclerosis), lo que refuerza la idea de que estos mecanismos no solo afectan las vías respiratorias, sino que podrían tener implicaciones sistémicas a largo plazo. La activación de IL-17 y NF-κB en nuestros resultados sugiere que, además de la vía Th2 (IL-4), existe un componente Th17 que agrava la inflamación, ofreciendo nuevas dianas terapéuticas para los pacientes. Finalmente, podríamos decir

que estos hallazgos nos ofrecen un panorama más complejo de la interacción entre alergia, infecciones virales e inflamación crónica en niños con sibilancia recurrente.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

- Se desarrolló un modelo predictivo robusto basado en perfiles de expresión génica, capaz de distinguir con precisión a niños con sibilancias recurrentes asociadas a aeroalérgenos frente a aquellos sin dicha condición. El modelo implementado mediante regresión Lasso alcanzó una precisión (accuracy) de 0.9887 y un puntaje F1 de 0.9879, con una variabilidad mínima, lo que evidencia su fiabilidad y estabilidad en la clasificación.
- El estudio de la expresión diferencial mostró un subgrupo de genes significativamente regulados, los cuales participan en procesos inflamatorios mediados por quimiocinas, citocinas e interleucinas, lo cual corrobora su importancia biológica en la fisiopatología de las sibilancias.
- El balanceo de muestras que se realizó ayudó a mitigar la distribución desigual de los datos ya que se eliminó 48 muestras de la clase PW, para evitar el sesgo, que puede conllevar a problemas estadísticos, de sobreajuste en el modelo matemático y en problemas de representación de datos.
- La fusión de métodos estadísticos como Mahalanobis, PCA, la filtración de genes por su baja expresión, variabilidad y finalmente el modelo Lasso posibilitó una elección eficaz de genes, disminuyendo la dimensionalidad y optimizando el rendimiento del modelo sin pérdida de precisión.
- El enriquecimiento funcional basado en las bases de datos GO, KEGG, Disease Ontology y Reactome, posibilitó un contexto biológico para los genes detectados, vinculándolos con procesos inmunológicos, inflamatorios y patologías respiratorias, confirmando su utilidad como posibles biomarcadores.

- El modelo predictivo Lasso identificó genes clave cuya modulación está asociada significativamente con la presencia de sibilancias recurrentes. En particular, genes como SOCS3 y PTGS2 mostraron sobreexpresión en el grupo PW, lo que sugiere una activación persistente de rutas inflamatorias.

5.2. RECOMENDACIONES

- Incrementar la cantidad de muestras en estudios futuros para optimizar la generalización del modelo y valorar su rendimiento.
- En caso de desbalance significativo entre clases, se recomienda considerar estrategias de compensación como la aplicación de técnicas de sobremuestreo (SMOTE) o la asignación de pesos adaptativos durante el entrenamiento del modelo, con el objetivo de mejorar la sensibilidad.
- Como los datos transcriptómicos empleados en esta investigación se obtuvieron de una base pública (GEO), se sugiere que, en futuros estudios, los genes detectados como diferencialmente expresados y pertinentes al modelo predictivo sean validados en una cohorte independiente, es decir, un conjunto distinto de muestras biológicas o pacientes que no formaron parte de los datos originales. Esta validación permitiría comprobar si los genes detectados mantienen su patrón de expresión en condiciones biológicas similares, lo que fortalecería la reproducibilidad y aplicabilidad de los resultados. Además, se sugiere realizar estudios con muestras propias de pacientes locales, lo que permitiría adaptar y optimizar el modelo a las características genéticas de la población objetivo.

DECLARACIÓN DE DISPONIBILIDAD DEL CÓDIGO

Los scripts y datos utilizados en este trabajo se encuentras disponibles en el siguiente repositorio:

https://github.com/axl1441/TESIS-REYES-SANTANDER.git

BIBLIOGRAFÍA

Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2, 433-459. http://dx.doi.org/10.1002/wics.101

Akalin, A. (2020, septiembre 30). *5.13 Logistic regression and regularization*. Github.io. https://compgenomr.github.io/book/logistic-regression-and-regularization.html

Akdis, M. (2006). Therapies for allergic diseases in the 21st century. Nature Reviews Immunology, 6(2), 148-161.

Bacharier, L. B., Boner, A., Carlsen, K. H., собственик, P. C., Clark, J. E., Dahl, R., ... & Zar, H.J. (2014). Diagnosis and management of asthma in preschool children: a European Respiratory Society/American Thoracic Society consensus task force report. *European Respiratory Journal*, *43*(3), 668-685

Banchereau, J., Pascual, V., & O'Garra, A. (2011). From translational research to a new era of immunology: human immunology and the Immunological Genome Project. Immunity, 34(4), 550-565.

Barnes, P. J. (2008). The cytokine network in asthma and chronic obstructive pulmonary disease. *The Journal of Clinical Investigation*, *118*(11), 3546–3556. https://doi.org/10.1172/JCI36130

BioDatev. (2023, junio 23). Log Fold Change cálculo e interpretación. *BioDatev, la forma de hacer ciencia*. https://biodatev.com/log-fold-change

Boulesteix, AL., Wilson, R. & Hapfelmeier, A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. BMC Med Res Methodol 17, 138 (2017). https://doi.org/10.1186/s12874-017-0417-2

Bousquet, J., Lockey, R. F., Malling, H. J., & World Health Organization. (2001). Allergen immunotherapy: therapeutic vaccines for allergic diseases. Journal of Allergy and Clinical Immunology, 108(5 Suppl), S147-S334.

BxINFO L. L. C. (2024, abril 13). *DEGs: What are Differentially Expressed Genes?* Olvtools.com; BxINFO LLC. https://olvtools.com/en/documents/degs

Cansiz, S. (2023, marzo 6). *Mahalanobis distance & multivariate outlier detection in R*. Built In. https://builtin.com/data-science/mahalanobis-distance

Carow, B., & Rottenberg, M. E. (2014). SOCS3, a major regulator of infection and inflammation. *Frontiers in Immunology*, *5*, 58. https://doi.org/10.3389/fimmu.2014.00058

Carey, I. M., Critchley, J. A., DeWilde, S., Harris, T., Hosking, F. J., & Cook, D. G. (2018). Risk of infection in type 1 and type 2 diabetes compared with the general population: A matched cohort study. *Diabetes Care*, *41*(3), 513–521. https://doi.org/10.2337/dc17-2131

Castro-Rodríguez, J. A., Holberg, C. J., Wright, A. L., Martinez, F. D. (2000). A clinical index to define risk of asthma in young children with recurrent wheezing. American Journal of Respiratory and Critical Care Medicine, 162(4), 1403–1406. https://doi.org/10.1164/ajrccm.162.4.9912111

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., ... & Stein, L. (2011). Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Research, *39*(Suppl. 1).https://doi.org/10.1093/nar/gkq1018

Djukanović, R., Sterk, P. J., Fahy, J. V., & Hargreave, F. E. (2002). Standardised methodology for the collection and processing of induced sputum for cell analysis. American Journal of Respiratory and Critical Care Medicine, 165(4), 588-590.

Fitzpatrick, A. M., Teague, W. G., Meyers, D. A., Cruikshank, W. W., Busse, W. W., Castro, M., & Childhood Asthma Research and Education Network (CARE). (2011). Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of clinical, physiological, and inflammatory features in the National Heart, Lung, and Blood Institute Severe Asthma Research Program. Journal of Allergy and Clinical Immunology, 127(6), 1427-1433.e1-13.

Garcia, J. (2023). Métodos de enriquecimiento funcional para la evaluación de la significación biológica en análisis bioinformáticos. Uva.es. https://uvadoc.uva.es/bitstream/handle/10324/74184/TFG-G7390.pdf;jsessionid=A03A0ACB103D6279CA86125AAD1AF68D?sequence=1

Garcia-Moreno, A., López-Domínguez, R., Villatoro-García, J. A., Ramirez-Mena, A., Aparicio-Puerta, E., Hackenberg, M., Pascual-Montano, A., & Carmona-Saez, P. (2022). Functional Enrichment Analysis of Regulatory Elements. *Biomedicines*, *10*(3). https://doi.org/10.3390/biomedicines10030590

Gamazon, E.R., Segrè, A.V., van de Bunt, M. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat Genet* 50, 956–967 (2018). https://doi.org/10.1038/s41588-018-0154-4

Geeksforgeeks. (2025, enero 7). *Cross-validation in R programming*. GeeksforGeeks. https://www.geeksforgeeks.org/r-language/cross-validation-in-r-programming/

GINA (Global Initiative for Asthma). (2023). Global Strategy for Asthma Management and Prevention. https://ginasthma.org/

Godwin, James Andrew. *Ridge, LASSO, and ElasticNet Regression*. Publish AI, ML & data-science insights to a global community of data professionals., 2021, https://towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3/.

Global Initiative for Chronic Obstructive Lung Disease (GOLD). (2023). Global Strategy for the Diagnosis, Management, and Prevention of 1 Chronic Obstructive Pulmonary Disease. 2

Google for Developers. (2025, mayo 23). *Clasificación: ROC y AUC*. *Machine Learning.* https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. Bioinformatics and Biology Insights, 9(Suppl 1), 29–46. https://doi.org/10.4137/BBI.S28991

Hao, Y., Wang, B., Zhao, J., Wang, P., Zhao, Y., Wang, X., Zhao, Y., & Zhang, L. (2022). Identification of gene biomarkers with expression profiles in patients with allergic rhinitis. Allergy, Asthma, and Clinical Immunology: Official Journal of the Canadian Society of Allergy and Clinical Immunology, 18(1), 20. https://doi.org/10.1186/s13223-022-00656-4

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., … White, R. (2004). The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(DATABASE ISS.). https://doi.org/10.1093/nar/gkh036

High-Throughput Sequencing in Respiratory, Critical Care, and Sleep Medicine Research. An Official American Thoracic Society Workshop Report - National Institutes of Health (NIH), fecha de acceso: abril 7, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6812157/

IBM. (2025, enero 18). Qué es la regresión Lasso. *Ibm.com*. https://www.ibm.com/es-es/think/topics/lasso-regression

IBM. (2025, abril 1). ¿Qué es el análisis de componentes principales (PCA)? *Ibm.com*. https://www.ibm.com/es-es/think/topics/principal-component-analysis

Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y., & Ishiguro-Watanabe, M. (2025). KEGG: biological systems database as a model of the real world. Nucleic Acids Research, 53(D1), D672–D677. https://doi.org/10.1093/nar/gkae909

Lead Up Collective. (2017, junio 2). *Statistics & high performers: Studying the outliers*. Lead Up Collective. https://leadupcollective.org/2017/06/02/statistics-high-performers-studying-the-outliers/

Lee, H. Y., Park, J. W., Baek, S. H., Jung, Y. J., Kim, J. H., Jang, A. S., ... & Park, C. S. (2006). Gene expression profiling of nasal epithelial cells in patients with allergic rhinitis after house dust mite challenge. Journal of Allergy and Clinical Immunology, 117(6), 1287-1294.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Makrinioti, H., Fainardi, V., Bonnelykke, K., Custovic, A., Cicutto, L., Coleman, C., Eiwegger, T., Kuehni, C., Moeller, A., Pedersen, E., Pijnenburg, M., Pinnock, H., Ranganathan, S., Tonia, T., Subbarao, P., & Saglani, S. (2024). European Respiratory Society statement on preschool wheezing disorders: updated definitions, knowledge gaps and proposed future research directions. The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology, 64(3), 2400624. https://doi.org/10.1183/13993003.00624-2024

Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Rodriguez, E., & Cookson, W. O. (2010). A large-scale, genome-wide association study of asthma identifies new risk oci. Nature, 464(7289), 907-911.

Murel, J., & Kavlakoglu, E. (2025, febrero 18). ¿Qué es la regresión Ridge? *Ibm.com*. https://www.ibm.com/es-es/think/topics/ridge-regression

OMS. (2024, mayo 6). *Asma*. Who.int. https://www.who.int/es/news-room/fact-sheets/detail/asthma

Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. Ann Appl Stat. 2016 Jun;10(2):946-963. doi: 10.1214/16-AOAS920. Epub 2016 Jul 22. PMID: 28367255; PMCID: PMC5373812

Platts-Mills, T. A. E., Sporik, R., Wheatley, L. M., Hussain, M. A., & Heymann, P. W. (1997). The role of indoor allergens in asthma. New England Journal of Medicine, 337(12), 805-811.

Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10(3): e0118432. https://doi.org/10.1371/journal.pone.0118432

Savenije, O. E., Mahachie John, J. M., Granell, R., Kerkhof, M., Dijk, F. N., de Jongste, J. C., Smit, H. A., Brunekreef, B., Postma, D. S., Van Steen, K., Henderson, J., & Koppelman, G. H. (2014). Association of IL33-IL-1 receptor-like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood. *The Journal of Allergy and Clinical Immunology*, *134*(1), 170–177. https://doi.org/10.1016/j.jaci.2013.12.1080

Sempértegui R, Bautista P. Estudio descriptivo transversal: Asma en niños de 2 a 5 años identificados con los criterios API en dos hospitales de la ciudad de Cuenca en el periodo Junio 2015 – Enero 2016. Rev Med HJCA. 2020; 12(1): 30-37. DOI: http://dx.doi.org/10.14410/2020.12.1.ao.05

Sharma, P. (2017, September 13). article NGS: empowering infectious disease research beyond reality.Drugtargetreview.com. https://www.drugtargetreview.com/article/25646/ngs-empowering-infectious-disease-research-beyond-reality/

Tanigaki, K., & Honjo, T. (2007). Regulation of lymphocyte development by Notch signaling. *Nature Immunology*, *8*(5), 451–456. https://doi.org/10.1038/ni1453

Tenero, L., Piazza, M., & Piacentini, G. (2016). Recurrent wheezing in children. Translational Pediatrics, 5(1), 31–36. https://doi.org/10.3978/j.issn.2224-4336.2015.12.01

Tiew, P. Y., Meldrum, O. W., & Chotirmall, S. H. (2023). Applying next-generation sequencing and multi-omics in chronic obstructive pulmonary disease. International Journal of Molecular Sciences, 24(3). https://doi.org/10.3390/ijms24032955

Tipney, H., & Hunter, L. (2010). An introduction to effective use of enrichment analysis software. In Human Genomics (Vol. 4, Issue 3). https://doi.org/10.1186/1479-7364-4-3-202

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63.

Williams, D. J., Gautam, S., Creech, C. B., Jimenez, N., Anderson, E. J., Bosinger, S. E., Grimes, T., Arnold, S. R., McCullers, J. A., Goll, J., Edwards, K. M., Ramilo, O., & 16-0036 Study Team. (2025). Transcriptomic biomarkers associated with microbiological etiology and disease severity in childhood pneumonia. The Journal of Infectious Diseases, 231(2), e277–e289. https://doi.org/10.1093/infdis/jiae491

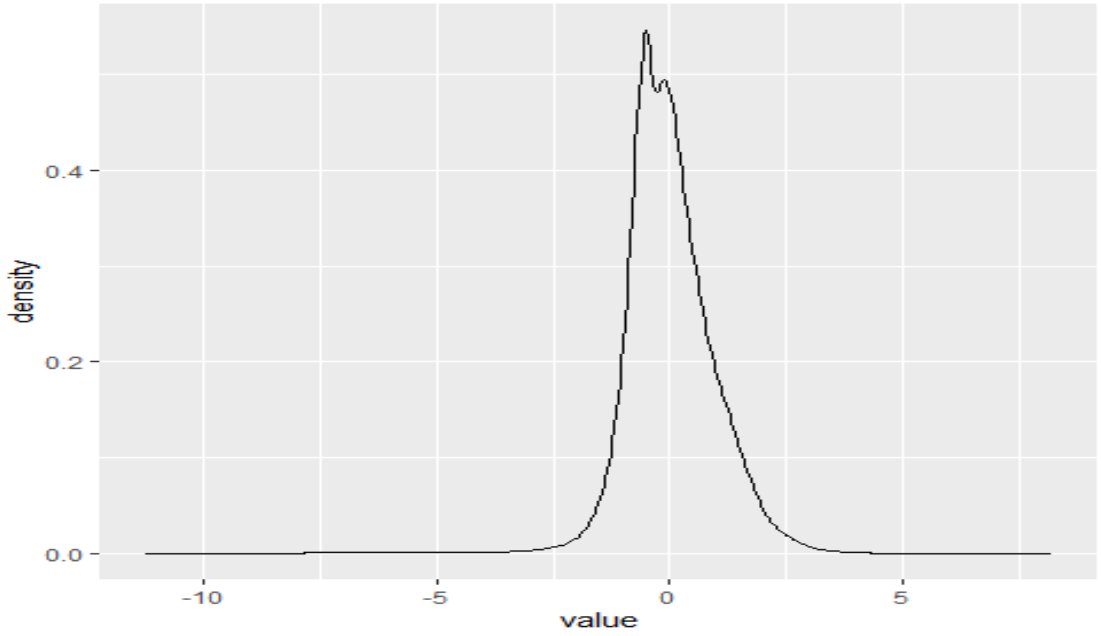
Yang, K. D., Ou, C.-Y., Chang, J.-C., Chen, R.-F., Liu, C.-A., Liang, H.-M., Hsu, T.-Y., Chen, L.-C., & Huang, S.-K. (2007). Infant frequent wheezing correlated to Clara cell protein 10 (CC10) polymorphism and concentration, but not allergy sensitization, in a perinatal cohort study. *The Journal of Allergy and Clinical Immunology*, *120*(4), 842–848. https://doi.org/10.1016/j.jaci.2007.07.009

Zhou, G., Soufan, O., Ewald, S. E., et al. (2021). Transcriptomic analysis reveals key immune dysregulation in pediatric asthma. Frontiers in Immunology, 12, 643776. https://doi.org/10.3389/fimmu.2021.643776

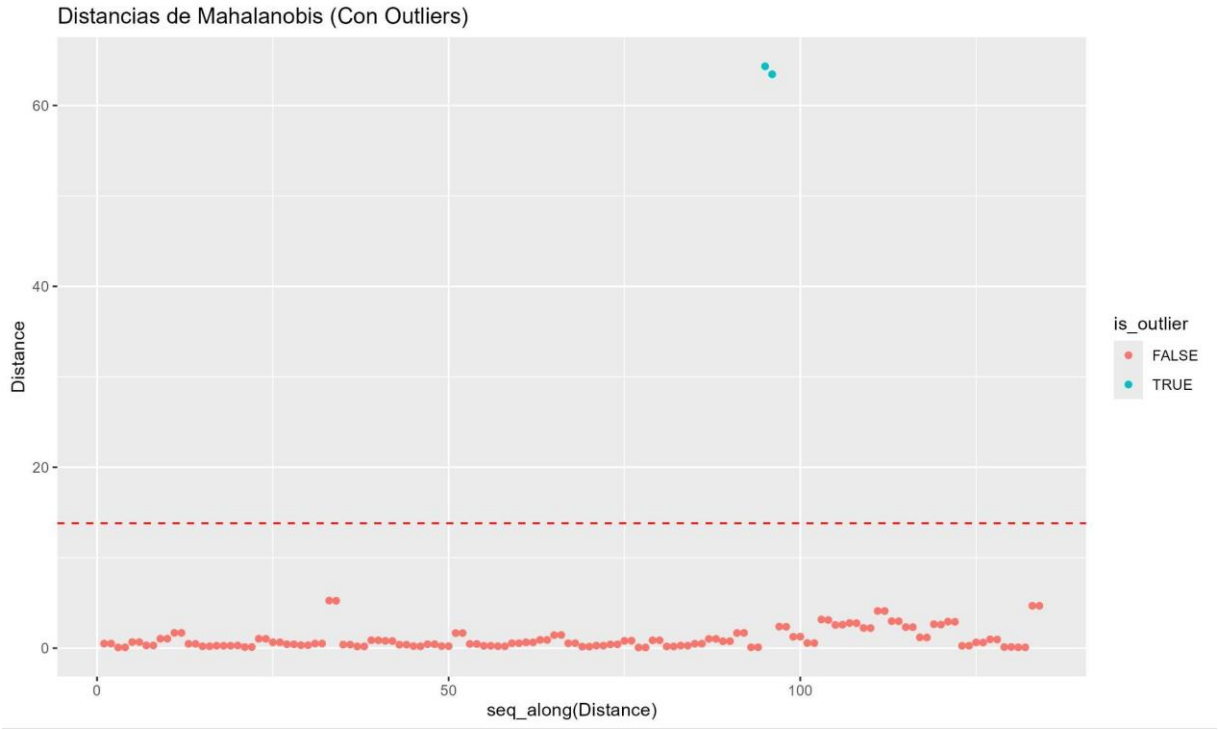
Zuurbier, C. J., Abbate, A., Cabrera-Fuentes, H. A., Cohen, M. V., Collino, M., De Kleijn, D. P. V., Downey, J. M., Pagliaro, P., Preissner, K. T., Takahashi, M., & Davidson, S. M. (2019). Innate immunity as a target for acute cardioprotection. *Cardiovascular Research*, *115*(7), 1131–1142. https://doi.org/10.1093/cvr/cvy304

ANEXOS

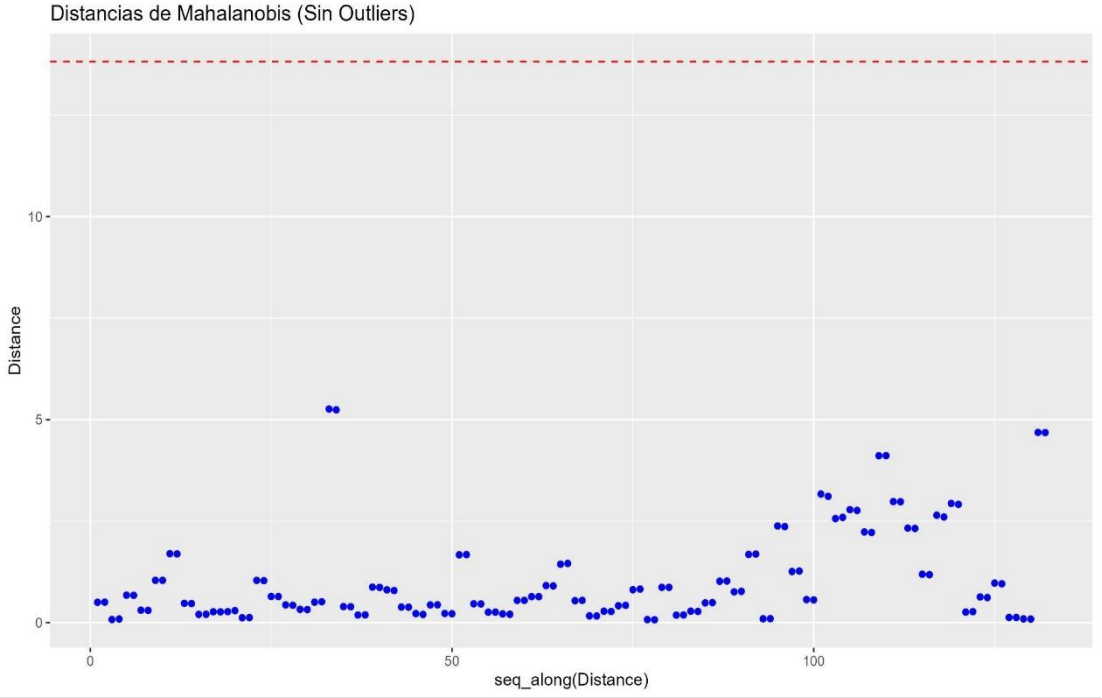
ANEXO 1. Distribución de densidad de los genes diferencialmente expresados antes de la reducción de dimensionalidad



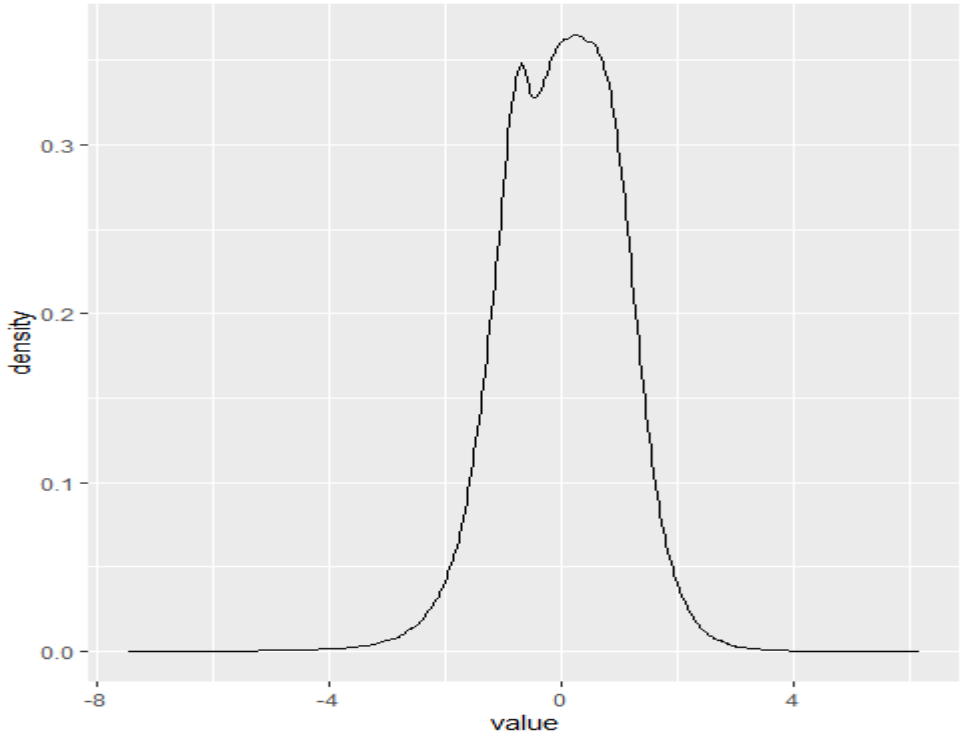
ANEXO 2. Distancia de Mahalanobis con outliers



ANEXO 3. Distribución de las distancias de Mahalanobis



ANEXO 4. Distribución de densidad de la expresión génica normalizada y transformada



ANEXO 5. Contribución de los genes basado en PCA balanceado