UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Ciencia Animal

Rendimiento de modelos machine learning en la identificación de especies microbianas causales

Trabajo Fin de Máster

Máster Universitario en Mejora Genética Animal y Biotecnología de la Reproducción

AUTOR/A: Duro Vizcaíno, Alba

Tutor/a: Ibáñez Escriche, Noelia

Cotutor/a: Casto Rebollo, Cristina

CURSO ACADÉMICO: 2023/2024

RESUMEN

El estudio del microbioma ha generado interés en el campo de la mejora genética debido a su influencia sobre caracteres clave en producción animal. Es por ello que es muy importante estudiar cómo el microbioma de los individuos afecta a la varianza fenotípica de estos caracteres e influye en su respuesta a la selección. Sin embargo, aunque hay estudios que consideran que la contribución del microbioma es similar a la del genoma, todavía estamos bastante lejos de descifrarlo. La herencia del microbioma es compleja y se ve influenciada por múltiples factores, ya que en sí podría considerarse como otro fenotipo del animal. Todavía no se ha establecido cuál es la mejor metodología para identificar aquellas especies bacterianas que influyen directamente sobre el fenotipo del animal. El uso de datos de simulación podría ayudar a determinar cuál de todas las metodologías propuestas en la literatura es la óptima.

El objetivo de este estudio es evaluar el rendimiento de diversos modelos de *machine learning* para identificar las especies microbianas con efecto en el fenotipo, así como evaluar su capacidad predictiva. Para ello se usaron datos del microbioma simulados con simuGMsel, una herramienta (basada en AlphaSimR) que permite simular la evolución del microbioma y el genoma de una población bajo selección. La herramienta simula el fenotipo de los individuos como una suma del efecto del genoma, del microbioma y del ambiente. En este estudio se simuló una población base de 1000 individuos y 600 especies bacterianas bajo un proceso de selección divergente por tamaño de camada durante 13 generaciones. De las 600 especies se asignaron 100 con efecto directo sobre el fenotipo simulado. Se simularon tres escenarios distintos: (i) escenario M donde el fenotipo depende únicamente del microbioma; (ii) escenario NMH (*Non-Microbial Heritability*) donde el microbioma no es afectado por el genotipo de los individuos; y un escenario HMH (*High-Microbial Heritability*) con un efecto alto (0.6) del genotipo de los individuos sobre el microbioma. Cada escenario fue simulado con una heredabilidad (h^2) del carácter de 0.15 y una microbiabilidad (m^2) variable de 0.15 o 0.5. También se varió el porcentaje de bacterias adquiridas del ambiente por la descendencia con valores de 0%, 20% y 50%. La matriz de abundancias microbianas tras 13 generaciones fue usada para ajustar modelos *machine learning* de clasificación por líneas (PLS-DA, Gaussian Naive Bayes, Random Forest y CatBoost) y cuatro modelos de regresión por fenotipo (PLS, LASSO, Random Forest y CatBoost). Se evaluó la capacidad predictiva de cada uno de ellos para clasificar las poblaciones divergentes o predecir directamente el fenotipo. Se determinaron las especies bacterianas que más contribuían a cada modelo, y se compararon con las especies con efecto simulado en el fenotipo. Posteriormente se realizaron distintas estrategias para reducir el número de especies identificadas sin efecto sobre el fenotipo, que fueron la combinación de metodologías, regresiones dentro de cada línea y repeticiones de los modelos alternando la composición de los grupos de testeo y entrenamiento.

Los resultados de los análisis mostraron que todos los modelos utilizados tuvieron un buen rendimiento de predicción, destacando los modelos de clasificación que predijeron la línea de los individuos con un 100% de precisión. Sin embargo, se obtuvieron porcentajes de falsos positivos (especies seleccionadas sin efecto en el fenotipo) muy elevados, de 81 ± 7% en modelos de clasificación y 75 ± 12% en modelos de regresión. Mediante las diferentes estrategias utilizadas, se logró reducir este porcentaje a un 66 ± 22% mediante combinación de modelos de regresión y a un 37 ± 20% en las repeticiones alternando los grupos de testeo y entrenamiento. Las regresiones dentro de línea permitieron identificar especies que no habían sido identificadas por los análisis iniciales.

Aunque la capacidad predictiva de los modelos fue moderadamente alta, la mayoría de las variables predictoras seleccionadas fueron especies sin efecto en el fenotipo. Estas especies fueron buenas predictoras debido a un efecto de deriva generado por la selección de los animales en cada generación. Una gran parte de estas especies sin efecto se fijaron en una línea, mientras que se perdieron en la otra. Otras son especies correlacionadas con las que tienen efecto en el fenotipo. Estos efectos, unidos a otros factores de confusión, dificultan la correcta identificación de las especies con efecto en el fenotipo, y promueven la selección de un gran número de falsos positivos. Combinar los resultados de los modelos de regresión, con y sin iteraciones cambiando los grupos de entrenamiento y testeo, y complementar estos resultados con regresiones dentro de cada línea puede ayudar a reducir el número de falsos positivos. Sin embargo, hay que tener cuidado a la hora de sacar conclusiones biológicas relacionadas con las variables seleccionadas por dichos modelos.

Palabras clave: microbioma, simulación, selección divergente, selección de variables, deriva

ABSTRACT

The study of microbiome has raised interest in the genetic improvement field due to its influence over key traits in animal production. As such, it is important to study the effects of an individual's microbiome on the traits' phenotypic variance and how it affects the response to selection. However, even though there are studies that consider that microbiome contribution is similar to genome contribution, we are still far from deciphering it. Microbiome heritability is complex, as it is influenced by multiple factors, and it could even be considered as another phenotypic trait by itself. It has not yet been established which methodology performs best to identify which bacterial species inluence the animal's phenotype directly. The use of simulated data may help determine which literature-proposed methodology is the most optimal.

The aim of this study is to evaluate the performance of various machine learning models to identify microbial species with an effect on the phenotype, as well as to assess their predictive capacity. For this purpose, microbiome data simulated with simuGMsel, a tool (based on AlphaSimR) that allows simulating the evolution of the microbiome and the genome of a population under selection, was used. The tool simulates the phenotype of individuals as a sum of the effect of the genome, the microbiome, and the environment. In this study, a base population of 1000 individuals and 600 bacterial species was simulated under a divergent selection process for litter size over 13 generations. Of the 600 species, 100 were assigned a direct effect on the simulated phenotype. Three distinct scenarios were simulated: (i) scenario M where the phenotype depends solely on the microbiome; (ii) scenario NMH (Non-Microbial Heritability) where the microbiome is not affected by the individuals' genotype; and a scenario HMH (High-Microbial Heritability) with a high effect (0.6) of the individuals' genotype on the microbiome. Each scenario was simulated with a heritability (h^2) of the trait of 0.15 and a microbiability (m^2) which varied between values of 0.15 or 0.5. The percentage of bacteria acquired from the environment by the offspring was also varied with values of 0%, 20%, and 50%. The microbial abundance matrix after 13 generations was used to fit four machine learning classification models (PLS-DA, Gaussian Naive Bayes, Random Forest, and CatBoost) and four phenotype regression models (PLS, LASSO, Random Forest, and CatBoost). The predictive capacity of each model to classify the divergent populations or directly predict the phenotype was evaluated. The bacterial species that most contributed to each model were determined and compared with the species with a simulated effect on the phenotype. Subsequently, different strategies were carried out to reduce the number of species identified without effect on the phenotype, which were the combination of methodologies, regressions within each line, and repetitions of the models changing the composition of the test and training groups.

The results of the analyses showed that all the models used had good predictive performance, with classification models predicting the line of individuals with 100% accuracy.

However, very high false positive (species selected without effect on the phenotype) rates were obtained, with percentages of 81 ± 7% false positives in classification models and 75 ± 12% in regression models. Through the different strategies used, this percentage was reduced to 66 ± 22% by combining regression models and to 37 ± 20% in repetitions changing the test and training groups. The regressions within each line allowed the identification of species that had not been selected by the initial analyses.

Although the predictive capacity of the models was moderately high, most of the predictor variables selected were species without effect on the phenotype. These species were good predictors due to a drift effect generated by the selection of animals in each generation. A large part of these non-effect species became fixed in one line, while they were lost in the other. Others are species correlated with those that influence the phenotype. These effects, combined with other confounding factors, make it difficult to correctly identify species with an effect on the phenotype and promote the selection of a large number of false positives. Combining the results of regression models, with and without iterations changing the training and test groups, and complementing these results with regressions within each line can help reduce the number of false positives. However, caution must be taken when drawing biological conclusions related to the variables selected by these models.

Keywords: microbiome; simulation; divergent selection; variable selection; machine learning

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El microbioma compone una parte esencial de los animales, con un tamaño similar al número total de células de los organismos. El microbioma contiene una gran diversidad de especies bacterianas, que además varían según el tipo de tejido que habitan (Rosenberg & Zilber-Rosenberg, 2018a). Actualmente, se sabe que el microbioma tiene un papel importante en la digestión, detoxificación (Cammack et al., 2018), el sistema inmune y en la salud (Kaur et al., 2023,Clavijo & Flórez, 2018) de humanos y animales. Al influir en estos procesos biológicos esenciales, el microbioma ha coevolucionado con humanos y animales (Hoffmann et al., 2016), y se ha transmitido de una población a la siguiente. La transmisión del microbioma está fuertemente influenciada por factores ambientales, siendo los principales el efecto materno, la dieta y el entorno (Rosenberg & Zilber-Rosenberg, 2018b). En mamíferos, la transmisión por parte de la madre sucede por varios mecanismos. La primera oleada de colonización bacteriana tiene lugar durante la gestación, debido a bacterias que logran pasar las barreras placentarias y amnióticas (Funkhouser & Bordenstein, 2013). Posteriormente hay una segunda oleada por el contacto con el microbioma vaginal e intestinal de la madre presente en el canal del parto (Senn et al., 2020a). Sobre este aspecto influye el modo de parto, dado que en los individuos nacidos por cesárea se interrumpe esta transmisión de bacterias de madre a hijo por el canal del parto (Bäckhed et al., 2015a) y en su lugar la cría adquiere especies bacterianas de la piel de la madre (Akagawa et al., 2019). La siguiente contribución directa de la madre está también ligada al efecto de la dieta. Durante la lactancia, el individuo adquiere especies bacterianas presentes en la leche materna (Gilbert, 2014). Los individuos que no se alimentan de leche materna tendrán una composición del microbioma más diferente a la de sus madres, y diferente a la de los individuos que sí lo hacen (Senn et al., 2020a). Por otra parte, las especies microbianas presentes en el entorno del animal al inicio de su vida son adquiridas por los individuos e influencian la composición de su microbioma (Spor et al., 2011), dado que ésta es diferente entre individuos crecidos en instalaciones cerradas e individuos en espacios abiertos (Mulder et al., 2009). Aunque hay indicios de que la aportación materna al microbioma de los hijos es, de media, más importante para su composición, no está claro realmente en qué proporciones se adquieren las especies bacterianas maternas y ambientales. En diversos estudios, realizados por Maqsood et al., (2019), Drell et al., (2017) y Bäckhed et al., (2015b) se obtuvieron porcentajes de especies transmitidas de la madre a los hijos variables, en un rango del 30-70%.

Tras esta adquisición inicial de especies bacterianas por los individuos, el microbioma evoluciona desde el nacimiento hasta que se estabiliza y se asemeja más al microbioma adulto (Senn et al., 2020b). Tanto durante este periodo de estabilización como cuando el individuo es adulto, el microbioma se ve afectado por múltiples factores (Fig. 1), entre los que se incluyen:

i) El crecimiento bacteriano: las especies bacterianas tienen un potencial de crecimiento propio dependiendo de varios factores, como el tejido en el que se encuentran o el sistema inmune del hospedador.

ii) El genoma del individuo: el microbioma se ve afectado por el genoma mediante locus del rasgo cualitativo (QTLs), que influyen en la abundancia de las distintas especies que componen el microbioma. También se han descrito genes relacionados con el metabolismo y el sistema inmune que afectan al desarrollo del microbioma (Spor et al., 2011). El genoma del individuo puede afectar a un 5-10% de la varianza del microbioma (Hall et al., 2017). A esta proporción de varianza genética que explica la varianza del microbioma se la define como heredabilidad microbiana (h_m^2).

iii) Las interacciones entre bacterias: las especies bacterianas del microbioma interaccionan entre sí mediante relaciones de simbiosis de diversos tipos (comensalismo, mutualismo o competencia por explotación, entre otras). Además, mediante las relaciones entre ellas, pueden ocurrir transferencias horizontales de genes que aumentan la variabilidad dentro de cada especie (Kern et al., 2021).

iv) Factores ambientales: el entorno y la dieta, entre otros, añaden variabilidad al microbioma tanto durante su desarrollo como en el microbioma adulto (Senn et al., 2020b).

v) El sistema inmune: muchos de los genes que afectan a la composición del microbioma son componentes del sistema inmune (Rosenberg & Zilber-Rosenberg, 2018b), y el propio sistema inmune limita la incursión de bacterias del ambiente en el organismo de los individuos (Cerutti et al., 2011).

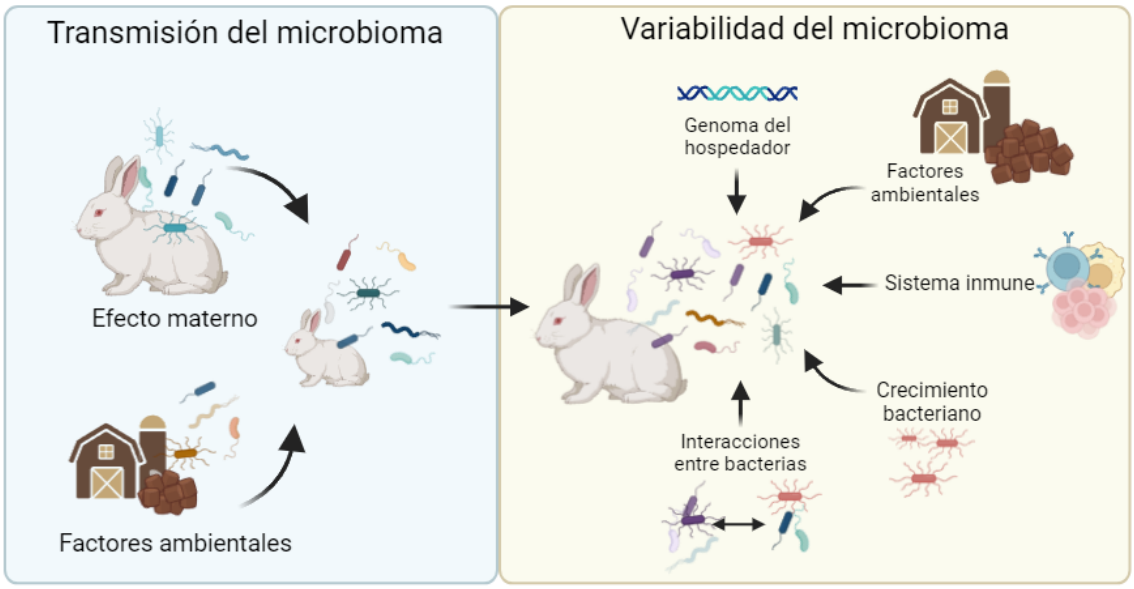

Figura 1. Esquema de la transmisión inicial del microbioma y de los factores que afectan a su desarrollo. El microbioma se hereda inicialmente de la madre (por transmisión placentaria, canal del parto

y leche materna) y *de factores ambientales* tales como la dieta y el entorno. El microbioma de los individuos cambia durante el crecimiento hasta que se estabiliza, influenciado por el genoma del hospedador, factores ambientales, el sistema inmune del hospedador, la capacidad de crecimiento de las propias baterías y por interacciones entre bacterias.

En diversas especies de producción animal se ha observado que el microbioma tiene un efecto sobre múltiples fenotipos de interés. Algunos ejemplos son el consumo de alimento en aves de corral, cerdos y rumiantes (Wessels, 2022,Bergamaschi et al., 2020), eficiencia alimentaria en pollos (Liu et al., 2021) y bovino (Cammack et al., 2018), el rendimiento de la canal en cerdo (Maltecca et al., 2021) y la grasa intramuscular (Martínez-Álvaro et al., 2021) y varianza ambiental en conejo (Casto-Rebollo et al., 2023a), entre otros. La varianza del microbioma explica una parte relevante de la varianza fenotípica, lo que se conoce como microbiabilidad (m^2) (Difford et al., 2018), y se han descrito valores de este parámetro en estos caracteres en diversos estudios (Tabla 1). Estos estudios han mostrado que el microbioma tiene un efecto entre bajo y moderado sobre los caracteres, y tiene valores diferentes para cada uno. Por estos motivos, el estudio del microbioma y la identificación de los géneros y especies que afectan a estos caracteres suscita interés en el área de mejora genética animal. Sin embargo, se trata de un carácter complejo que se ve influenciado por múltiples factores tanto en su transmisión como en su desarrollo, lo que dificulta su estudio.

Tabla 1. Parámetros de microbiabilidad en caracteres de especies de producción animal

Especie	Carácter	Microbiabilidad (m^2)	Referencia
Porcino	Grasa intramuscular	0.03-0.06	(Khanal et al., 2021)
	Ganancia diaria promedio de la canal	0.18	(Khanal et al., 2021)
	Consumo de alimento residual	0.11-0.12	(Aliakbari et al., 2022)
Bovino	Consumo residual de alimento	0.19	(Martinez Boggio et al., 2024)
	% Proteico de la leche	0.08	(Buitenhuis et al., 2019)
	% Grasa de la leche	0.08	(Buitenhuis et al., 2019)
Ovino	Contenido proteico de la leche	0.03	(Boggio et al., 2023)
Gallina ponedora	Índice de conversión	0.03	(Zhou et al., 2023)
	Consumo de alimento residual	0.16	(Zhou et al., 2023)

1.1. ANÁLISIS DE DATOS DE MICROBIOMA

La composición del microbioma de los individuos se puede identificar mediante la secuenciación del rRNA 16S (Schloss & Westcott, 2011) o mediante técnicas de secuenciación masiva del metagenoma (Breitwieser et al., 2019). Mediante la secuenciación del rRNA 16S, las diferentes especies del microbioma se identifican agrupando las lecturas secuenciadas con una similitud mayor al 97% en grupos llamados unidades taxonómicas operativas (OTUs) (Schloss & Westcott, 2011). Este método tiene un coste computacional alto (Schloss & Handelsman, 2005), y, además, el uso del 97% para determinar diferencias entre especies es cuestionable dado que asume que la diferencia entre grupos taxonómicos es del 3% (Konstantinidis et al., 2006). Por estos problemas se desarrolló otra aproximación, mediante el uso de variantes de secuencias de amplicón (ASVs) que agrupan las secuencias que solo se diferencian en un nucleótido, de forma que no se asume cuál es la variabilidad entre taxones (Caruso et al., 2019). Sin embargo, mediante la secuenciación del rRNA 16S no se pueden determinar los genes bacterianos y sus funciones. En este aspecto, la secuenciación masiva del metagenoma proporciona más información sobre la composición del microbioma. Estas técnicas permiten conocer la taxonomía y la información genética de todos los microorganismos de un entorno específico (Breitwieser et al., 2019). Esta aproximación se hace por secuenciación *shotgun*, secuenciando fragmentos de ADN al azar, y las lecturas obtenidas se ensamblan en fragmentos más grandes llamados *contigs*. Los *contigs* se alinean con secuencias presentes en bases de datos y se determinan su taxonomía y la función del gen al que pertenecen (Lapidus & Korobeynikov, 2021). Tras la secuenciación, se determinan las abundancias totales de cada grupo taxonómico utilizando el número de lecturas de cada OTU, ASV o *contigs*, y así se obtiene la base de datos final del microbioma.

Las abundancias absolutas de cada especie presente en el microbioma son imposibles de obtener, debido al sesgo generado por todo el proceso de obtención de muestra y de las técnicas de secuenciación, que restringen la cantidad de lecturas que se pueden obtener. Las abundancias de las especies detectadas por estas técnicas no deben entenderse como datos de abundancia absoluta sino relativa al resto de especies presentes. (Gloor et al., 2017a). A este tipo de datos, en los que la información proviene de las abundancias relativas y no absolutas, se les denomina datos composicionales (Fernandes et al., 2014). Para tener en cuenta la composicionalidad de los datos existen diversas aproximaciones matemáticas que permiten lidiar con ello. Una forma válida de hacer esto, propuesta por Aitchison (1982), es mediante transformaciones logarítmicas. Se pueden utilizar distintas transformaciones: (i) *centered log-ratio* (CLR), (ii) *isometric log-ratio* (ILR) y (iii) *additive log-ratio* (ALR). El CLR transforma las variables expresándolas en función de la media geométrica. La media geométrica incluye a todas las variables, por lo que la transformación es isométrica, es decir, que tiene la misma geometría que la de todos los pares de log-ratios (Aitchison, 1982). Esta transformación es muy útil a nivel computacional, pero su interpretación

es más complicada dado que todas las variables están incluidas en el denominador (Greenacre et al., 2021). El ILR expresa las variables según las medias geométricas de amalgamaciones de variables relacionadas entre ellas. También es una transformación isométrica, pero es costosa a nivel computacional y tiene la dificultad de necesitar saber qué grupos de variables son interesantes para realizar los ratios (Greenacre et al., 2021). El ALR expresa las variables en función de una sola variable escogida entre toda la base de datos (Aitchison, 1982), siendo la transformación más sencilla de interpretar (Greenacre, 2018). Un problema del ALR a tener en cuenta es que no es estrictamente isométrico, por lo que se debe comprobar que la transformación mantiene su geometría asegurando que su correlación procruster con datos transformados isométricos es elevada (Greenacre et al., 2021). Otro factor a tener en cuenta al realizar esta transformación es la elección de la variable de referencia, que debe cumplir con tres requisitos: (i) tener un coeficiente de variación bajo, (ii) tener un alto número de lecturas y (iii) que la transformación mantenga una elevada correlación procruster (Greenacre, 2018,Greenacre et al., 2021).

Al ser necesario realizar transformaciones logarítmicas debido a la composicionalidad de los datos, surge el problema de tener que lidiar con los datos ausentes, dado que el log(0) no está definido. Los datos de microbioma en particular suelen presentar una gran cantidad de datos ausentes (Paulson et al., 2013). Estos ceros son principalmente de dos tipos. Por un lado, están los ceros biológicos o estructurales, que son aquellos causados por la biología subyacente y que por tanto son reales (Kaul, Davidov, et al., 2017). Por otro lado, están los ceros técnicos, causados por los procesos de preparación de la muestra (Silverman et al., 2020). Por último, están los ceros de muestreo, causado por los límites de los procesos de secuenciación para obtener lecturas (Kaul, Mandal, et al., 2017), por los que muchas especies que están presentes en bajas cantidades no son detectadas (Silverman et al., 2020). Hay diversas estrategias para eliminar los ceros sin que eso afecte a los resultados. Una opción sencilla y más comúnmente utilizada es reemplazar los ceros por valores positivos pequeños, sumando una lectura a todos los datos (Greenacre, 2018). Otras aproximaciones incluyen sustituir los ceros por la media de las variables o por el valor más pequeño de la variable, modelos de imputación bayesiana, regresiones generalizadas o regresiones lineales múltiples, entre otros (Lubbe et al., 2021). La gran cantidad de ceros añade otros problemas a la hora de analizar los datos de microbioma. El exceso de ceros en la base de datos altera su distribución, creando picos en el 0 y alejándola de la normalidad que muchos modelos estadísticos asumen al realizar los análisis (Jiang et al., 2019). A causa de la escasez de los datos, puede ser necesario realizar filtros eliminando las especies menos abundantes (por ejemplo, presentes en menos del 70% u 80% de las muestras) para mitigar el efecto de los ceros. Al realizar estos filtros se asume que las especies más relevantes para el carácter estudiado son muy abundantes, mientras que las de menor importancia son poco abundantes y su eliminación

en un filtrado no afecta a los resultados y de hecho beneficia al rendimiento de los análisis (Busato et al., 2023).

Las dos aproximaciones estadísticas más se clásicas para analizar datos del microbioma son los análisis de abundancia diferenciales y los índices de alfa- y beta-diversidad (Lutz et al., 2022). Los análisis de abundancia diferenciales estudian si hay diferencias entre el microbioma de varios grupos de muestras. Dentro estos tipos de análisis uno de los más utilizados son los t-test corregidos por el *false Discovery rate* (*FDR*), que permiten identificar diferencias significativas en la abundancia de especies bacterianas entre grupos, controlando los falsos positivos (Benjamini & Hochberg, 1995). Mediante la alfa- y beta-diversidad se identifican diferencias en el microbioma dentro de una población (alfa-diversidad) y entre poblaciones (beta-diversidad). Las medidas más comunes para estos índices incluyen los índices de Shannon para alfa-diversidad (Lemos et al., 2011) y la distancia de Bray-Curtis (Lemos et al., 2011) para beta-diversidad, entre otras. Una alternativa al uso de los análisis estadísticos clásicos son los algoritmos *machine learning* (ML), que son metodologías de análisis computacionales y estadísticos que construyen y adaptan modelos para mejorar su rendimiento. Es decir, hacen inferencias sobre los datos para aprender y predecir en nuevos datos. Las metodologías ML presentan ciertas ventajas sobre los análisis estadísticos clásicos: pueden inferir relaciones entre variables para la identificación de patrones y son capaces de lidiar con datos multidimensionales (Marcos-Zambrano et al., 2021), por lo que son útiles para bases de datos de la magnitud del microbioma. El uso de modelos ML para identificar variables causales en el microbioma conlleva principalmente tres pasos: (i) selección del tipo de modelo a utilizar, (ii) optimización del modelo y (iii) selección de variables.

Los modelos ML se pueden clasificar en no supervisados o supervisados (Namkung, 2020). Los modelos no supervisados realizan agrupaciones de las variables basándose en similitudes entre las muestras creando clústeres o realizando reducciones de dimensionalidad. Los modelos supervisados utilizan datos con resultados conocidos para realizar las predicciones. Dentro de los modelos supervisados se pueden dividir entre modelos de clasificación y modelos de regresión dependiendo del tipo de dato que se predice. Los modelos de clasificación realizan una predicción sobre el grupo o clase a la que pertenecen las muestras, utilizando valores discretos; mientras que los modelos de regresión predicen sobre un valor continuo, por ejemplo, el fenotipo del individuo (Namkung, 2020).

Los modelos supervisados son los más utilizados en datos de microbioma. Algunos de los modelos más comunes son:

i) Regresiones penalizadas: estos modelos seleccionan las variables más importantes mientras se construye el modelo, penalizando los coeficientes de regresión de las variables de menor importancia para la predicción. La penalización puede ser de tipo L1,

usada por el modelo *least absolute shrinkage and selection operator* (LASSO), que elimina por completo las variables poco importantes; o tipo L2, usada por la regresión *Ridge*, que reduce los coeficientes menos importantes a valores cercanos a 0 pero no los elimina (Namkung, 2020,Tibshirani, 1996).

ii) Modelos basados en árboles de decisión: estos modelos agrupan las variables en grupos, llamados árboles, a partir de los cuales sacan más grupos de variables hasta determinar la combinación de variables que mejor predice el resultado. Hay múltiples modelos basados en esto, como Random Forest o CatBoost, que también utiliza *gradient boosting* y aprende de los errores de cada árbol generado para mejorar el siguiente (Prokhorenkova et al., 2017).

iii) Modelos probabilísticos: los modelos probabilísticos realizan las predicciones basándose en distribuciones de probabilidad. A este grupo pertenece el modelo Gaussian Naïve Bayes (GNB), basado en el teorema de Bayes, que calcula la probabilidad de que cada variable pertenezca a un grupo u otro, asumiendo que los datos siguen una distribución normal y que las variables son independientes (Domingos & Pazzani, 1997).

iv) Modelos lineales con reducción de dimensionalidad: a este grupo pertenecen el *partial least squares regression* (PLS) y *partial least squares discriminant analysis* (PLS-DA). Ambos combinan la reducción de dimensionalidad con una regresión múltiple (PLS) o un análisis discriminante (PLS-DA) (Barker & Rayens, 2003).

Para el ajuste y entrenamiento de los modelos ML, la base de datos se divide en un set de entrenamiento y un set de testeo, típicamente en proporciones 70/30 o 75/25. De esta forma, el modelo se optimiza y ajusta sobre el set de entrenamiento, y su rendimiento se evalúa sobre el set de testeo (Papoutsoglou et al., 2023). Cada modelo de ML tiene sus propios hiperparámetros que deben ser optimizados para realizar la mejor predicción posible según la medida de rendimiento que se elija. Hay múltiples formas de optimizar los hiperparámetros, como la búsqueda aleatoria (asigna valores al azar a cada parámetro hasta encontrar la mejor combinación), la búsqueda en cuadrícula (determina la mejor combinación dentro de parámetros indicados por el usuario) u optimización bayesiana (construye un modelo probabilístico que decide qué parámetros evaluar), entre otros (Feurer & Hutter, 2019). Los hiperparámetros se ajustan sobre el set de entrenamiento mediante validación cruzada de 5-10 *folds* (valor que puede ajustarse dependiendo del tamaño de los datos). Este proceso de optimización y división de los datos es problemático cuando el número de muestras es muy reducido, dado que el rendimiento del modelo sobre el grupo de testeo será muy variable (Papoutsoglou et al., 2023).

Mediante estos métodos se pueden llevar a cabo procesos de selección de variables para determinar qué especies o géneros del microbioma tienen un efecto sobre los fenotipos de interés. Cada modelo tiene varios posibles métodos de selección. Algunos, como el modelo LASSO,

seleccionan las variables a la vez que construyen el modelo. En otros, se deben establecer criterios como la importancia de variable en modelos basados en árboles de decisión (Genuer et al., 2010). Sin embargo, estos análisis se ven influenciados por el tamaño de la muestra, las correlaciones entre las variables, la normalización de los datos, el tipo de modelo a utilizar, la optimización de parámetros y los procesos de selección de variables (Moreno-Indias et al., 2021). Debido a la cantidad de opciones y variabilidad entre modelos ML y dentro de cada modelo, no hay un consenso claro sobre qué metodología es la más adecuada para analizar los datos de microbioma e identificar las especies bacterianas que influyen sobre los fenotipos de interés. Por ello, el uso de datos de simulación podría ayudar a determinar cuál de las metodologías propuestas en literatura es la más óptima.

1.2. SIMULACIÓN DE DATOS DE MICROBIOMA

En este estudio se ha utilizado SimuGMSel, una herramienta que permite simular la evolución del genoma, microbiota y fenotipo de una población sometida a un programa de selección fenotípica a través de varias generaciones (Casto-Rebollo et al., 2022). SimuGMSel está programada en R, utilizando funciones del paquete AlphaSimR (Gaynor et al., 2021). La simulación se realiza teniendo en cuenta todos los factores que afectan a la herencia y desarrollo del microbioma.

En la simulación, el valor fenotípico de cada individuo depende del genoma, la microbiota y el ambiente (Figura 2) siguiendo la siguiente fórmula:

$$y_i = \mu_p + gv_i + mv_i + e_i \ , \quad (1)$$

donde y_i es el fenotipo del individuo i, μ_p es la media fenotípica de la población base, gv_i es el valor genético de un individuo, mv_i es el valor de la microbiota del individuo y e_i es el residuo. Los efectos genéticos gv_i y de la microbiota mv_i sobre el fenotipo están regulados por un valor de heredabilidad (h^2) y microbiabilidad (m^2), que es la proporción de varianza fenotípica determinada por el genoma y la microbiota, respectivamente. El efecto genético gv se simula como:

$$gv_i = \sum_{j=1}^{n} z_{ij}\alpha_j \ , \quad (2)$$

donde z_{ij} es el genotipo del individuo i para el SNP j, y α_j es el efecto del alelo de sustitución del locus del rasgo cualitativo (QTLs). Por otra parte, el efecto de la microbiota mv_i se simuló como:

$$mv_i = \sum_{k=1}^{m} x_{ik}\omega_k \ , \quad (3)$$

donde x_{ik} es la abundancia de la especie k en el individuo i, y ω_k es el efecto de la especie k sobre el fenotipo del individuo. Este efecto ω se asigna a un grupo de especies dentro del microbioma, y puede ser tanto positivo como negativo.

La interacción entre el genoma y el microbioma se regula mediante el parámetro de heredabilidad de la microbiota (h_m^2), que determina el efecto de los QTLs del genoma del individuo sobre la abundancia de las especies de su microbiota. Este efecto se asigna a un grupo de bacterias, de las cuales la mitad también pertenecen al grupo de especies con efecto sobre el fenotipo.

La herramienta establece una población inicial de especies de bacterias. Dado que la microbiota se encuentra en los individuos y en el ambiente, se genera una distribución de las especies tanto a nivel de microbioma parental como a nivel de microbioma ambiental. De todas las especies se establece un grupo que se encuentran presentes en la población base, mientras que el resto solo pueden ser adquiridas a través del ambiente. Por otra parte, la herramienta tiene en cuenta la complejidad de la herencia de la microbiota, simulando tanto la transmisión inicial de especies de madres a hijos como la adquisición de especies del ambiente. Se regula, para cada camada, cuántas especies bacterianas se heredan de la madre y cuántas se heredan del ambiente. En ambos casos, las especies heredadas dependen de la abundancia de las especies en la madre o en el ambiente.

Finalmente, una vez establecidas las poblaciones con sus valores genéticos, fenotípicos y de la microbiota, la simulación realiza un programa de selección fenotípica en líneas divergentes. Al final de la simulación se tienen dos poblaciones: una seleccionada para fenotipo alto y otra para fenotipo bajo. De ambas poblaciones se obtienen los datos de abundancia de cada especie de la microbiota simulada, con las cuales se pueden realizar los análisis de identificación de especies casuales en el microbioma y determinar cuál es la metodología más óptima para lidiar con este tipo de datos.

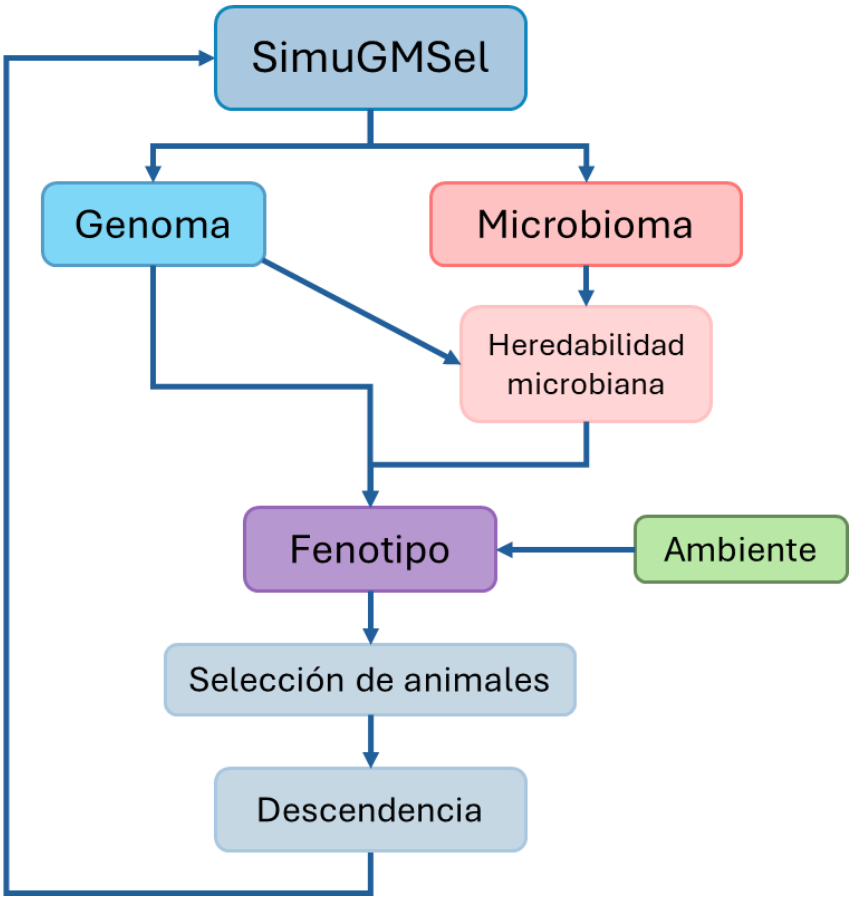

Figura 2. Esquema de funcionamiento de SimuGMSel. El fenotipo se computa a partir del genoma, el microbioma y el ambiente. El genoma tiene un *efecto directo* sobre el fenotipo y un efecto sobre la heredabilidad microbiana. El microbioma tiene un efecto directo sobre el fenotipo, que se ve afectado por el genoma. Una vez generado el fenotipo, se seleccionan los animales para producir la descendencia de la siguiente generación.

2. OBJETIVOS

El objetivo de este estudio es determinar qué modelo *machine learning* (ML) tiene un mejor rendimiento y una mayor capacidad para la identificación de las variables del microbioma que tienen un efecto sobre caracteres de interés en producción animal. Para ello se plantearon los siguientes objetivos parciales:

i) Simular dos poblaciones divergentes diferenciadas por la composición del microbioma, con especies microbianas que afectan positiva y negativamente a sus fenotipos, bajo diferentes escenarios.
ii) Ajustar modelos ML de clasificación y de regresión y seleccionar las variables más importantes consideradas por cada uno, evaluando el número de verdaderos y falsos positivos en cada caso.
iii) Utilizar diversas estrategias para reducir la cantidad de falsos positivos y determinar las mejores aproximaciones y modelos ML.

3. MATERIALES Y MÉTODOS

3.1. SIMULACIÓN DE LOS DATOS

Los datos de este estudio fueron simulados con la herramienta simuGMSel (Casto-Rebollo et al., 2022) basada en el paquete de R *AlphaSimR* (Gaynor et al., 2021). La población fundadora fue simulada usando la historia demográfica de la especie *Oryctolagus Cunniculus* según Carneiro et al., (2014), generándose una población base de 1000 individuos. El carácter simulado fue tamaño de camada (TC) siguiendo una distribución normal con una media (μ_p) 8 y varianza (σ_p) de 6. El genoma de los animales se simuló con un total de 21 autosomas y 127 Mb con 10000 sitios de segregación por cromosoma. La heredabilidad (h^2) del carácter se fijó a 0.15. Se consideraron 100 QTLs por cromosoma con efecto aditivo en el fenotipo. El mínimo de sitios de segregación en cada cromosoma fue de 1 para los QTLs y 4250 para los SNPs. El microbioma se compuso por 1000 especies, con abundancias iniciales (x_k) generadas por una distribución binomial negativa $x_k \sim$ NB (r=2, q=0.0001), de las cuales 600 se fijaron en la población base. De las 600 especies presentes en la población base, a 100 se les asignó un efecto sobre el fenotipo, el cual siguió una distribución gamma $\omega \sim \Gamma$ (k=1.4, θ=3.4). La población se sometió a un proceso de selección fenotípica divergente durante 13 generaciones para alto y bajo TC. En el programa de selección, se seleccionaron un total de 125 hembras y 25 machos en cada generación para cada línea (Blasco et al., 2017).

3.1.1. Escenarios

El fenotipo del tamaño de camada se simuló considerando diferentes escenarios, según los efectos que contribuyen a dicho fenotipo; Genoma (G), Microbiota (M) y Ambiente (A):

- Escenario microbiota (M): el fenotipo se calcula únicamente en función del efecto de la microbiota y el ambiente, sin efecto del genoma.
- Escenario sin heredabilidad microbiana (NMH; *Non-microbial heritability*): el fenotipo se calcula en función de la microbiota, el genoma y el ambiente. No hay efecto del genoma del individuo sobre la abundancia de la microbiota (heredabilidad microbiana).
- Escenario de alta heredabilidad microbiana (HMH; *High microbial heritability*): el fenotipo se calcula en función del efecto de la microbiota, el genoma y el ambiente, considerando una heredabilidad microbiana de 0.6. En este escenario se estableció que 100 de las 1000 especies estaban influenciadas por el genoma del individuo, forzándose a que al menos 50 fueran especies con efecto en el fenotipo. El efecto de los QTLs sobre la microbiota se generó mediante una distribución gamma $\beta \sim \Gamma$ (k=0.2, θ=1).

Cada escenario se simuló bajo distintos parámetros de microbiabilidad (m^2), y se determinaron diferentes porcentajes para la proporción de especies adquiridas del ambiente (EM) y de la madre (PM) por cada individuo en cada generación. Los valores utilizados para la m^2

fueron de 0.15 y 0.5; y los valores de EM y PM variaron según las siguientes proporciones (Tabla 1):

Tabla 2. Proporciones de especies adquiridas por el ambiente (EM) y de la madre (PM)

	EM	PM
Ambiente alto	0.5	0.5
Ambiente bajo	0.2	0.8
Sin Ambiente	0	1

3.2 PROCESADO DE DATOS

Los datos fueron procesados usando el lenguaje de programación R (R Core Team, 2023). Primero, se eliminaron todas las especies bacterianas que se encontraban ausentes en todos los individuos de ambas poblaciones. Aunque las abundancias microbianas simuladas son las abundancias reales esperadas, se realizó una transformación de los datos para emular la metodología utilizada en este tipo de análisis (Casto-Rebollo et al., 2023; Zubiri-Gaitán et al., 2023), ya que se consideran datos composicionales. Por esta razón, se aplicó un "*additive logarithm-ratio*" (ALR) (Gloor et al., 2017b). Antes de realizar esta transformación, se añadió un *count* a toda la base de datos (Greenacre, 2018). Los ALR fueron calculados según Greenacre et al., (2021):

$$ALR(j|ref) = \log\left(\frac{X_j}{X_{ref}}\right),\ j = 1, \dots, J,\ j \neq ref, \quad (1)$$

siendo X_{ref} la abundancia de la variable de referencia y X_j la abundancia de la especie j. Como variable de referencia se utilizó aquella especie con el menor coeficiente de variación y con una abundancia media elevada. Posteriormente, se evaluó si la base de datos mantenía la isometría en sus medidas a través del cálculo de la correlación *procrustes* usando los paquetes de R *easyCODA* 0.34.3 (Greenacre, 2018) y *vegan* 2.6-4 (Oksanen J et al, 2022). Finalmente, los datos transformados se centraron y escalaron a una media de 0 y desviación típica de 1. Se realizaron dos análisis de componentes principales (PCA) para comprobar si la variabilidad observada en la matriz de microbiota permitía diferenciar entre los individuos de las poblaciones divergentes. El primero se realizó con el microbioma completo, y el segundo sólo con las especies con efecto en el fenotipo. Además, se realizó un estudio de correlaciones entre las especies con y sin efecto en el fenotipo. Las visualizaciones gráficas de estos dos estudios se realizaron con los paquetes de R *factoextra* 1.0.7 (Kassambara, A. 2020) y *ComplexHeatmap* (Gu et al., 2016).

3.3. MODELOS *MACHINE LEARNING*

Diferentes metodologías machine learning (ML) se utilizaron para desarrollar modelos predictivos para clasificar las líneas divergentes y predecir el TC usando la matriz de abundancias

de la microbiota. Para ello, se emplearon cuatro algoritmos de clasificación: *Partial Least Squares Discriminant Analysis* (PLS-DA) (Barker & Rayens, 2003), *Gaussian Naïve Bayes* (GNB) (Domingos & Pazzani, 1997), Random Forest (Breiman, 2001) y CatBoost (Prokhorenkova et al., 2017); y cuatro modelos de regresión y predicción del fenotipo: *Partial Least Squares* (PLS), *Least Absolute Shrinkage and Selection Operator* (LASSO) (Tibshirani, 1996), *Random Forest* y CatBoost. Para cada uno de los escenarios, la base de datos del microbioma se dividió en un set de entrenamiento y un set de testeo (Figura 1), según una ratio de 70/30. Los hiperparámetros de cada algoritmo fueron optimizados en el set de entrenamiento usando una validación cruzada de 5 *folds.* Los hiperparámetros que se ajustaron fueron los siguientes: el número de componentes para PLS-DA y PLS; el suavizado en GNB; el parámetro λ en LASSO; el número de árboles, profundidad máxima, número máximo de nodos por árbol y el mínimo de variables por partición en Random Forest; y el número de iteraciones, tasa de aprendizaje y profundidad en CatBoost. La capacidad predictiva de los modelos se evaluó en el set de testeo (Figura 1). Para los modelos de clasificación se utilizó la precisión y para los modelos de regresión se utilizaron el error cuadrático medio (RMSE) y la Q^2. Los análisis por PLS-DA, PLS y LASSO se realizaron en R con el paquete *mixOmics* 6.24.0 (Rohart et al., 2017) y *glmnet* 4.1-8 (Friedman et al., 2010), respectivamente. Los modelos GNB, Random Forest y CatBoost se realizaron en Python 3.11.5, con el módulo Sci-kit learn 1.3.0 (Pedregosa et al., 2012).

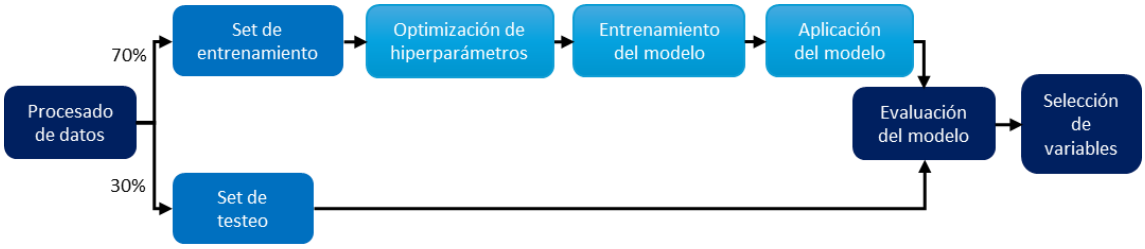

Figura 3. Esquema de procesado de datos y aplicación de los modelos. Los datos procesados se separan en dos grupos: un 70% conforman el set de entrenamiento y el 30% restante el set de testeo. Con el primer set se optimizan los hiperparámetros, se entrenan y se aplican los modelos. Con el segundo set se evalúa la capacidad predictiva de los modelos entrenados. Una vez entrenado, aplicado y evaluado el modelo, se procede a la selección de las variables más importantes.

Una vez entrenado y evaluado los modelos, se realizó una selección de las variables más importantes para las predicciones de línea y fenotipo. El proceso se llevó a cabo de forma distinta según el modelo. En PLS-DA y PLS se utilizó la importancia de las variables en la proyección (VIP) (Galindo-Prieto et al., 2014), seleccionando aquellas especies con un VIP > 1. En LASSO se seleccionaron las variables con un coeficiente de regresión distinto de 0 (Susin et al., 2020). En Random Forest y CatBoost se seleccionaron las especies con mayores valores de importancia de la variable dados por los propios modelos, seleccionando los mayores a 0.01 en el RandomForest y mayores a 0.1 en CatBoost. En Gaussian Naïve Bayes se utilizó la importancia

de permutación, realizando 20 permutaciones, seleccionando las variables con una importancia mayor a 0. Para valorar si la transformación por ALR afectaba a la selección e identificación de especies, se realizó un proceso de selección de variables con un modelo de clasificación y uno de regresión, que fueron el PLS-DA y el PLS, en el escenario intermedio NMH, con los datos con y sin transformar. Todos estos análisis se repitieron 10 veces cambiando la composición de los sets de testeo y entrenamiento para evitar sesgos en la predicción debido al set de entrenamiento y testeo utilizado. De cada iteración se obtuvo la media del número de especies totales seleccionadas y el número de especies seleccionadas con efecto sobre el fenotipo. En todos los análisis realizados, se determinó el número de falsos positivos identificados por cada modelo y para cada uno de los escenarios simulados.

Una vez comprobado el número de falsos positivos, se evaluaron diferentes aproximaciones para la selección de variables con el objetivo de reducir el número de falsos positivos identificados por los algoritmos ML. En todas estas aproximaciones se calculó el porcentaje de falsos positivos identificados. Se utilizaron tres aproximaciones:

i) Combinación de resultados entre modelos: se realizó una comparación por parejas de las especies seleccionadas entre modelos de clasificación, entre modelos de regresión y entre modelos de clasificación y regresión. Aquellas especies solapantes fueron consideradas las más importantes para la clasificación entre las líneas y/o predicción del fenotipo.
ii) Predicciones de TC dentro de cada línea divergente: se entrenaron tres modelos de regresiones para predecir el TC dentro de cada línea. Los modelos utilizados fueron PLS, LASSO y CatBoost, ya que presentaron los mejores resultados tras la combinación de metodologías. Las especies relevantes para la predicción del TC fueron aquellas especies coincidentes, identificadas por cada uno de los modelos ML para ambas líneas.
iii) Iteraciones alternando la composición de los sets de entrenamiento y testeo: con los modelos PLS, LASSO y CatBoost también se realizaron repeticiones de los análisis 100 veces alternando en cada iteración los individuos pertenecientes a los sets de entrenamiento y testeo. Se consideraron como variables más importantes aquellas que fueron seleccionadas en más de 95 iteraciones.

4. RESULTADOS

4.1. ANÁLISIS EXPLORATORIO DE LOS DATOS SIMULADOS

La evaluación de las tendencias fenotípicas, del efecto de la microbiota (mv) (Fig. 4) y del efecto del genoma (Fig. suplementaria 1) mostraron una clara separación de las líneas de alto y bajo tamaño de camada (TC) tras 13 generaciones de selección fenotípica (Fig. 4). El fenotipo de alto TC alcanzó una media de 9.92 ± 1.67 gazapos entre todos los escenarios, mientras que en la línea baja la media alcanzada fue de 7.62 ± 0.67. Las líneas seleccionadas para TC alto alcanzaron valores más alejados de la media fenotípica inicial que las líneas de TC bajo (Fig. 4A). Los valores fenotípicos más extremos alcanzados fueron de 16.07 ± 2.14 gazapos (8.07 por encima de la media) para la línea alta, y de 5.63 ± 1.92 gazapos en la línea baja (2.37 por debajo de la media). Ambos se registraron bajo condiciones de microbiabilidad de 0.5, pero con distintos valores de proporción de especies adquiridas del ambiente (Fig. 4A). La respuesta a la selección en las poblaciones simuladas bajo m2 de 0.15 fue menor, alcanzando valores máximos de 12.31 ± 2.39 gazapos de media y mínimos de 6.09 ± 2.26 gazapos. En los valores de mv, la población de alto TC alcanzó valores de 1.84 ± 0.74 y la de bajo TC de -0.38 ± 0.54 con m^2 de 0.15. Para valores de m2 de 0.5, la población de alto TC alcanzó valores de mv de 5.01 ± 1.43, mientras que la población de bajo TC de -0.73 ± 1.32 (Fig. 4B). La proporción de especies adquiridas por el ambiente (EM) también afectó a los valores máximos y mínimos alcanzados tanto para el fenotipo como para el mv. A mayores valores de EM, menor fue la respuesta fenotípica y los valores de mv fueron más similares entre escenarios, especialmente cuando la microbiabilidad era de 0.15. A nivel fenotípico, el escenario M mostró una respuesta a la selección menor en comparación a los escenarios NMH y HMH, que fueron más parecidos entre ellos (Fig. 4A). A nivel de mv, los tres escenarios mostraron tendencias más similares entre ellos. Las mayores respuestas a la selección se presentaron siempre en el escenario HMH, es decir, cuando el genotipo del propio individuo influía en la abundancia de un determinado número de especies (Fig. 4B).

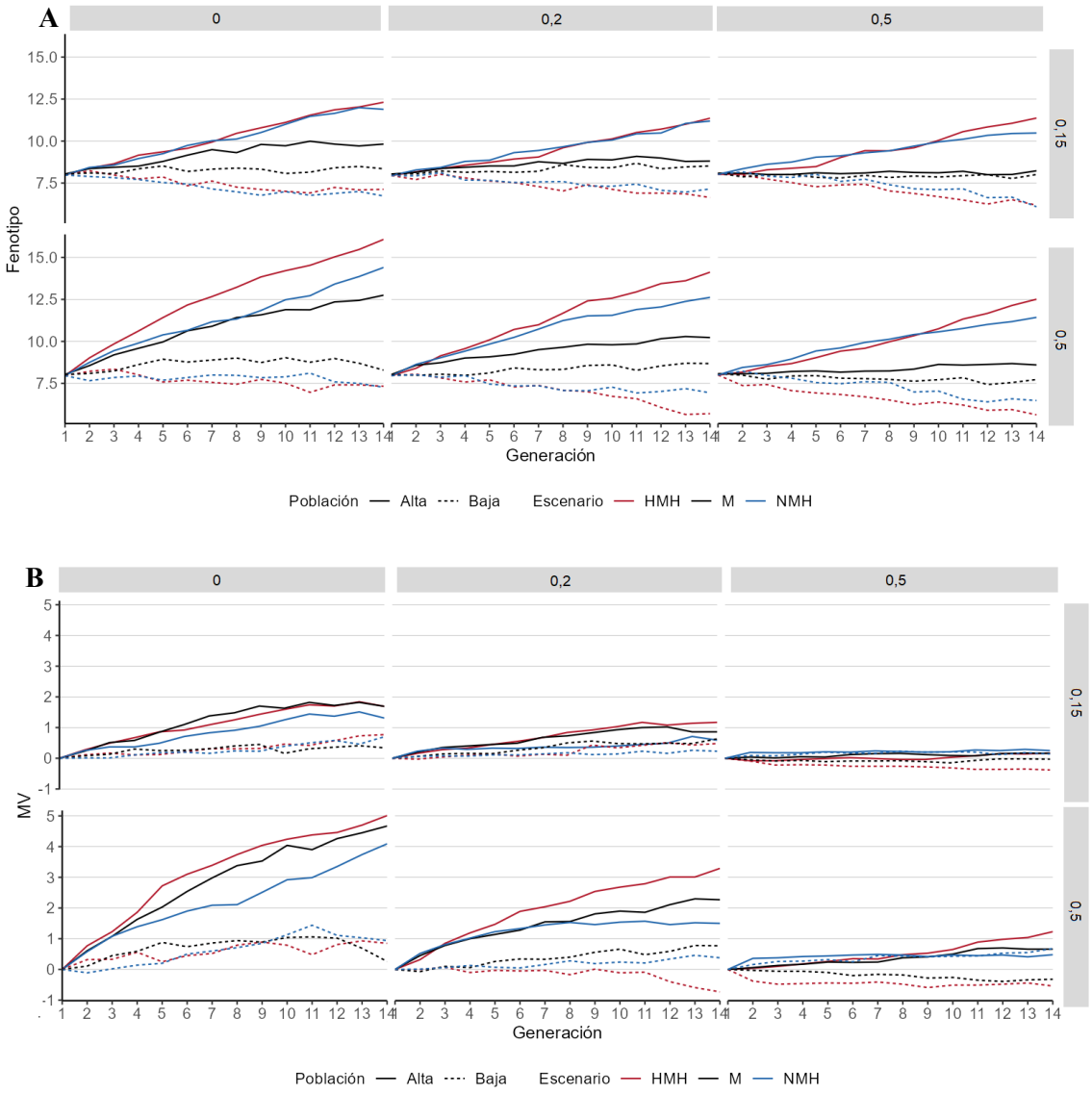

Figura 4. Evolución del fenotipo (A) y valor de la microbiota (B) de las poblaciones divergentes tras 13 generaciones de selección. Tendencias fenotípicas y mv de los escenarios M (negro), NMH (azul) y HMH (roja) de las poblaciones de alto (línea continua) y bajo (línea discontinua) tamaño de camada (TC) para valores de microbiabilidad de 0.15 y 0.5.

En la simulación se generaron un total de 1000 especies, de las cuales 600 se encontraban inicialmente en la microbiota de la población base de los individuos y en el ambiente. En la última generación, el máximo de especies presentes en la microbiota de la población fue de 791 para el escenario M y el mínimo fue de 540 para el escenario HMH y la línea de bajo TC (Tabla Suplementaria 1). Las diferencias en el total de especies presentes en las poblaciones dependieron principalmente de la proporción de especies adquiridas por el ambiente (Tabla suplementaria 1). La mayoría de las especies con efecto en el fenotipo se mantuvieron a lo largo de las generaciones, con un mínimo de 81 especies en la línea de alto TC para el escenario NMH y m^2 de 0.15. Por otra parte, se observó diferencias en la presencia/ausencia de determinadas especies dependiendo

de la línea divergente. El máximo de especies presentes sólo en la línea de alto TC fue de 47, mientras que para la baja fue de 89 especies (Tabla suplementaria 1). En todos los casos, la mayoría de las especies no tenían efecto en el fenotipo.

El análisis de componentes principales (PCA) usando la microbiota simulada tras 13 generaciones de selección divergente (Fig. 2) mostró el mismo patrón en todos los escenarios. Las poblaciones divergentes se separaron en dos grupos definidos (Fig. 2), aunque en algunos escenarios aparecieron subgrupos dentro línea. El porcentaje de varianza explicado por los dos primeros componentes y usando la matriz de microbiota completa fue en media de 5.88 ± 2.37%, siendo el máximo de 8.2% (Tabla suplementaria 2). Los PCA realizados únicamente con las especies con efecto sobre el fenotipo también separaron las dos líneas divergentes, y el porcentaje de varianza explicado fue en media de 8.67 ± 2.11% con un máximo del 12%. En estos PCA, los subgrupos presentes en el microbioma completo desaparecieron o se vieron reducidos.

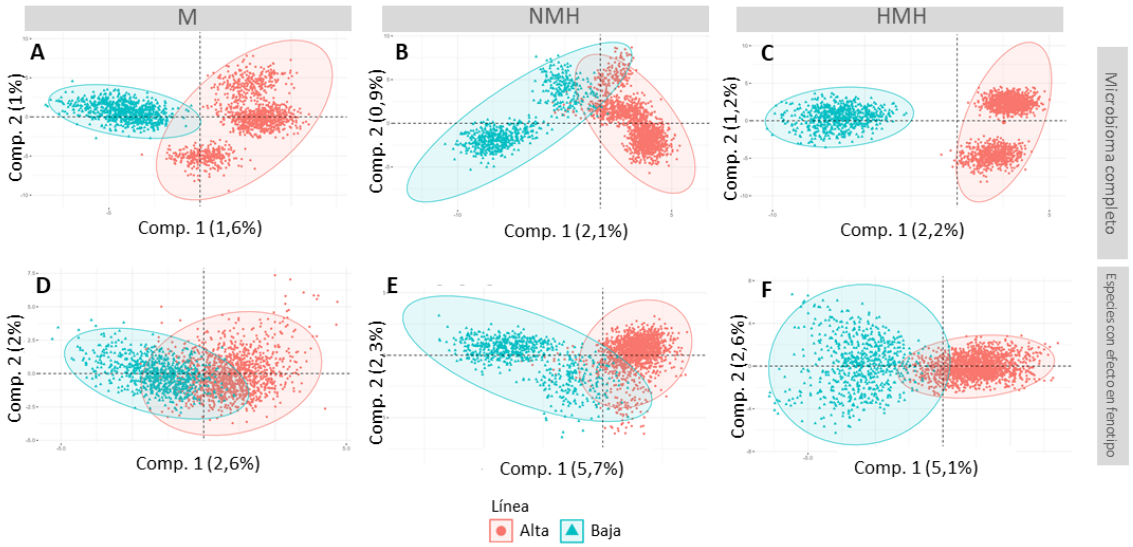

Figura 5. Análisis de componentes principales de las poblaciones divergentes tras 13 generaciones de selección. El análisis de componentes principales (PCA) se realizó utilizando la matriz de la microbiota completa (A, B, C) y únicamente la matriz de abundancias de las especies con efecto en el fenotipo (D, E, F). Los escenarios representados son el M (A, D), NMH (B, E) y HMH (C, F) con una microbiabilidad de 0.15 y un porcentaje de especies adquiridas por el ambiente del 50%. En rojo se destaca la población de alto tamaño de camada (TC), mientras que en azul se representa la población de bajo TC.

Se observaron correlaciones elevadas de 0.99 y -0.99 en las abundancias de las especies con y sin efecto sobre el fenotipo presentes en la población (Fig. 6). Estas correlaciones variaban según la proporción de especies adquiridas del ambiente. A valores más altos de EM, menor cantidad de especies con correlaciones altas (mayor de 0.4 o menor de -0.4, que fueron los valores donde se encontraron la mayoría de las especies) (Fig. 6). En los escenarios sin especies adquiridas del ambiente se obtuvo una media de 109 ± 23 especies con correlaciones altas (Fig.

6A, Fig. 6B, Fig. 6C), mientras que en los escenarios con EM de 0.2, la media fue de 77 ± 8 (Fig. 6D, Fig. 6E, Fig. 6F), y con EM de 0,5 fue de 35 ± 10 (Fig. 6G, Fig. 6H, Fig. 6I).

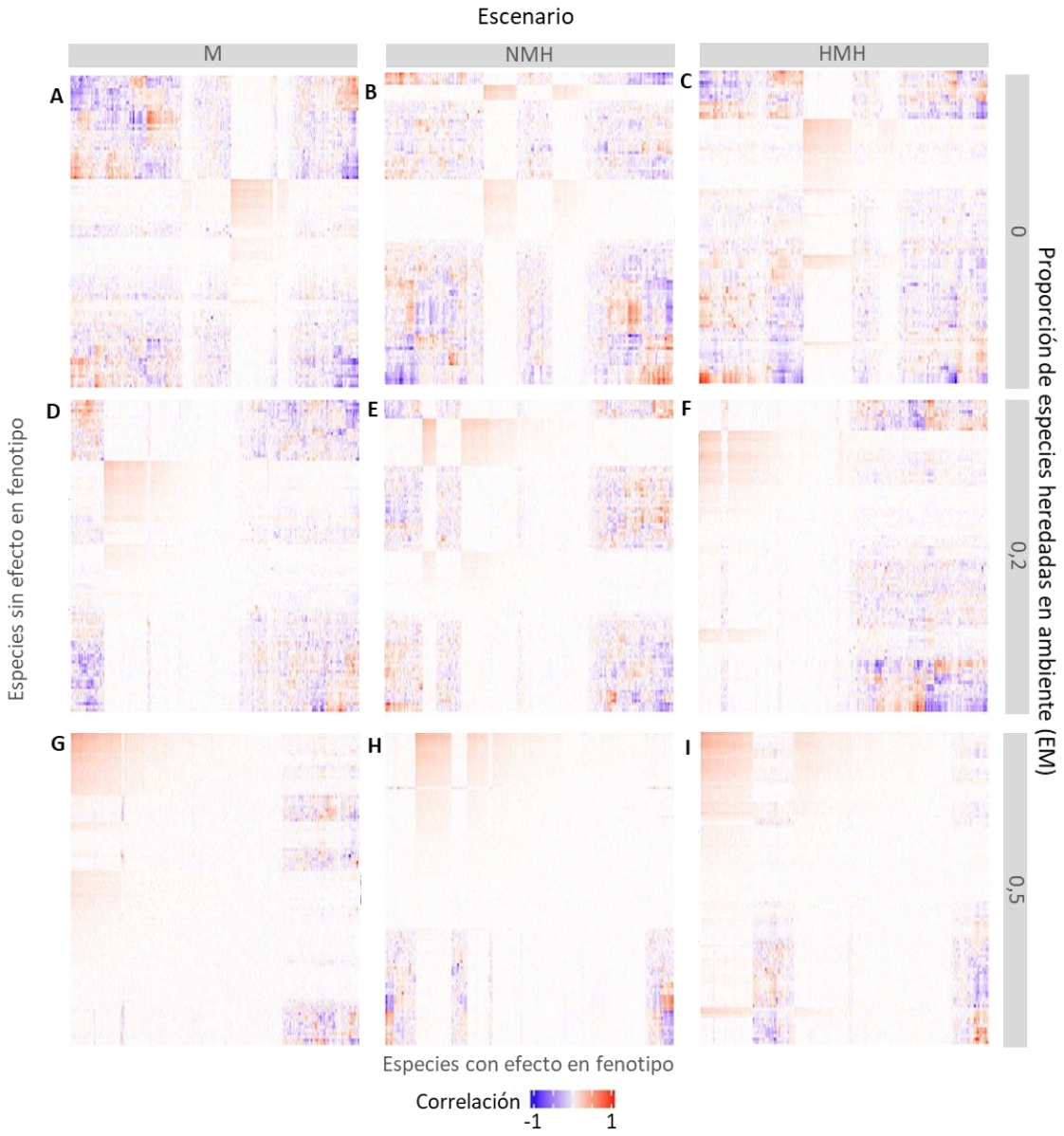

Figura 6. Heatmaps de correlaciones entre especies con y sin efecto en el fenotipo. Los escenarios representados fueron los escenarios con microbiabilidad de 0.15 en escenarios M (A, D, G), NMH (B, E, H) y HMH (C, F I) con EM de 0 (A, B, C), 0.2 (D, E, F) y 0.5 (G, H, I).

4.1.1. Efecto de la transformación por ALR

Una transformación ALR fue aplicada a la matriz de abundancias ya que es uno de los procedimientos utilizados para lidiar con la composicionalidad de este tipo de datos (Gloor et al., 2017a). No se observaron diferencias en la comparación entre el porcentaje de especies con efecto identificadas con y sin transformación ALR, utilizando tanto PLS como PLS-DA. En el caso específico de PLS-DA, se identificó que el 19% de las especies presentaban efecto con relación

al total de especies identificadas, independientemente de si se aplicó la transformación ALR o no. En el PLS, los porcentajes de especies con efecto sobre el total fueron de 19% con transformación y de 18% sin transformación. De las especies identificadas por el PLS-DA, el número de falsos y verdaderos positivos fue similar con y sin transformación ALR (Fig. 7). En cambio, en el PLS sí varió el número de especies con y sin efecto en el fenotipo, con un número total de especies identificadas de 115 ± 2 especies con ALR y 175±6 sin ALR. De todas las especies identificadas, se encontraron 22 ± 2 especies con efecto sobre el fenotipo utilizando ALR, y 32 ± 3 especies con los datos sin transformar (Fig. 7).

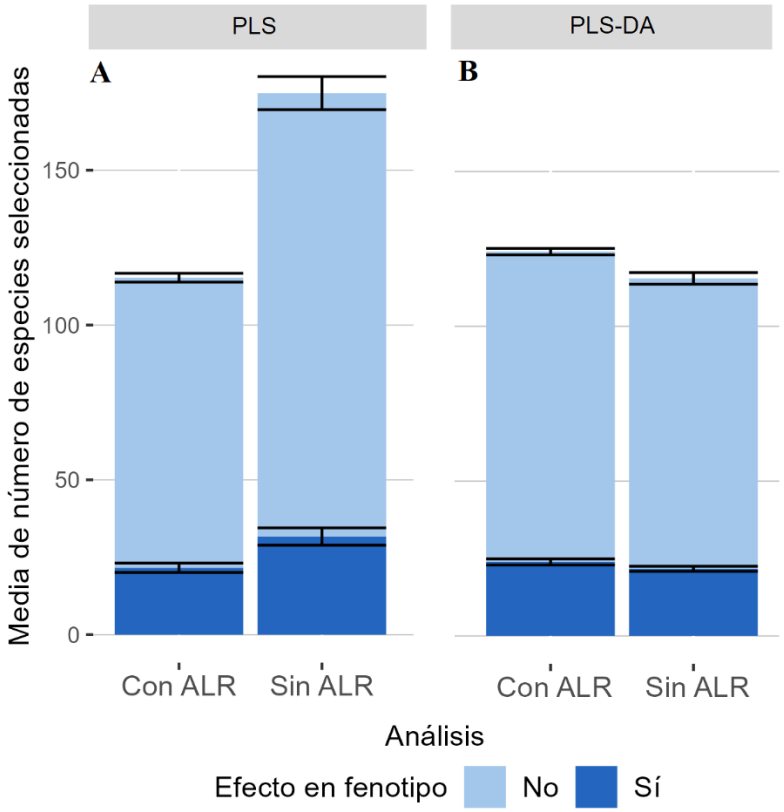

Figura 7. Comparación de media ± desviación típica de especies seleccionadas por los modelos PLS (A) y PLS-DA (B). Los gráficos muestran los resultados con y sin transformación por additive log-ratio (ALR) en el escenario NMH bajo parámetros de microbiabilidad de 0.15 y EM de 0.5.

4.2. MODELOS MACHINE LEARNING

Diferentes modelos de *machine learning* (ML) se implementaron para evaluar la capacidad de identificación de especies con efecto en el fenotipo mediante el desarrollo de modelos de clasificación de las poblaciones divergentes y de predicción del fenotipo. Para ello se utilizaron modelos ML de clasificación como *Partial Least Squares Discriminant Análisis* (PLS-DA), *Gaussian Naive Bayes* (GNB), *Random Forest* (RFC) y CatBoost (CBC) y de regresión como *Partial Least Squares* (PLS), *Least Absolute Shrinkage and Selection Operator* (LASSO), *Random Forest* (RFR) y CatBoost (CBR).

4.2.1. Modelos de clasificación

Todos los modelos de clasificación mostraron un rendimiento de clasificación del 100% independientemente de los parámetros y escenarios simulados. Los modelos identificaron una media de 19 ± 7% especies con efecto en el fenotipo sobre el total de especies identificadas por cada método (Fig. 8). Ninguno de los modelos identificó las 100 especies con efecto en el fenotipo. Los modelos PLS-DA, CatBoost y Random Forest identificaron una cantidad similar de especies con y sin efecto en el fenotipo entre todos los escenarios independientemente de los parámetros de m^2 y EM. El GNB fue el algoritmo que presentó una mayor variabilidad entre escenarios. Los modelos GNB ajustados para microbiomas simulados con valores de EM de 0.2 (Fig. 8B, 8E) y 0.5 (Fig. 8C, 8F) identificaron una gran cantidad de especies para los escenarios de NMH y M. Sin embargo, para EM de 0, el GNB identificó una proporción muy baja de especies, cercana a cero, en comparación con los demás escenarios. El modelo que seleccionó una mayor cantidad de especies con efecto en el fenotipo en todos los escenarios y parámetros de simulación fue el PLS-DA, con una media de 25 ± 4 especies con efecto en el fenotipo, pero con una media de 135 ± 18 falsos positivos. El modelo que menos especies identificó fue el CBC, con 1 ± 2 especies con efecto y 4 ± 1 especies sin efecto. Gran parte las especies seleccionadas entre las 10 primeras de mayor importancia en cada modelo fueron falsos positivos (Fig. 8). El modelo que más verdaderos positivos incluyó de media fue el RFC con una media de 2 ± 2 especies con efecto en el fenotipo entre todos los escenarios, y el que menos el CBC con una media de 1 ± 2 especies. El máximo de verdaderos positivos dentro del top 10 fueron identificados por el CatBoost en el escenario HMH con condiciones de m^2 de 0.5 y EM de 0.2, con 6 especies (Fig. Suplementaria 2). Sin embargo, este modelo también identificó el valor mínimo de especies con efecto en el top 10, y hubo escenarios y parámetros en los que no incluyó ninguna especie (Fig. Suplementaria 2).

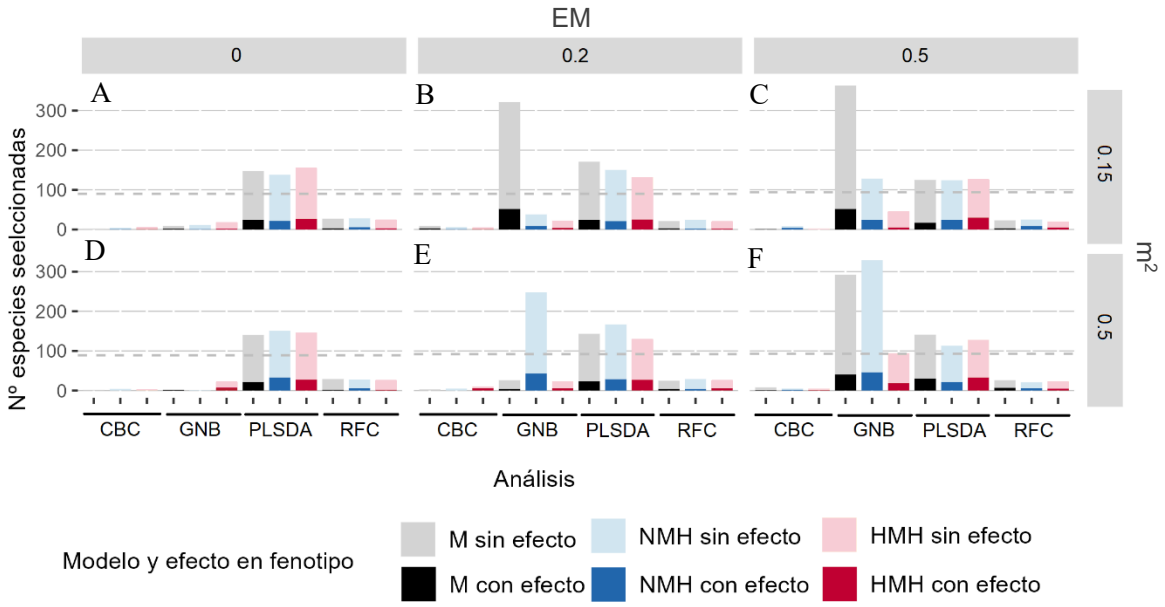

Figura 8. Número de especies seleccionadas por los modelos de clasificación. Número de especies seleccionadas en escenarios M (negro, gris), NMH (azul) y HMH (rojo, rosa) con efecto en el fenotipo (negro, azul oscuro, rojo) y sin efecto en el fenotipo (gris, azul claro, rosa) para valores de microbiabilidad de 0.15 (A, B y C) y 0.5 (D, E y F) y valores de ambiente de 0 (A y D), 0,2 (B y E) y 0,5 (C y F) para los modelos CatBoost de clasificación (CBC), *Gaussian Naive Bayes* (GNB), *Partial Least Squares Discriminant Analysis* (PLSDA) y *Random Forest* de clasificación (RFC)

4.2.2. Modelos de regresión

El rendimiento de los modelos de regresión se evaluó mediante el error cuadrático medio (RMSE) y la Q^2 (Tabla 3). Los modelos usando el microbioma simulado con una m^2 de 0.15 tuvieron errores más elevados que aquellos con una m^2 de 0.5. Independientemente del escenario, para m^2 de 0.15 los valores medios de RMSE fueron de 2.31 ± 0.1 y de 1.95 ± 0.2 para m^2 de 0.5. Respectivamente, los valores de Q^2 fueron de 0.31 ± 0.21 y de 0.56 ± 0.22 para m^2 de 0.15 y de 0.5. Se observaron diferencias en los valores de Q^2 dependiendo del escenario simulado. El escenario M obtuvo los peores ajustes, siendo su mejor resultado una media de 0.36 ± 0.22 obtenida con el CBR con microbiabilidad de 0.5. Los modelos NMH y HMH se parecieron más entre ellos, siendo el mejor ajuste el obtenido por el modelo CBR en el escenario HMH con m^2 de 0.5 con una media de 0.8 ± 0.04. En este escenario, el peor modelo fue el PLS, con una Q^2 de 0.71 ± 0.017.

Tabla 3. Errores cuadráticos medios (RMSE) y Q^2 en cada escenario (M, NMH y HMH), con parámetros de microbiabilidad (m^2) de 0,15 y 0,5.

Modelo	Análisis	RMSE		Q^2	
		m^2=0.15	m^2=0.5	m^2=0.15	m^2=0.5
M	PLS	2.45±0.33	2.19±0.57	-0.07±0.13	0.18±0.29
	LASSO	2.35±0.65	2.1±0.13	0.0002±0.05	0.23±0.25
	Random Forest	2.48±0.26	1.85±0.43	0.13±0.15	0.37±0.16
	CatBoost	2.17±0.5	1.78±0.47	0.04±0.08	0.36±0.22
NMH	PLS	2.46±0.45	2.06±0.57	0.4±0.08	0.61±0.08
	LASSO	2.39±0.57	1.97±0.65	0.41±0.06	0.64±0.11
	Random Forest	2.36±0.32	1.9±0.45	0.42±0.05	0.67±0.05
	CatBoost	2.24±0.54	1.8±0.16	0.45±0.04	0.66±0.09
HMH	PLS	2.41±0.45	2.2±0.59	0.46±0.04	0.71±0.017
	LASSO	2.39±0.63	1.91±0.4	0.47±0.02	0.75±0.05
	Random Forest	2.36±0.41	1.89±0.38	0.45±0.03	0.78±0.03
	CatBoost	2.2±0.73	1.78±0.27	0.5±0.02	0.8±0.04

Los modelos de regresión identificaron una media de un 25 ± 12% de especies con efecto en el fenotipo sobre el total de especies identificadas por cada método (Fig. 9). Al igual que en los modelos de clasificación, ningún modelo fue capaz de identificar las 100 especies presentes con efecto sobre el fenotipo. En todos los modelos, los escenarios NMH y HMH identificaron un número similar de especies entre ellos, mientras que el escenario M mostró más variabilidad en los modelos PLS y CBR, seleccionando un mayor número de falsos positivos al aumentar el efecto del ambiente (Fig. 9). El PLS fue el modelo que identificó de media un mayor número de especies en todos los escenarios, con una media de 33 ± 5 especies con efecto en el fenotipo. Sin embargo, también fue el modelo que más falsos positivos identificó (205 ± 31). Por otra parte, los modelos CBR y RFR identificaron la menor cantidad de especies con efecto en el fenotipo (Fig. 9), ambos con una media de 4 ± 2 especies, pero el RFR tuvo una mayor cantidad de falsos positivos con una media de 20 ± 3 frente a una media de 6 ± 4 falsos positivos en CBR. Los modelos de regresión incluyeron una mayor cantidad de verdaderos positivos entre las 10 primeras especies seleccionadas con respecto a los modelos de clasificación. El modelo con un mayor número de media fue el modelo LASSO, con 6 ± 2 especies, y el que menos fue el Random Forest con 2 ± 1 (Fig. suplementaria 3). El máximo de especies en el top 10 lo obtuvo el modelo LASSO, con 9 especies en todos los escenarios con los parámetros de m^2 de 0.5 y sin efecto del ambiente; y en

el escenario M con parámetros de m^2 de 0.5 y EM de 0.2. El mínimo de especies lo obtuvo el PLS y RFR, con 1 sola especie en varios escenarios y parámetros.

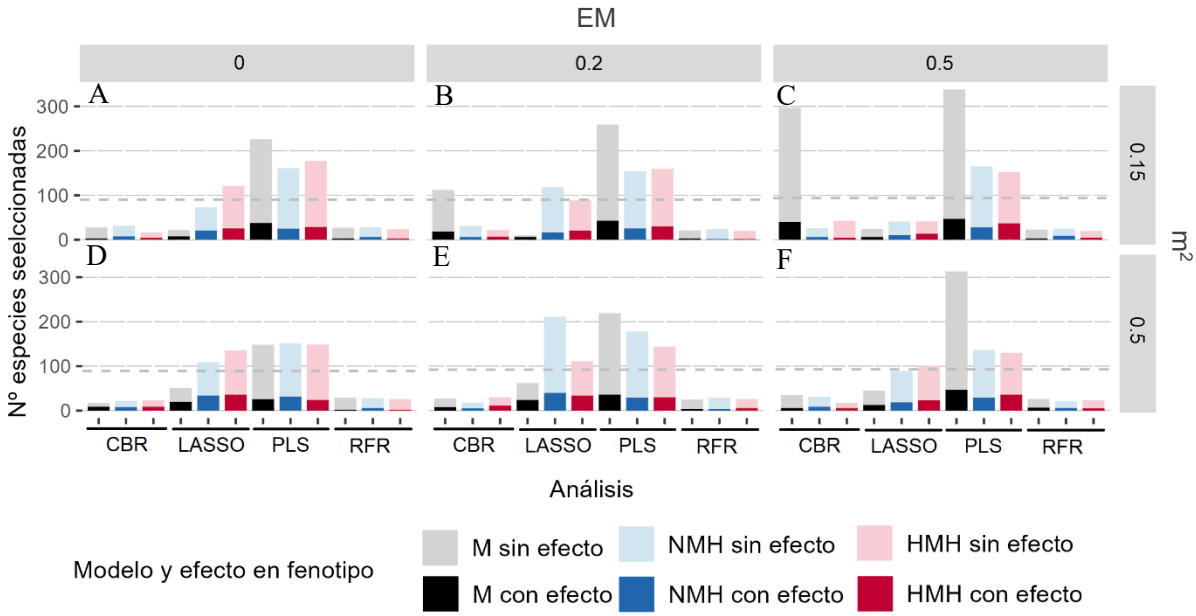

Figura 9. Número de especies seleccionadas por los modelos de regresión. Número de especies seleccionadas en escenarios M, NMH y HMH con efecto en el fenotipo (negro, azul oscuro y rojo) y sin efecto en el fenotipo (gris, azul claro, rosa) para valores de microbiabilidad de 0.15 (A, B y C) y 0.5 (D, E y F) y valores de ambiente de 0 (A y D), 0,2 (B y E) y 0,5 (C y F) para los modelos CatBoost de regresión (CBR), *Least Absolute Shrinkage Selection Operator* (LASSO), *Partial Least Squares* (PLS) y *Random Forest de regresión* (RFR).

4.2.3. Combinación de metodologías

Todos los modelos mostraron un elevado número de falsos positivos con respecto a los verdaderos positivos. Por tanto, para reducir el número de falsos positivos, se combinaron los resultados de los modelos por parejas. Las combinaciones se realizaron entre los modelos de clasificación, entre los modelos de regresión, y entre modelos de clasificación y de regresión. Debido a las pocas diferencias que se observaron entre los distintos parámetros de simulación, estos análisis se realizaron únicamente bajo las dos combinaciones de parámetros más extremas: las condiciones más favorables con m^2 de 0.5 y EM de 0; y en las condiciones más desfavorables con m^2 de 0.15 y EM de 0.5.

Al combinar las especies identificadas en modelos de clasificación, se observó que la media de falsos positivos se redujo a 14 ± 20, con un máximo de 119. Sin embargo, hubo un gran número de combinaciones en las que no coincidía ninguna especie entre los modelos (Figura Suplementaria 4). Esto sucedió en todos los escenarios, tanto en condiciones favorables como desfavorables, por lo que la eficiencia general de esta metodología fue baja. En los casos en los que sí coincidieron especies, el porcentaje de verdaderos positivos sobre el total aumentó

ligeramente con respecto a los modelos usados individualmente, con un valor de 21±16%. La combinación de modelos de regresión sí que mostró mejores resultados (Fig. 10). La media de falsos positivos fue de 10 ± 11, con un máximo de 71 especies sin efecto en el fenotipo; aunque la media de verdaderos positivos fue más baja, de 4 ± 4, con un máximo de 19. El porcentaje de verdaderos positivos sobre el total fue de 36 ± 24%, lo que implica que aumentó con respecto al análisis de los modelos de regresión individualmente. Todas las combinaciones realizadas tuvieron especies coincidentes, a excepción de RFR+CBR en los escenarios HMH y M usando las condiciones desfavorables. Hubo ocho combinaciones de modelos de regresión que mostraron un número de verdaderos positivos mayor al de falsos positivos. Las combinaciones de LASSO+RFR y PLS+RFR en las condiciones desfavorables y LASSO+CBR y PLS+CBR (Figura XX). Los modelos LASSO, CBR y PLS siempre tuvieron especies coincidentes entre ellos independientemente del escenario y parámetros simulados. También se observó una reducción del número de falsos positivos cuando se combinaban modelos de clasificación y de regresión. La media de falsos positivos fue de 14 ± 26, con un máximo de 125, pero el porcentaje de verdaderos positivos fue más bajo que cuando se combinaban modelos de regresión entre sí, ya que fue de 23 ± 20%. Debido al bajo rendimiento de los modelos de clasificación, hubo varias combinaciones en las que no coincidió ninguna especie (Fig. suplementaria 5). Bajo los parámetros favorables de m^2 de 0,5 y EM de 0, la combinación que mayor cantidad de especies con efecto sobre el fenotipo obtuvo manteniendo un número menor de falsos positivos fueron LASSO+CBR y PLS+CBR, con 6 verdaderos positivos en los escenarios HMH y NMH. Bajo parámetros más desfavorables se obtuvo un número menor, siendo el máximo 2, independientemente de la metodología implementada. Considerando los resultados obtenidos, los siguientes análisis se realizaron sobre los modelos PLS, LASSO y CBR, que podrían ofrecer los mejores resultados.

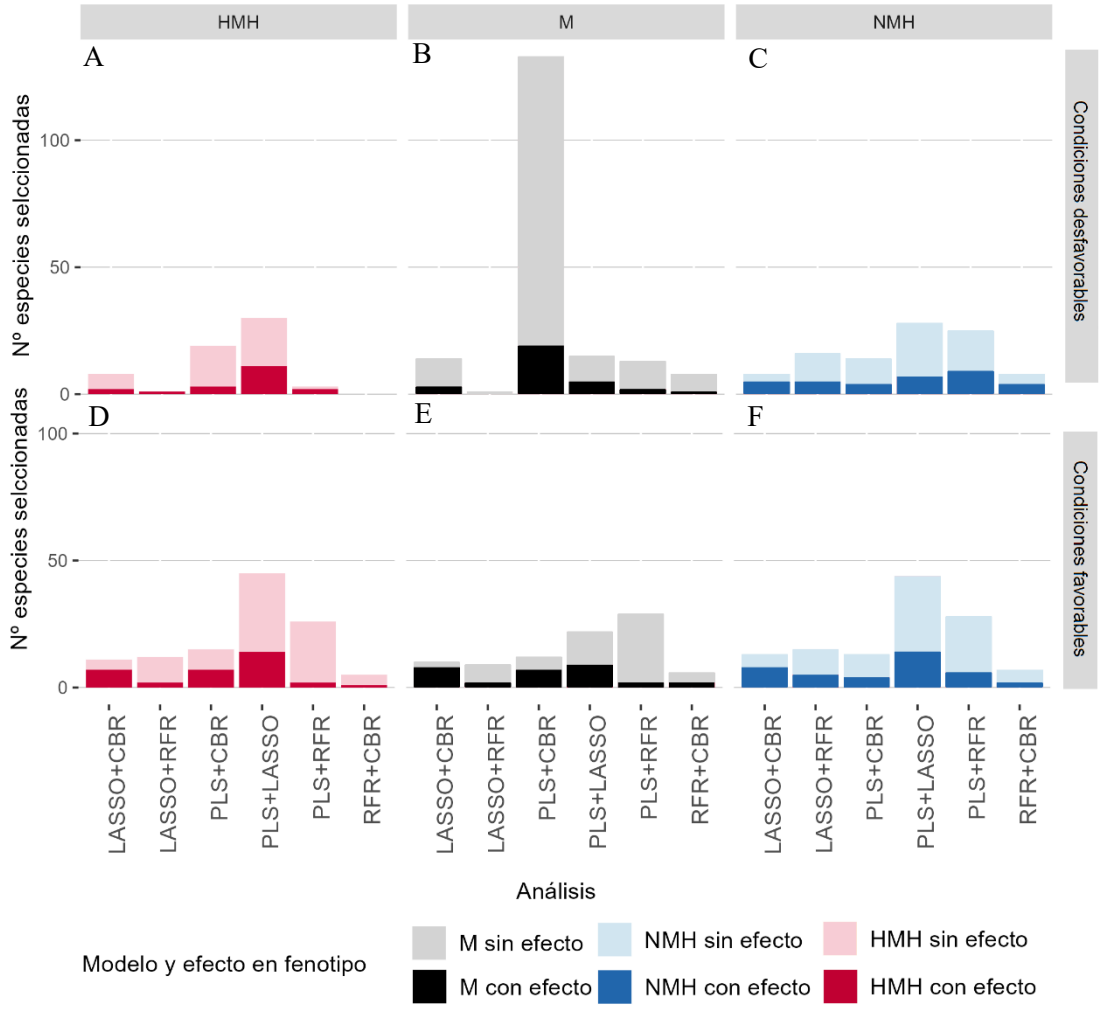

Figura 10. Número de especies coincidentes por combinación de modelos de regresión. En el eje X se muestran las combinaciones de modelos, y en el eje Y el número de especies seleccionadas. Las gráficas superiores muestran los escenarios M (negro, gris), NMH (azul) y HMH (rojo, rosa) realizados bajo condiciones desfavorables (m^2 de 0.15 y EM de 0.5; gráficas A, B y C), mientras que las gráficas inferiores muestran los escenarios realizados bajo condiciones favorables (m^2 de 0.5 y EM de 0; gráficas D, E y F). Especies con efecto (rojo, negro, azul oscuro) y sin efecto (gris, azul, rosa) en el fenotipo

4.2.4. Regresiones dentro de línea

La combinación de modelos de regresión redujo el número de falsos positivos identificados, sin reducir drásticamente el número de especies con efecto identificadas. Sin embargo, el porcentaje de falsos positivos siguió siendo elevado (66 ± 23%). En este caso, se realizaron modelos de predicción del fenotipo dentro de cada población divergente utilizando los tres modelos de regresión que mejores resultados obtuvieron en la combinación de metodologías: LASSO (Fig. suplementaria 6), PLS (Fig. suplementaria 7) y CBR (Fig. suplementaria 8). En general, las especies identificadas por los modelos dentro de cada línea no solaparon cuando el microbioma fue simulado bajo las peores condiciones de una m^2 de 0.15 y un EM de 0.5. Con los parámetros más favorables (m^2 de 0.5 y EM de 0), los modelos PLS y CBR identificaron especies

solapantes entre ambas líneas (Fig. 11). El número de especies solapantes fue bajo, pero el porcentaje de falsos positivos se redujo considerablemente, siendo de 0% para CBR (Fig. 11). El número de especies solapantes por CBR fue de 2 para el escenario M y de 3 para el HMH, no habiéndose identificado ninguna especie para el modelo NMH. El PLS sí que identificó un total de 6 especies solapantes en el escenario NMH (Fig. 11). Sin embargo 4 de ellas fueron falsos positivos. En el escenario M se identificó 5 especies, todas con efecto en el fenotipo, y en el HMH 9 especies de las cuales solo 3 fueron falsos positivos (Fig.11). Por otro lado, se evaluó si las especies solapantes entre las líneas ya habían sido identificadas previamente por los modelos de regresión realizados. En este caso, sólo el PLS para los escenarios M y HMH con los parámetros más favorables identificó 3 especies solapantes entre las líneas y el modelo de regresión previamente desarrollado (Fig. suplementaria 7).

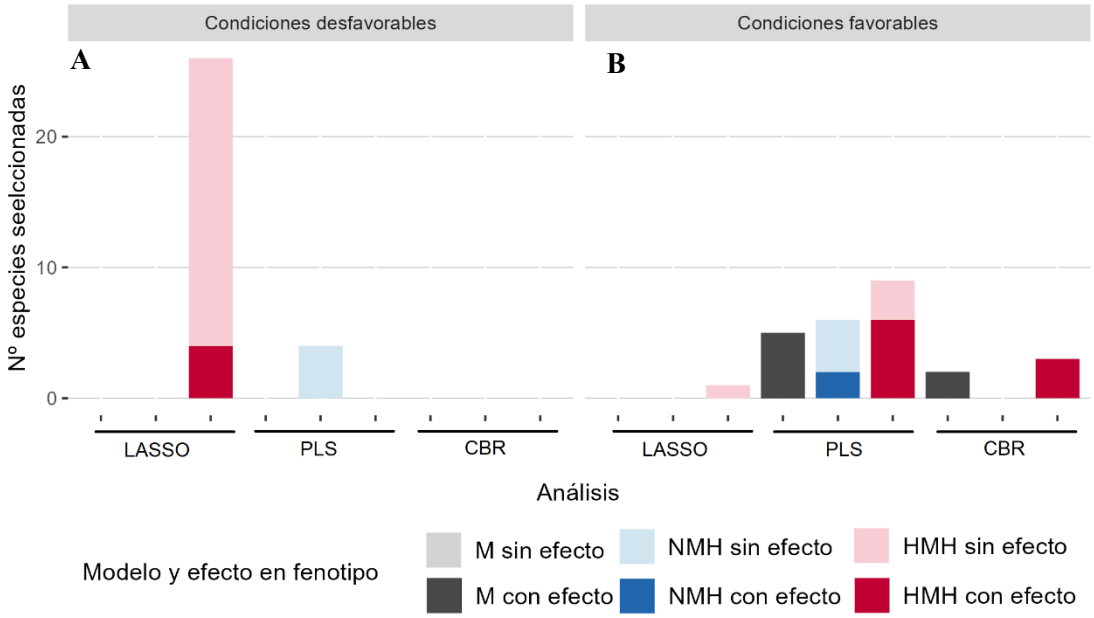

Figura 11. Especies coincidentes entre regresiones por línea bajo los parámetros favorables (A) (m^2 de 0.5 y EM de 0) y desfavorables (B) (m^2 de 0.15 y EM de 0.5). Especies con efecto (colores negro, azul oscuro y rojo) y sin efecto (colores gris, azul claro y rosa) en el fenotipo en escenarios M (negro/gris), NMH (azul) y HMH (rojo)

4.2.5. Modelos de regresión alternando sets de entrenamiento y testeo

Otra aproximación para reducir los falsos positivos fue evaluada. En este caso se utilizaron los modelos PLS, CBR y LASSO 100 veces, alternando los individuos que componen los sets de entrenamiento y testeo para comprobar si el sesgo generado por la división en estos grupos influía al número de falsos positivos, y evaluando las especies seleccionadas por cada modelo y las coincidentes entre ellos (Fig. 12). En general, los resultados fueron mejores bajo las condiciones favorables de microbiabilidad 0.5 y sin efecto del ambiente. Con estos parámetros, los modelos LASSO y CBR tuvieron pocos o ningún falso positivo en los tres escenarios (Fig.

12), mientras que el PLS tuvo resultados similares a los obtenidos en los análisis iniciales. Al combinar las metodologías, en todas las combinaciones coincidieron especies entre los modelos y el porcentaje de verdaderos positivos en estas combinaciones fue, de media, de 63 ± 20%. Sin embargo, en las condiciones más desfavorables los modelos LASSO y CBR identificaron pocas o ninguna especie con efecto en el fenotipo, y el PLS mantuvo resultados similares, destacando que en el escenario M solo el PLS identificó especies con efecto. Al combinar los modelos, en el escenario NMH las combinaciones con CBR o no coincidieron especies o solo coincidieron falsos positivos. En el escenario HMH sí coincidieron especies entre todos los métodos, pero el porcentaje de especies con efecto en el fenotipo fue de 25 ± 19%, mucho menor que en las condiciones favorables.

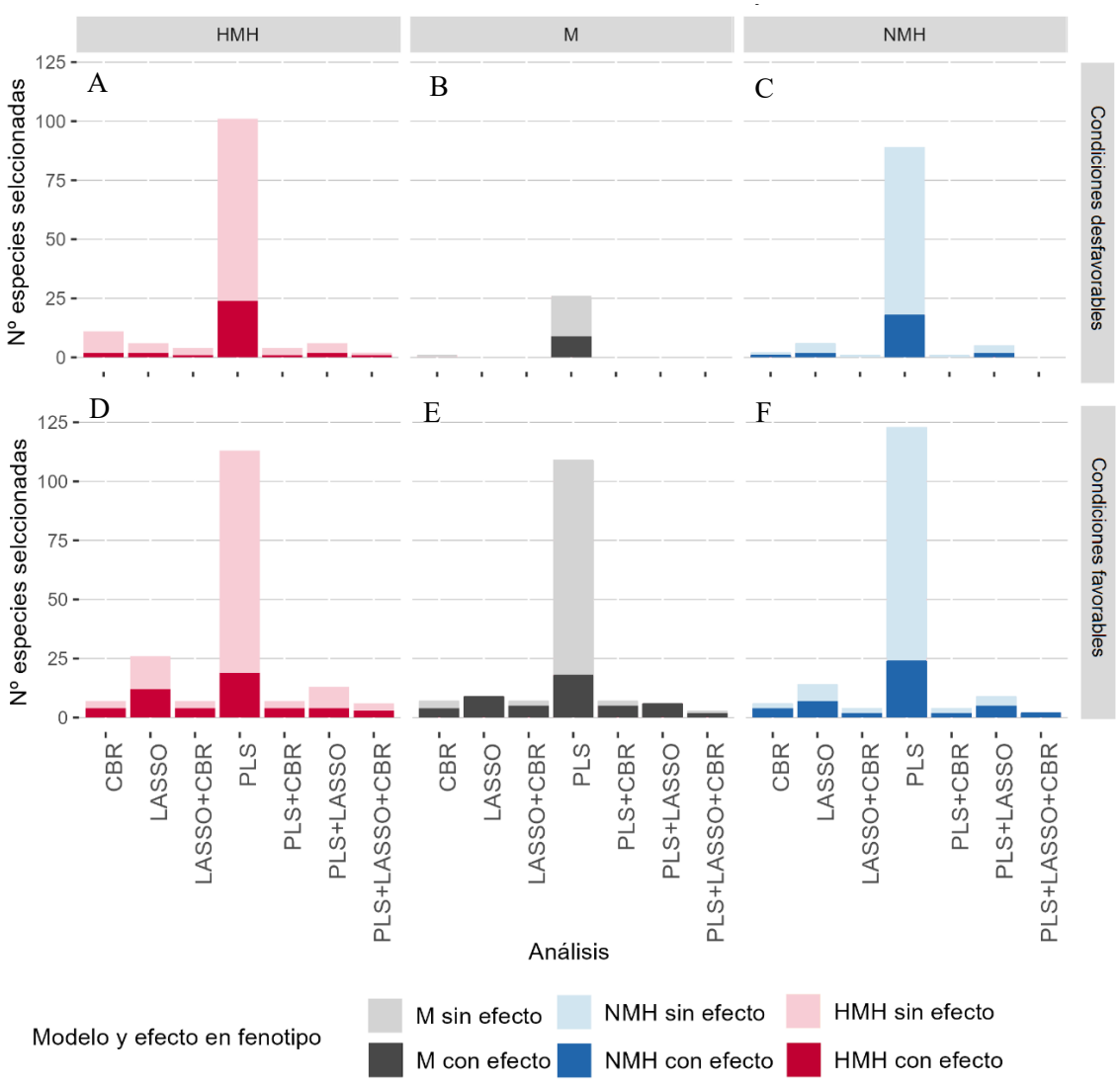

Figura 12. Especies seleccionadas alternando los sets de entrenamiento y testeo en los modelos CBR, PLS y LASSO. Se muestran las especies coincidentes entre ellos, en los escenarios HMH (A, D), M (B, E) y NMH (C, F) bajo parámetros favorables (m^2 de 0.5 y EM de 0) y desfavorables (m^2 de 0.15 y EM de 0.5). Especies con efecto (azul oscuro) y sin efecto (azul claro) en el fenotipo.

5. DISCUSIÓN

En este estudio se evaluó el rendimiento de diferentes modelos machine learning a la hora de identificar las especies bacterianas con efecto en el fenotipo. Para ello, se utilizó una microbiota simulada asumiendo una selección fenotípica divergente para alto y bajo TC durante 13 generaciones. La simulación de las dos poblaciones divergentes mostró que las dos líneas se diferenciaron entre ellas a nivel fenotípico (Fig. 5), genético (Fig. suplementaria 1) y de la microbiota (Fig. 5, Fig. 6). La respuesta fue mayor en la línea alta que en la línea baja (Fig. 5). De las 100 especies con efecto, muy pocas se perdieron durante el proceso de selección. Además, las dimensiones del microbioma se mantuvieron similares entre ambas líneas (Tabla suplementaria 1). Esta diferencia no fue debida a la ausencia de especies con efecto negativo sobre el fenotipo. En cambio, la mayor respuesta por parte de la línea alta se debió a que en las poblaciones base el efecto del microbioma fue mayoritariamente positivo, de modo que la tendencia inicial de la selección sería al alza, y la selección en la línea baja tendría mayores dificultades para reducir sus valores fenotípicos y de microbioma. Sin embargo, en el ambiente las especies con efecto negativo fueron más abundantes, por lo que en los escenarios con mayor efecto ambiental sí hubo una mayor respuesta en la línea baja (Fig. 5B). Por otra parte, las especies afectadas por el genoma con efecto positivo sobre el fenotipo fueron más abundantes en la población base, mientras que en el ambiente fueron más abundantes las que tenían efecto negativo. Por ello el escenario HMH sin ambiente obtuvo una mayor respuesta positiva, y una mayor respuesta negativa en la línea baja (Fig. 5B). Por otra parte, los cambios en los valores de la microbiabilidad (m^2) y la proporción de especies adquiridas por el ambiente tuvieron el efecto esperado sobre el fenotipo y la microbiota. A mayor microbiabilidad, mayor fue la respuesta de la microbiota. En cuanto al efecto ambiental, distintos estudios han mostrado que las crías comparten microbioma con sus madres (Maqsood et al., 2019b). Esta diferencia se incrementa cuanto menos contacto hay entre las crías y sus madres, ya sea en el nacimiento (Dominguez-Bello et al., 2010, Drell et al., 2017), o en la lactancia (Bäckhed et al., 2015, Drell et al., 2017). Por tanto, a mayor efecto ambiental menor es el parecido entre el microbioma materno y el de las crías, lo cual coincide con los resultados obtenidos en este estudio.

La separación entre las poblaciones también se pudo observar en el PCA (Fig. 2). En los PCA usando la microbiota completa se observaron subgrupos dentro de poblaciones. Sin embargo, estos subgrupos desaparecieron o se redujeron cuando el PCA se realizó sólo con las especies con efecto en el fenotipo (Fig. 6). Esto podría indicarnos un efecto deriva, que podría estar asociado a la selección de determinadas familias a lo largo de las generaciones, ya que esta variabilidad dentro de cada línea viene dada por especies que no influyen al carácter. Por otra parte, considerando las especies con efecto en el fenotipo (Fig. 5D, Fig. 5E, Fig. 5F), las poblaciones divergentes se separaban. Esto nos estaría indicando que la variabilidad observada

en el microbioma de los animales ha sido un efecto de la selección divergente aplicada (Fig. 5). Este efecto de la selección en el microbioma ya se había observado previamente en conejos seleccionados para alta y baja variabilidad del tamaño de camada (Casto-Rebollo et al., 2023c) y en conejos seleccionados para alta y baja grasa intramuscular (Zubiri-Gaitán et al., 2023). No obstante, estos estudios no tienen en cuenta la deriva asumiendo que todas las especies identificadas por los modelos son especies causales, importantes para la regulación del fenotipo de interés. Sin embargo, en este estudio observamos que una elevada proporción de las especies seleccionadas por los modelos ML de clasificación y de regresión fueron falsos positivos (Fig. 9 y Fig.10), a pesar del buen rendimiento de predicción observado en algunos de ellos. Todos los modelos utilizados han sido considerados como buenas opciones para el análisis de datos de microbioma (Topçuoğlu et al., 2020, Li et al., 2022, Moreno-Indias et al., 2021), pero estos estudios se han basado mayoritariamente en los rendimientos de predicción, y no en las variables que estaban siendo seleccionadas. Por ello, estudios de simulación como este son necesarios para evaluar la capacidad de identificación de las especies causales por los diferentes algoritmos de ML.

Todos los modelos de clasificación predijeron con un 100% de precisión la línea a la que pertenecía cada individuo. Además, estos modelos incluyeron muy pocas especies con efecto sobre el fenotipo entre las 10 más importantes de cada método (Fig. Suplementaria 2). Esto indica que el elevado número de falsos positivos obtenido se debe a que los modelos no han considerado las especies con efecto sobre el fenotipo como las más importantes. Estos modelos seleccionaron especies sin efecto sobre el fenotipo con grandes diferencias de abundancia entre líneas, generalmente debido a presencia/ausencia de especies en una población con respecto a la otra. La mayoría de estas especies no tenían efecto en el fenotipo y fueron consecuencia de la deriva. Al no necesitar las especies causales para realizar la clasificación, se explican las pocas diferencias en rendimiento de cada modelo, independientemente de los escenarios y parámetros simulados. Los modelos de regresión, por otra parte, sí tuvieron diferencias de rendimiento dependiendo del escenario y del parámetro de microbiabilidad. El escenario M, en el que se había visto una menor respuesta a la selección (Fig. 4, Fig. 5) tuvo peores ajustes, lo cual se vio reflejado en la selección de variables, ya que los modelos no han generalizado correctamente. Los modelos de regresión incluyeron de media un mayor número de especies con efecto en el fenotipo entre las 10 especies más importantes para la predicción de cada modelo (Fig. suplementaria 3), y el porcentaje de verdaderos positivos que identificaron fue mayor que en los modelos de clasificación. Por tanto, las especies con efecto sobre el fenotipo habrían tenido un mayor peso en las predicciones con modelos de regresión. Al utilizar directamente el fenotipo de los individuos, las especies diferentes entre líneas por presencia/ausencia, no son seleccionadas por los modelos de regresión

si esa presencia/ausencia no está relacionada con las diferencias fenotípicas, lo que explica su mejor rendimiento en comparación a los modelos de clasificación.

El elevado número de falsos positivos visto en todos los modelos puede tener diversas explicaciones. Un aspecto importante es que las abundancias de muchas especies del microbioma sin efecto sobre el fenotipo mostraron estar fuertemente correlacionadas con especies con efecto sobre el fenotipo (Fig. 6). Distintos tipos de modelos reaccionan a las variables correlacionadas de distintas maneras. La regresión LASSO, al ser un método de penalización tipo L1, al encontrar variables correlacionadas solo considera una de ellas como importante, asignando un coeficiente de regresión de 0 al resto (Namkung, 2020). El modelo CatBoost también elimina variables correlacionadas, dado que no incluye variables fuertemente correlacionadas entre ellas al generar los árboles. El resto de las modelos no eliminan las variables correlacionadas entre ellas, considerándolas todas importantes para la predicción. Dependiendo de si las variables correlacionadas se eliminan o no, puede suceder que o bien se pierda la especie con efecto en el fenotipo o que esa especie sea seleccionada junto con otros muchos falsos positivos. Esto se pudo observar especialmente en los modelos PLS y PLS-DA, que siempre seleccionaron una mayor cantidad de variables con respecto al resto de modelos (Fig. 8, Fig. 9) dado que no eliminan las correlaciones (Palermo, 2009). Esto ocasiona la selección de un elevado número de falsos positivos (Fig. 8, Fig. 9). Por otra parte, al tratarse de datos composicionales, una posibilidad inicial que explique los malos resultados puede venir de las dificultades de analizar este tipo de datos. En este estudio se realizó una transformación logarítmica por ALR, que previamente había sido considerada como una opción simple y efectiva para analizar datos composicionales (Greenacre et al., 2021). Se observó una diferencia en el número de falsos positivos identificados entre datos con y sin transformar (Fig. 7), pero el porcentaje de verdaderos positivos identificados por los modelos no se vio afectado. La transformación por ALR, por tanto, no afectó a la proporción de falsos positivos. Otra posible causa de los resultados observados son las fuerzas de deriva. A pesar de que las fuerzas de selección tienen un mayor efecto sobre el microbioma (Logares et al., 2018), la deriva tiene un fuerte efecto sobre caracteres que no tienen un gran impacto sobre el *fitness* (Mutumi et al., 2017), y causaría que unas especies, por azar, se fijen en una línea y no en la otra. El efecto de deriva se pudo observar en los PCA (Fig. 5) y afectaría particularmente a los modelos de clasificación, que no distinguen las especies identificadas por deriva de las causales. Otros aspectos que podrían estar afectando a los resultados estarían relacionados con parámetros propios de la simulación de los datos, como las abundancias iniciales de las especies, su variabilidad, la magnitud del efecto de las especies sobre el microbioma, o si están influenciadas por el genoma o no. Todos estos parámetros iniciales pueden beneficiar a unas especies por encima de otras y causar que por azar los modelos las seleccionen o no. Por ejemplo, si una especie con un efecto grande sobre el fenotipo tiene una abundancia inicial en la población

base muy pequeña, es posible que esa especie se pierda en alguna generación y no vuelva a incorporarse nunca o se incorpore a través del ambiente en una generación muy avanzada; de forma que su efecto se pierde en la población. Sin embargo, para determinar los motivos por los que unas especies se seleccionan y otras no sería necesario realizar otros estudios analizando los microbiomas de cada línea en cada generación. Por otra parte, también hay que considerar el rendimiento de los modelos. Los modelos con bajos rendimientos de predicción no generalizan bien, por lo que la proporción de falsos positivos aumenta al no aprender correctamente los patrones para predecir el fenotipo. Esto se observa en los modelos de regresión, que tuvo mejores rendimientos en los modelos ajustados con la microbiota simulada bajo condiciones favorables (Tabla 3). También obtuvieron una menor cantidad de falsos positivos bajo los parámetros más favorables (Fig. 10, Fig. 11, Fig. 12).

En este estudio se consideraron tres aproximaciones para intentar mitigar el número de falsos positivos identificados por los modelos. El primero fue la combinación por pares de los resultados obtenidos por cada modelo, ya fueran de regresión o de clasificación. Esta aproximación se basa en la idea de que cada metodología es capaz de aprender diferentes patrones para resolver un mismo problema, por lo que el uso de varios modelos es recomendable (Moreno-Indias et al., 2021,Papoutsoglou et al., 2023) . En este estudio se han obtenido resultados que respaldan el uso de la combinación de metodologías, dado que la combinación de modelos de regresión redujo el número de falsos positivos del 75 ± 12% inicial a un 66 ± 22%, lo que supone una diferencia de 34 ± 8 falsos positivos (Fig. 10). Por otra parte, se realizó una aproximación alternando las composiciones de los grupos de entrenamiento y testeo. Para ello se realizó 100 grupos de set de entrenamiento y test, con el objetivo de reducir la identificación de especies debido a la subdivisión de set de entrenamiento y test realizada. Los individuos que conforman el grupo de entrenamiento tienen una mayor influencia sobre las predicciones, ya que se usan para generar el modelo. En bases de datos donde el número de individuos es pequeño, este efecto puede ser determinante a la hora de la selección de variables. Esta aproximación funcionó bajo las condiciones favorables de microbiabilidad y ambiente, pero en las condiciones desfavorables presentó peores resultados, pues mantuvo un porcentaje de falsos positivos similar al de los análisis iniciales con un 75 ± 19%. Destacó especialmente el escenario M (Fig. 13B), en el que solo el PLS identificó especies causales. El mal rendimiento de este escenario pudo deberse a su baja respuesta de selección (Fig. 4) y a los bajos rendimientos de predicción que obtuvo (Tabla 3), lo que impide que los modelos generalicen y predigan correctamente. En el resto de los escenarios, la composición de los grupos de entrenamiento y testeo mostraron tener efecto sobre los resultados, principalmente en los modelos LASSO y CBR, pero solo tuvieron buenos resultados en las condiciones favorables donde se obtuvieron los mejores rendimientos de predicción, con tan solo un 37 ± 20% de falsos positivos. Finalmente, se realizaron modelos de

regresión dentro de cada línea divergente para intentar reducir efectos de deriva debido a la selección divergente de las dos líneas. Esta aproximación fue la que mejor rendimiento mostró, ya que las especies que coincidían entre las seleccionadas de cada población eran todas o casi todas verdaderos positivos para los modelos PLS y CBR, en los escenarios donde hubo especies coincidentes (Fig. 11). Sin embargo, hubo ocasiones en las que los modelos no pudieron identificar especies o identificaron muy pocas, algo que sucedió principalmente en el modelo LASSO y en la línea baja. Esto sucedió debido a que la línea baja mostró una menor respuesta a la selección, y el modelo LASSO no pudo encontrar ninguna relación entre el microbioma y el fenotipo de los individuos dentro de esta línea. Debido a estos resultados, esta aproximación no sería indicada en el modelo LASSO. En PLS y CBR, el número de especies identificadas al coincidir las seleccionadas en la línea alta y en la línea baja fue reducido (máximo de 9 en condiciones favorables para el modelo PLS) pero con un número de falsos positivos casi nulo (máximo de 4 en PLS). También cabe destacar que con esta aproximación los verdaderos positivos seleccionados generalmente no habían sido previamente identificados en los modelos ML de regresión. Por ello, una buena aproximación sería complementar los análisis ML de regresión globales con análisis dentro de población. Además, estos resultados pueden combinarse con el top 10 de especies identificadas en los análisis de regresión iniciales (Fig. suplementaria 3), especialmente con el top 10 del modelo LASSO que fue el que más especies con efecto incluyó. De esta forma, también pueden reducirse los falsos positivos. A pesar de que todas estas metodologías redujeron el número de falsos positivos, en los análisis que identificaron más especies con efecto en el fenotipo que falsos positivos el número de especies identificadas fue muy reducido, de 5 especies como máximo. Por tanto, si se quieren obtener resultados con el menor número de falsos positivos posible, se debe asumir que una gran cantidad de especies que serían biológicamente relevantes se van a perder en los análisis. Aun así, los falsos positivos tienen un papel importante en estos análisis y se deben tener en cuenta a la hora de interpretar biológicamente los resultados.

6. CONCLUSIÓN

En este estudio se han evaluado diversos métodos para identificar las especies con efecto sobre el fenotipo de datos de la microbiota de dos poblaciones divergentes seleccionadas para tamaño de camada alto y bajo, obtenidas mediante una herramienta que permite simular la evolución del fenotipo, genoma y microbiota a través de 13 generaciones. Se han evaluado la cantidad de verdaderos y falsos positivos seleccionados por modelos de clasificación y de regresión, y se han realizado análisis de los métodos combinados y regresiones dentro de cada línea divergente y en iteraciones cambiando los sets de entrenamiento y testeo. De todos los análisis realizados, se observó que los modelos *machine learning* tienen buenos rendimientos de predicción, pero identifican un elevado número de falsos positivos. Dichos falsos positivos pudieron aparecer por diversas causas como correlaciones entre especies, fuerzas de deriva o efectos propios de la simulación. Los modelos de regresión mostraron un mejor rendimiento seleccionando un menor número de falsos positivos y dando una mayor importancia a las especies con efecto en el fenotipo, y de todos los análisis realizados con los modelos de regresión los que menor cantidad falsos positivos mostraron fueron las combinaciones de los modelos PLS, LASSO y CatBoost de regresión, pudiendo complementarse con el top 10 de especies identificadas por el modelo LASSO y con regresiones dentro de línea en modelos CBR y PLS, bajo las condiciones favorables en las que los modelos tuvieron buenos rendimientos de predicción.

En conclusión, los modelos *machine learning* para identificación de especies causales seleccionan un elevado número de falsos positivos y las mejores estrategias encontradas para reducirlos fue seleccionar las especies coincidentes entre los modelos de regresión LASSO, PLS y CBR, con y sin iteraciones alternando los grupos de entrenamiento y testeo, complementando estos resultados con regresiones dentro de línea y con el top 10 de especies identificadas. Aun así, a pesar de que los modelos muestren buenos rendimientos, se debe tener cuidado al obtener conclusiones biológicas sobre las especies identificadas mediante estas metodologías.

7. REFERENCIAS

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *44*(2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Akagawa, S., Tsuji, S., Onuma, C., Akagawa, Y., Yamaguchi, T., Yamagishi, M., Yamanouchi, S., Kimata, T., Sekiya, S., Ohashi, A., Hashiyada, M., Akane, A., & Kaneko, K. (2019). Effect of Delivery Mode and Nutrition on Gut Microbiota in Neonates. *Annals of Nutrition and Metabolism*, *74*(2), 132–139. https://doi.org/10.1159/000496427

Aliakbari, A., Zemb, O., Cauquil, L., Barilly, C., Billon, Y., & Gilbert, H. (2022). Microbiability and microbiome-wide association analyses of feed efficiency and performance traits in pigs. *Genetics Selection Evolution*, *54*(1), 29. https://doi.org/10.1186/s12711-022-00717-7

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., … Wang, J. (2015a). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*, *17*(5), 690–703. https://doi.org/10.1016/j.chom.2015.04.004

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., … Wang, J. (2015b). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*, *17*(5), 690–703. https://doi.org/10.1016/j.chom.2015.04.004

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bergamaschi, M., Tiezzi, F., Howard, J., Huang, Y. J., Gray, K. A., Schillebeeckx, C., McNulty, N. P., & Maltecca, C. (2020). Gut microbiome composition differences among breeds impact feed efficiency in swine. *Microbiome*, *8*(1), 110. https://doi.org/10.1186/s40168-020-00888-9

Blasco, A., Martínez-Álvaro, M., García, M.-L., Ibáñez-Escriche, N., & Argente, M.-J. (2017). Selection for environmental variance of litter size in rabbits. *Genetics Selection Evolution*, *49*(1), 48. https://doi.org/10.1186/s12711-017-0323-4

Boggio, G. M., Christensen, O. F., Legarra, A., Meynadier, A., & Marie-Etancelin, C. (2023). Microbiability of milk composition and genetic control of microbiota effects in sheep. *Journal of Dairy Science*, *106*(9), 6288–6298. https://doi.org/10.3168/jds.2022-22948

Breiman, L. (2001). *Random Forests* (Vol. 45).

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1136. https://doi.org/10.1093/bib/bbx120

Buitenhuis, B., Lassen, J., Noel, S. J., Plichta, D. R., Sørensen, P., Difford, G. F., & Poulsen, N. A. (2019). Impact of the rumen microbiome on milk fatty acid composition of Holstein cattle. *Genetics Selection Evolution*, *51*(1), 23. https://doi.org/10.1186/s12711-019-0464-8

Busato, S., Gordon, M., Chaudhari, M., Jensen, I., Akyol, T., Andersen, S., & Williams, C. (2023). Compositionality, sparsity, spurious heterogeneity, and other data-driven challenges for machine learning algorithms within plant microbiome studies. *Current Opinion in Plant Biology*, *71*, 102326. https://doi.org/10.1016/j.pbi.2022.102326

Cammack, K. M., Austin, K. J., Lamberson, W. R., Conant, G. C., & Cunningham, H. C. (2018). Tiny but mighty: The role of the rumen microbes in livestock production. *Journal of Animal Science*. https://doi.org/10.1093/jas/skx053

Carneiro, M., Albert, F. W., Afonso, S., Pereira, R. J., Burbano, H., Campos, R., Melo-Ferreira, J., Blanco-Aguiar, J. A., Villafuerte, R., Nachman, M. W., Good, J. M., & Ferrand, N. (2014). The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit. *PLoS Genetics*, *10*(8), e1003519. https://doi.org/10.1371/journal.pgen.1003519

Caruso, V., Song, X., Asquith, M., & Karstens, L. (2019). Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *MSystems*, *4*(1). https://doi.org/10.1128/mSystems.00163-18

Casto-Rebollo, C., Argente, M. J., García, M. L., Pena, R. N., Blasco, A., & Ibáñez-Escriche, N. (2023a). Selection for environmental variance shifted the gut microbiome composition driving animal resilience. *Microbiome*, *11*(1). https://doi.org/10.1186/s40168-023-01580-4

Casto-Rebollo, C., Pocrnic, I., Gorjanc, G., & Ibáñez-Escriche, N. (2022). 503. Simulation of host-microbiome evolution throughout a divergent selection experiment. *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*, 2089–2092. https://doi.org/10.3920/978-90-8686-940-4_503

Cerutti, A., Chen, K., & Chorny, A. (2011). Immunoglobulin Responses at the Mucosal Interface. *Annual Review of Immunology*, *29*(1), 273–293. https://doi.org/10.1146/annurev-immunol-031210-101317

Clavijo, V., & Flórez, M. J. V. (2018). The gastrointestinal microbiome and its association with the control of pathogens in broiler chicken production: A review. *Poultry Science*, *97*(3), 1006–1021. https://doi.org/10.3382/ps/pex359

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., Wright, A.-D. G., Zhu, Z., Kristensen, L., Nielsen, H. B., Guldbrandtsen, B., & Sahana, G. (2018). Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLOS Genetics*, *14*(10), e1007580. https://doi.org/10.1371/journal.pgen.1007580

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2/3), 103–130. https://doi.org/10.1023/A:1007413511361

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, *107*(26), 11971–11975. https://doi.org/10.1073/pnas.1002601107

Drell, T., Štšepetova, J., Simm, J., Rull, K., Aleksejeva, A., Antson, A., Tillmann, V., Metsis, M., Sepp, E., Salumets, A., & Mändar, R. (2017). The Influence of Different Maternal Microbial

Communities on the Development of Infant Gut and Oral Microbiota. *Scientific Reports*, *7*(1), 9940. https://doi.org/10.1038/s41598-017-09278-y

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 15. https://doi.org/10.1186/2049-2618-2-15

Feurer, M., & Hutter, F. (2019). *Hyperparameter Optimization* (pp. 3–33). https://doi.org/10.1007/978-3-030-05318-5_1

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Funkhouser, L. J., & Bordenstein, S. R. (2013). Mom Knows Best: The Universality of Maternal Microbial Transmission. *PLoS Biology*, *11*(8), e1001631. https://doi.org/10.1371/journal.pbio.1001631

Galindo-Prieto, B., Eriksson, L., & Trygg, J. (2014). Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *Journal of Chemometrics*, *28*(8), 623–632. https://doi.org/10.1002/cem.2627

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. *G3 Genes|Genomes|Genetics*, *11*(2). https://doi.org/10.1093/g3journal/jkaa017

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gilbert, S. F. (2014). A holobiont birth narrative: the epigenetic transmission of the human microbiome. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00282

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017a). Microbiome datasets are compositional: And this is not optional. In *Frontiers in Microbiology* (Vol. 8, Issue NOV). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2017.02224

Greenacre, M. (2018). *Compositional Data Analysis in Practice*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429455537

Greenacre, M., Martínez-Álvaro, M., & Blasco, A. (2021). Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.727398

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nature Reviews Genetics*, *18*(11), 690–699. https://doi.org/10.1038/nrg.2017.63

Hoffmann, A. R., Proctor, L. M., Surette, M. G., & Suchodolski, J. S. (2016). The Microbiome. *Veterinary Pathology*, *53*(1), 10–21. https://doi.org/10.1177/0300985815595517

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., & Jiang, Y. (2019). Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00995

Kassambara A, Mundt F (2020). _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7, <https://CRAN.R-project.org/package=factoextra>

Kaul, A., Davidov, O., & Peddada, S. D. (2017). Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, kxw053. https://doi.org/10.1093/biostatistics/kxw053

Kaur, H., Kaur, G., Gupta, T., Mittal, D., & Ali, S. A. (2023). Integrating Omics Technologies for a Comprehensive Understanding of the Microbiome and Its Impact on Cattle Production. *Biology*, *12*(9), 1200. https://doi.org/10.3390/biology12091200

Kern, L., Abdeen, S. K., Kolodziejczyk, A. A., & Elinav, E. (2021). Commensal inter-bacterial interactions shaping the microbiota. *Current Opinion in Microbiology*, *63*, 158–171. https://doi.org/10.1016/j.mib.2021.07.011

Khanal, P., Maltecca, C., Schwab, C., Fix, J., & Tiezzi, F. (2021). Microbiability of meat quality and carcass composition traits in swine. *Journal of Animal Breeding and Genetics*, *138*(2), 223–236. https://doi.org/10.1111/jbg.12504

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1475), 1929–1940. https://doi.org/10.1098/rstb.2006.1920

Lapidus, A. L., & Korobeynikov, A. I. (2021). Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.613791

Lemos, L. N., Fulthorpe, R. R., Triplett, E. W., & Roesch, L. F. W. (2011). Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of Microbiological Methods*, *86*(1), 42–51. https://doi.org/10.1016/j.mimet.2011.03.014

Li, P., Luo, H., Ji, B., & Nielsen, J. (2022). Machine learning for data integration in human gut microbiome. *Microbial Cell Factories*, *21*(1), 241. https://doi.org/10.1186/s12934-022-01973-4

Liu, J., Stewart, S. N., Robinson, K., Yang, Q., Lyu, W., Whitmore, M. A., & Zhang, G. (2021). Linkage between the intestinal microbiota and residual feed intake in broiler chickens. *Journal of Animal Science and Biotechnology*, *12*(1), 22. https://doi.org/10.1186/s40104-020-00542-2

Logares, R., Tesson, S. V. M., Canbäck, B., Pontarp, M., Hedlund, K., & Rengefors, K. (2018). Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. *Environmental Microbiology*, *20*(6), 2231–2240. https://doi.org/10.1111/1462-2920.14265

Lubbe, S., Filzmoser, P., & Templ, M. (2021). Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, *210*. https://doi.org/10.1016/j.chemolab.2021.104248

Lutz, K. C., Jiang, S., Neugent, M. L., De Nisco, N. J., Zhan, X., & Li, Q. (2022). A Survey of Statistical Methods for Microbiome Data Analysis. *Frontiers in Applied Mathematics and Statistics*, *8*. https://doi.org/10.3389/fams.2022.884810

Maltecca, C., Dunn, R., He, Y., McNulty, N. P., Schillebeeckx, C., Schwab, C., Shull, C., Fix, J., & Tiezzi, F. (2021). Microbial composition differs between production systems and is associated with growth performance and carcass quality in pigs. *Animal Microbiome*, *3*(1), 57. https://doi.org/10.1186/s42523-021-00118-z

Maqsood, R., Rodgers, R., Rodriguez, C., Handley, S. A., Ndao, I. M., Tarr, P. I., Warner, B. B., Lim, E. S., & Holtz, L. R. (2019a). Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome*, *7*(1), 156. https://doi.org/10.1186/s40168-019-0766-7

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.634511

Martinez Boggio, G., Monteiro, H. F., Lima, F. S., Figueiredo, C. C., Bisinotto, R. S., Santos, J. E. P., Mion, B., Schenkel, F. S., Ribeiro, E. S., Weigel, K. A., & Peñagaricano, F. (2024). Host and rumen microbiome contributions to feed efficiency traits in Holstein cows. *Journal of Dairy Science*, *107*(5), 3090–3103. https://doi.org/10.3168/jds.2023-23869

Martínez-Álvaro, M., Zubiri-Gaitán, A., Hernández, P., Greenacre, M., Ferrer, A., & Blasco, A. (2021). Comprehensive functional core microbiome comparison in genetically obese and lean hosts under the same environment. *Communications Biology*, *4*(1), 1246. https://doi.org/10.1038/s42003-021-02784-w

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., Aydemir, O., Bakir-Gungor, B., Santa Pau, E. C., D'Elia, D., Desai, M. S., Falquet, L., Gundogdu, A., Hron, K., Klammsteiner, T., Lopes, M. B., Marcos-Zambrano, L. J., Marques, C., Mason, M., … Claesson, M. J. (2021). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.635781

Mulder, I. E., Schmidt, B., Stokes, C. R., Lewis, M., Bailey, M., Aminov, R. I., Prosser, J. I., Gill, B. P., Pluske, J. R., Mayer, C.-D., Musk, C. C., & Kelly, D. (2009). Environmentally-acquired bacteria influence microbial diversity and natural innate immune responses at gut surfaces. *BMC Biology*, *7*(1), 79. https://doi.org/10.1186/1741-7007-7-79

Mutumi, G. L., Jacobs, D. S., & Winker, H. (2017). The relative contribution of drift and selection to phenotypic divergence: A test case using the horseshoe bats *Rhinolophus simulator* and *Rhinolophus swinnyi*. *Ecology and Evolution*, *7*(12), 4299–4311. https://doi.org/10.1002/ece3.2966

Namkung, J. (2020). Machine learning methods for microbiome studies. *Journal of Microbiology*, *58*(3), 206–216. https://doi.org/10.1007/s12275-020-0066-8

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux

B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). _vegan: Community Ecology Package_. R package version 2.6-4, <https://CRAN.R-project.org/package=vegan>.

Palermo, G. (2009). Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry*, 57. https://doi.org/10.2147/AABC.S3619

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., Novielli, P., Tonda, A., Simeon, A., Shigdel, R., Béreux, S., Vitali, G., Tangaro, S., Lahti, L., Temko, A., Claesson, M. J., & Berland, M. (2023). Machine learning approaches in microbiome research: challenges and best practices. In *Frontiers in Microbiology* (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fmicb.2023.1261889

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202. https://doi.org/10.1038/nmeth.2658

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). *Scikit-learn: Machine Learning in Python*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: unbiased boosting with categorical features*.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, *13*(11). https://doi.org/10.1371/journal.pcbi.1005752

Rosenberg, E., & Zilber-Rosenberg, I. (2018a). The hologenome concept of evolution after 10 years. *Microbiome*, *6*(1), 78. https://doi.org/10.1186/s40168-018-0457-9

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology*, *71*(3), 1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005

Senn, V., Bassler, D., Choudhury, R., Scholkmann, F., Righini-Grunder, F., Vuille-dit-Bille, R. N., & Restin, T. (2020a). Microbial Colonization From the Fetus to Early Childhood—A Comprehensive Review. *Frontiers in Cellular and Infection Microbiology*, *10*. https://doi.org/10.3389/fcimb.2020.573735

Silverman, J. D., Roche, K., Mukherjee, S., & David, L. A. (2020). Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, *18*, 2789–2798. https://doi.org/10.1016/j.csbj.2020.09.014

Spor, A., Koren, O., & Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, *9*(4), 279–290. https://doi.org/10.1038/nrmicro2540

Susin, A., Wang, Y., Cao, K. A. L., & Luz Calle, M. (2020). Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/nargab/lqaa029

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., & Schloss, P. D. (2020). A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *MBio*, *11*(3). https://doi.org/10.1128/mBio.00434-20

Wessels, A. G. (2022). Influence of the Gut Microbiome on Feed Intake of Farm Animals. *Microorganisms*, *10*(7), 1305. https://doi.org/10.3390/microorganisms10071305

Zhou, Q., Lan, F., Gu, S., Li, G., Wu, G., Yan, Y., Li, X., Jin, J., Wen, C., Sun, C., & Yang, N. (2023). Genetic and microbiome analysis of feed efficiency in laying hens. *Poultry Science*, *102*(4), 102393. https://doi.org/10.1016/j.psj.2022.102393

Zubiri-Gaitán, A., Blasco, A., & Hernández, P. (2023). Plasma metabolomic profiling in two rabbit lines divergently selected for intramuscular fat content. *Communications Biology*, *6*(1), 893. https://doi.org/10.1038/s42003-023-05266-3

ANEXO

TABLAS SUPLEMENTARIAS

Tabla suplementaria 1. Dimensiones del microbioma en las poblaciones.

Escenario	m^2	Ambiente (EM)	Especies totales		Especies con efecto		Especies Ausentes	
			Línea					
			Alta	Baja	Alta	Baja	Alta	Baja
M	0.15	0%	561	556	91	90	33	28
		20%	667	667	92	93	12	12
		50%	791	776	95	93	19	4
	0.5	0%	509	554	82	88	26	61
		20%	651	643	93	92	30	22
		50%	791	791	97	96	5	5
NMH	0.15	0%	558	570	91	94	23	35
		20%	669	678	99	97	6	15
		50%	780	797	95	98	4	21
	0.5	0%	505	584	81	97	10	89
		20%	679	668	98	97	18	7
		50%	789	786	97	98	14	11
HMH	0.15	0%	562	540	86	82	42	20
		20%	623	623	89	83	47	7
		50%	757	788	88	94	3	34
	0.5	0%	573	584	88	97	10	89
		20%	634	663	84	93	7	36
		50%	774	773	92	90	14	13

Tabla suplementaria 2. Porcentaje de varianza explicado por los dos primeros componentes del análisis de componentes principales (PCA)

Escenario	m^2	Ambiente (EM)	% de varianza explicada por los dos primeros componentes	
			Microbiota completa	Especies con efecto sobre el fenotipo
M	0.15	0%	7.6	8.8
		20%	5.2	7.3
		50%	2.6	4.8
	0.5	0%	8.1	8.7
		20%	5.3	7
		50%	2.6	5.6
NMH	0.15	0%	8.2	10
		20%	5.7	7.7
		50%	3	8
	0.5	0%	8.9	11.5
		20%	5.2	8
		50%	2.9	5.9
HMH	0.15	0%	9.1	10.1
		20%	5.7	10.2
		50%	8.7	12
	0.5	0%	8.7	12
		20%	4.8	9.5
		50%	3.5	9

FIGURAS SUPLEMENTARIAS

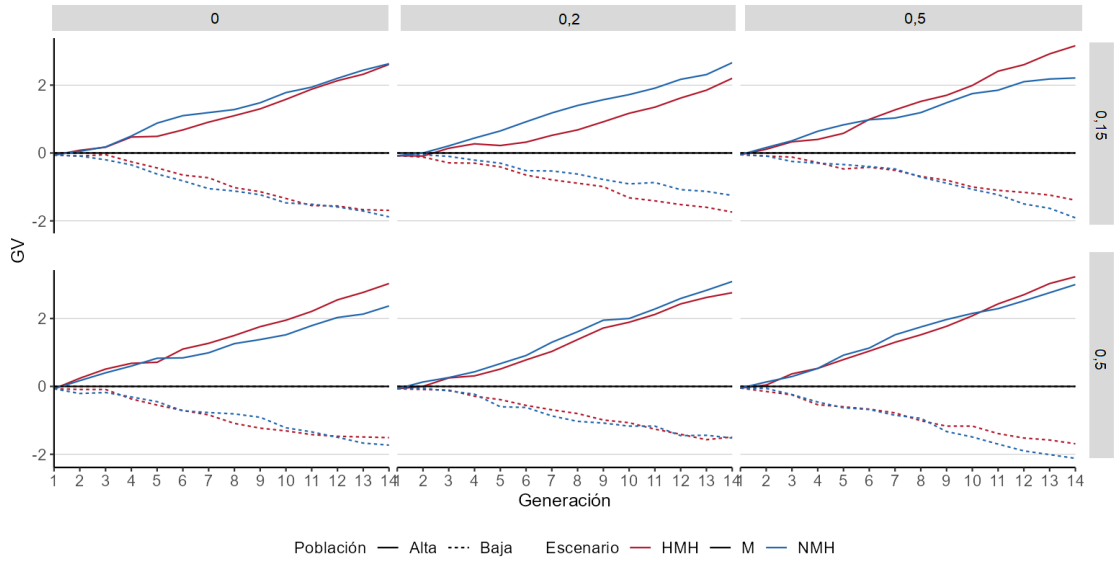

Figura suplementaria 1. Evolución del valor genético de las poblaciones divergentes tras 13 generaciones de selección. Escenarios M (negro), NMH (azul) y HMH (roja) de las poblaciones de alto (línea continua) y bajo (línea discontinua) tamaño de camada (TC) para valores de microbiabilidad de 0.15 y 0.5.

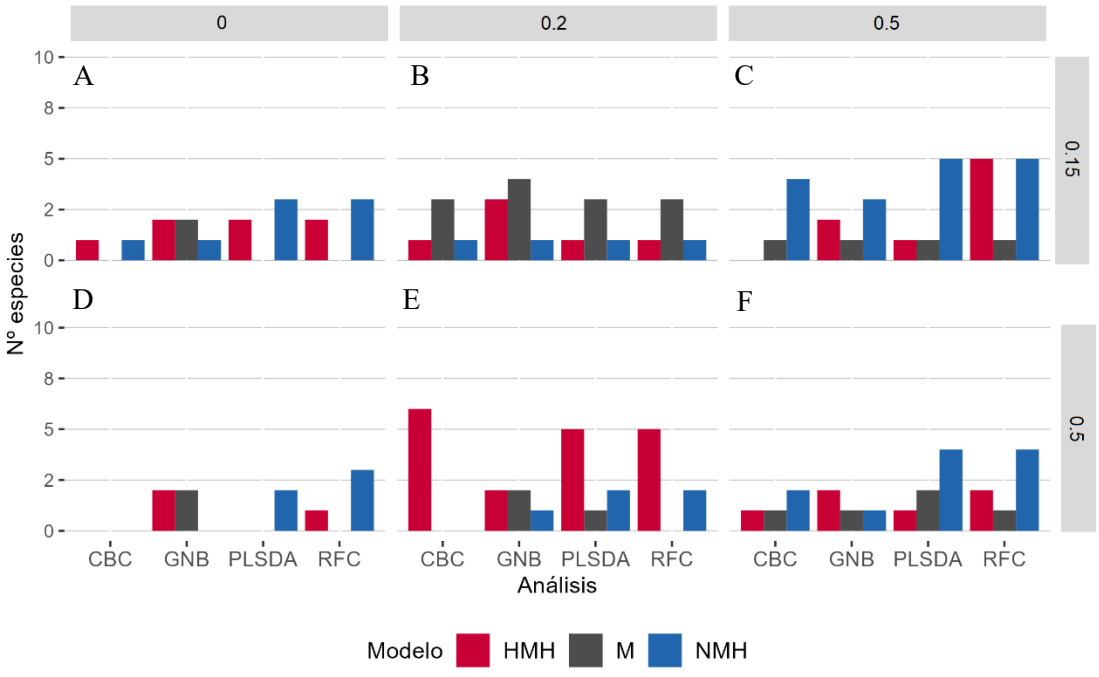

Figura suplementaria 2. Número de especies con efecto en el fenotipo entre las 10 primeras especies seleccionadas en modelos de clasificación. Escenarios M (negro), NMH (azul) y HMH (rojo) con microbiabilidad de 0.15 (A, B y C) y 0.5 (D, E y F) y valores de ambiente de 0 (A y D), 0,2 (B y E) y 0,5 (C y F) para los modelos CatBoost de clasificación (CBC), *Gaussian Naive Bayes* (GNB), *Partial Least Squares Discriminant Analysis* (PLSDA) y *Random Forest* de clasificación (RFC)

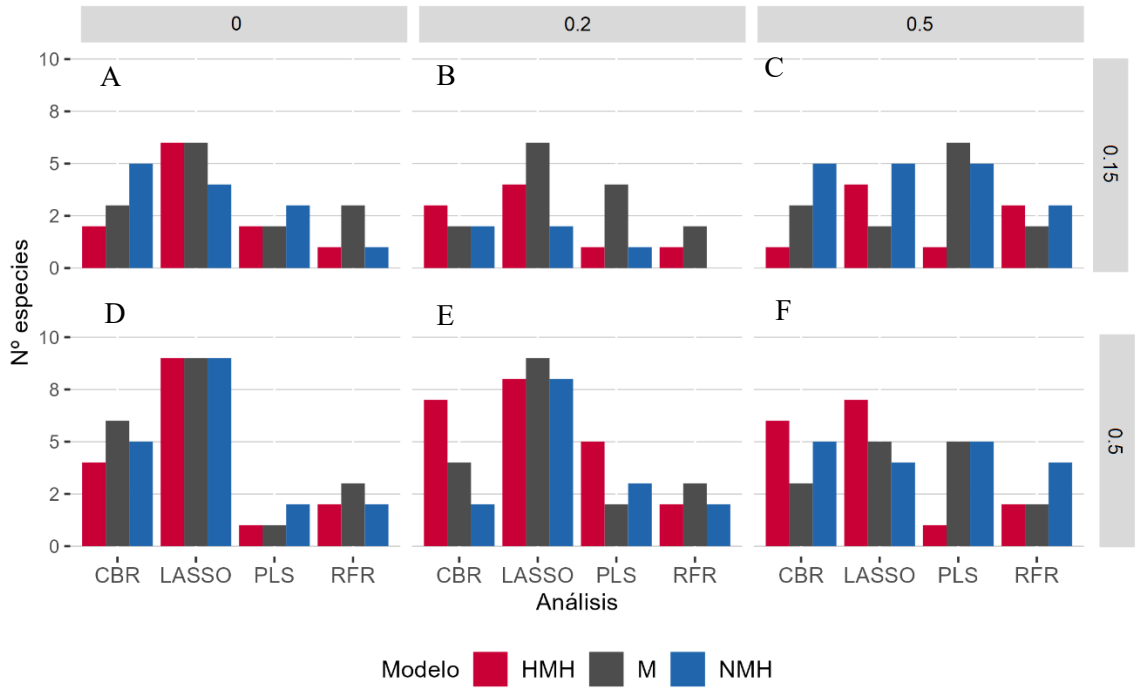

Figura suplementaria 3. Número de especies con efecto en el fenotipo entre las 10 primeras especies seleccionadas en modelos de regresión. Escenarios M (negro), NMH (azul) y HMH (rojo) con microbiabilidad de 0.15 (A, B y C) y 0.5 (D, E y F) y valores de ambiente de 0 (A y D), 0,2 (B y E) y 0,5 (C y F) para los modelos CatBoost de regresión (CBR), *Least Absolute Shrinkage and Selection Operator* (LASSO), *Partial Least Squares* (PLS) y *Random Forest* de regresión (RFR).

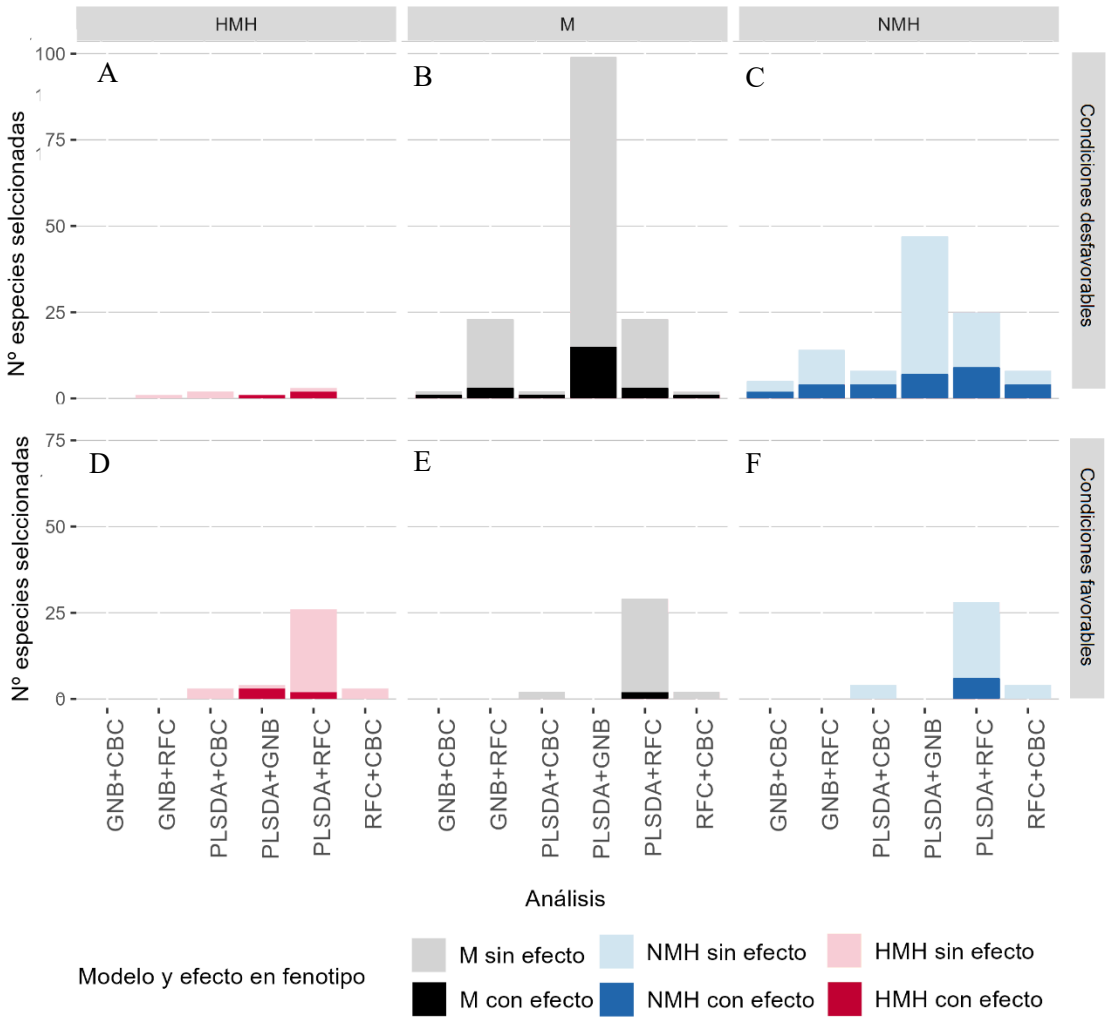

Figura suplementaria 4. Número de especies coincidentes entre modelos de clasificación por parejas. Escenarios HMH (rojo, rosa), M (negro, gris) y NMH (azul) bajo los parámetros desfavorables (A, B, C) con microbiabilidad (m^2) de 0.15 y proporción de especies adquiridas del ambiente (EM) de 0.5 en las tres gráficas superiores, y los parámetros favorables (D, E, F) con m^2 de 0.5 y EM de 0 en las tres gráficas inferiores.

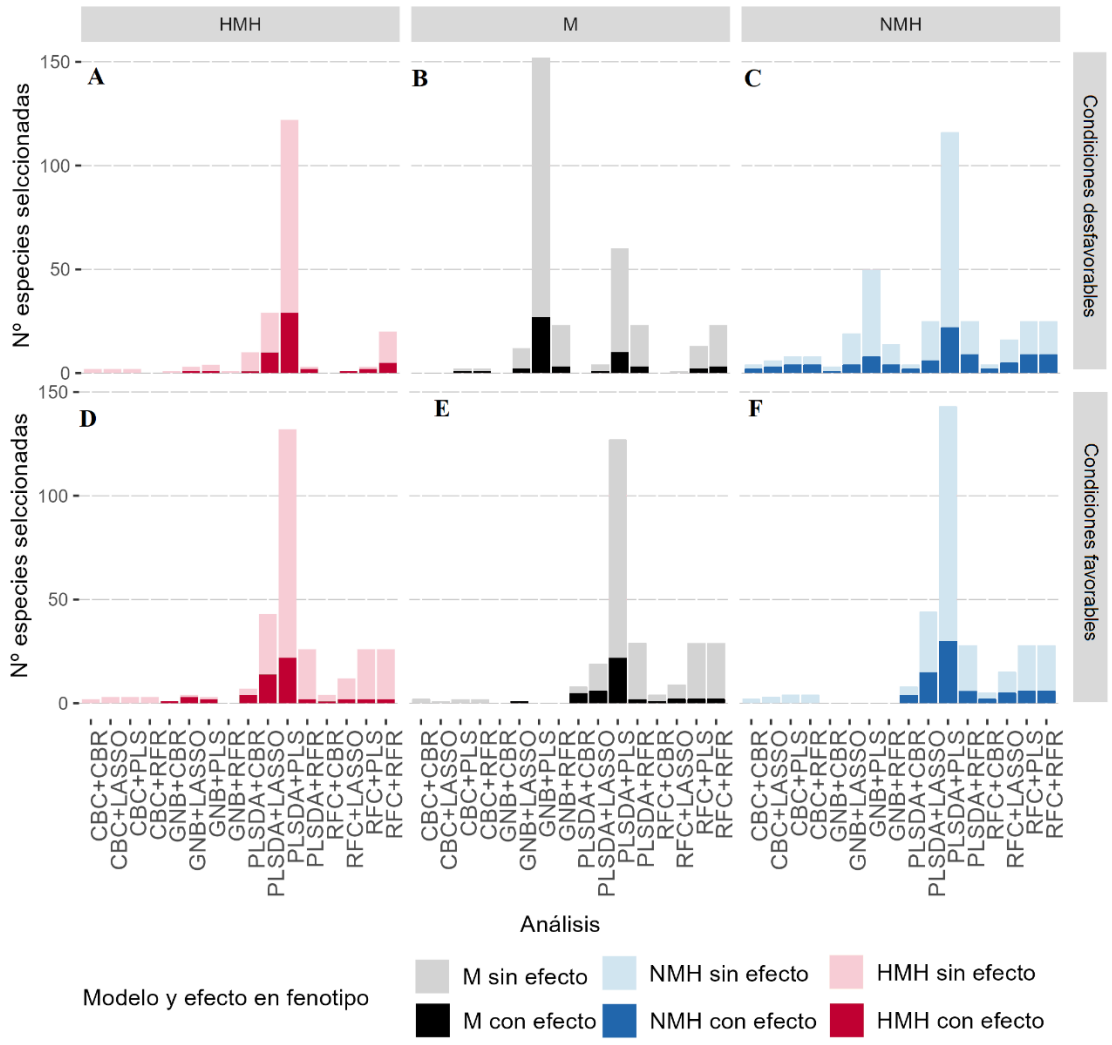

Figura suplementaria 5. Especies seleccionadas coincidentes entre metodologías de machine learning de clasificación y regresión por parejas de métodos. En la parte superior (A, B, C) las condiciones más desfavorables para el microbioma (m^2 de 0.15 y EM de 0.5), en la parte inferior (D, E, F) las más favorables (m^2 de 0.5 y EM de 0).

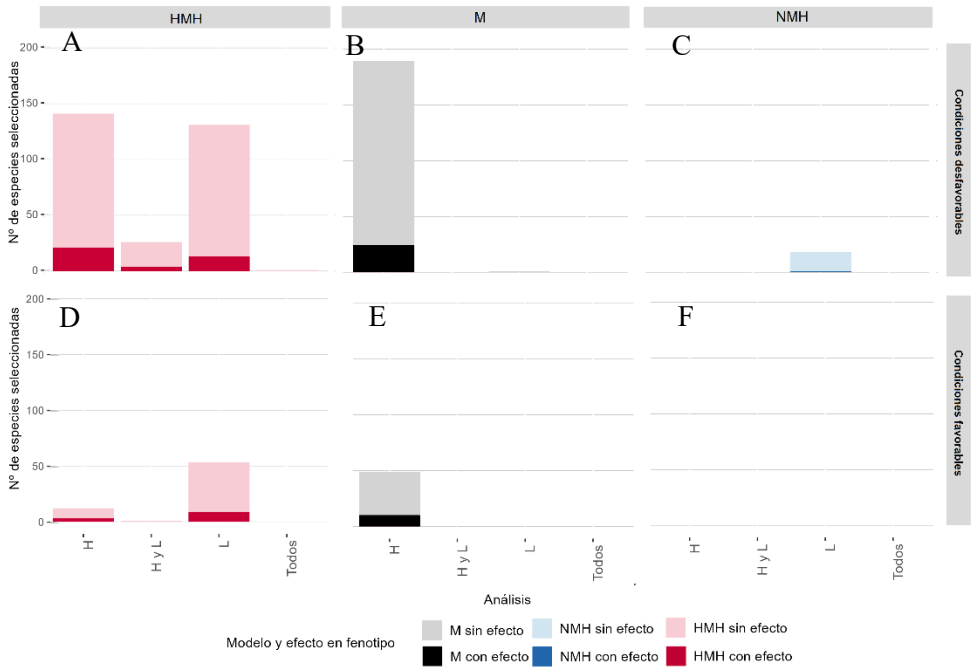

Figura suplementaria 6. Regresión por LASSO dentro de cada línea. Especies seleccionadas dentro de la línea alta (H), baja (L), coincidentes entre línea alta y baja (H y L) y coincidentes entre los análisis por línea y con ambas líneas (Todos) en los escenarios M (negro, gris), NMH (azul) y HMH (rojo, rosa) con parámetros desfavorables (m^2 de 0.15; EM de 0.5; gráficas A, B y C) y favorables (m^2 de 0.5; EM de 0; gráficas D, E y F).

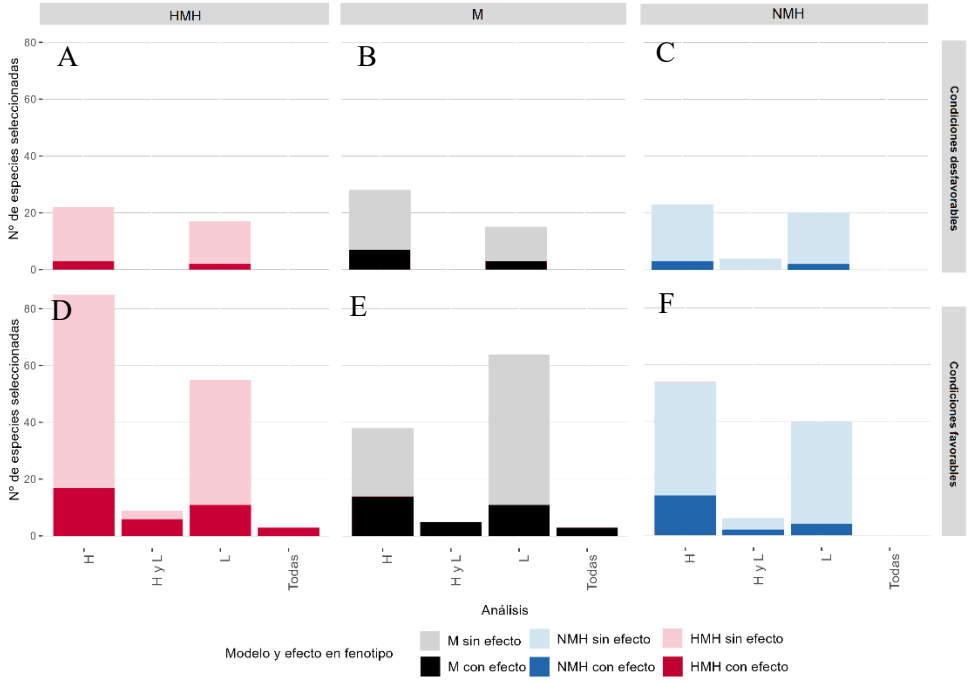

Figura suplementaria 7. Regresión por PLS dentro de cada línea. Especies seleccionadas dentro de la línea alta (H), baja (L), coincidentes entre línea alta y baja (H y L) y coincidentes entre los análisis por línea

y con ambas líneas (Todos) en los escenarios M (negro, gris), NMH (azul) y HMH (rojo, rosa) con parámetros desfavorables (m^2 de 0.15; EM de 0.5; gráficas A, B y C) y favorables (m^2 de 0.5; EM de 0; gráficas D, E y F).

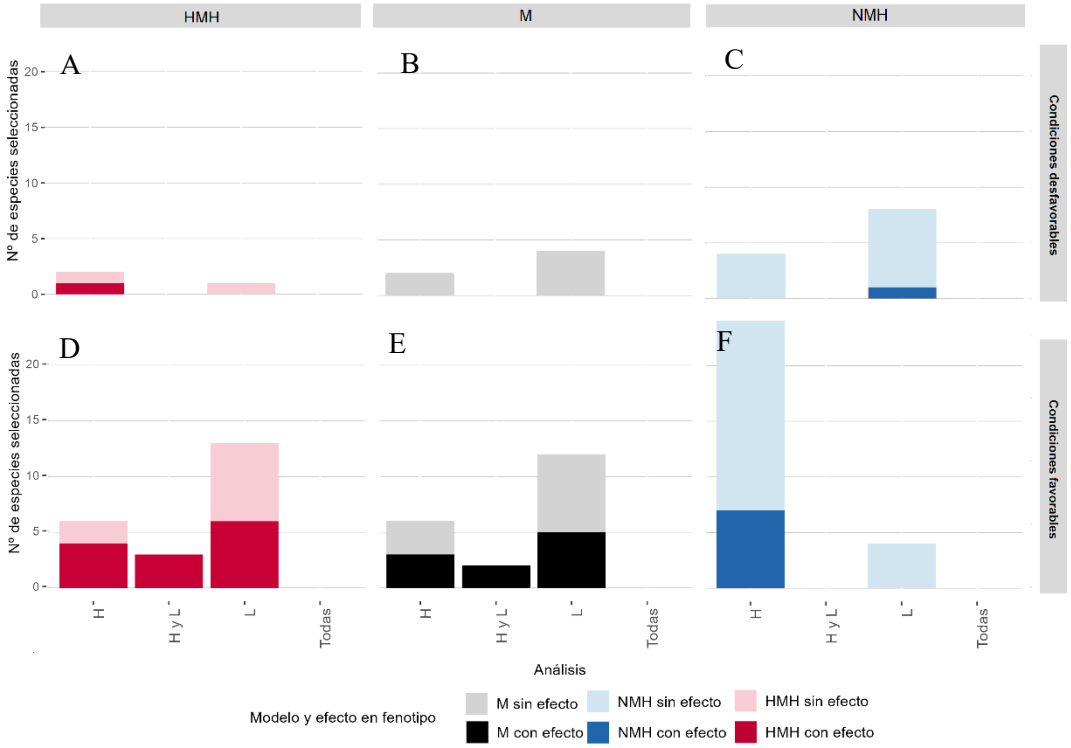

Figura suplementaria 8. Regresión por CBR dentro de cada línea. Especies seleccionadas dentro de la línea alta (H), baja (L), coincidentes entre línea alta y baja (H y L) y coincidentes entre los análisis por línea y con ambas líneas (Todos) en los escenarios M (negro, gris)), NMH (azul) y HMH (rojo, rosa) con parámetros desfavorables (m^2 de 0.15; EM de 0.5; gráficas A, B y C) y favorables (m^2 de 0.5; EM de 0; gráficas D, E y F).