Identificación de bacterias y falsos positivos: el papel de la deriva

A. Duro-Vizcaíno[*1], N. Ibáñez-Escriche[1], C. Casto-Rebollo[1]*

[1]Instituto de Ciencia y Tecnología animal. Universitat Politècnica de València, 46022, Spain
**Corresponding autor: alduviz@etsiamn.upv.es*

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 aprendizaje automático 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 cinco modelos de aprendizaje automático: PLS-DA, PLS, *Gaussian naive bayes* (GNB), *random forest* y LASSO. 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.

Los resultados de los análisis mostraron que todos los modelos de clasificación utilizados (PLS-DA, GNB y *random forest*) tuvieron un rendimiento para la clasificación de las poblaciones divergentes del 100% para todos los escenarios simulados. Aun así, en todos los supuestos, la mayoría de las especies seleccionadas no tuvieron efecto sobre el fenotipo. Los modelos de regresión PLS y LASSO mostraron un error cuadrático medio entre 4.54±0.52 para m^2=0.15 y 5.8±0.36 para m^2=0.5. El modelo de *random forest* obtuvo valores de 1.13±0.12 para m^2=0.15 y 2.1±0.22 para m^2=0.5. Se evaluó también cuántas especies con efecto en el fenotipo se encontraban entre el top 10 con mayor contribución en cada método. En general, los modelos de clasificación incluían una o ninguna, mientras que los de regresión incluían una media de 2-3 especies, destacando el modelo de regresión LASSO con una media de 6±1 especies. 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 son 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 resultados basados en

diferentes metodologías, así como la comparación de resultados entre modelos de clasificación y regresión podría ayudar a reducir los falsos positivos identificados. Los modelos de aprendizaje automáticos usados para predicción son óptimos. Sin embargo, hay que tener cuidado a la hora de sacar conclusiones biológicas relacionadas con las variables seleccionadas por dichos modelos.

Keywords: microbioma, simulación, selección divergente, selección de variables, deriva