APLICACIÓN DE TÉCNICAS DE ANÁLISIS DE DATOS COMPOSICIONALES AL MICROBIOMA DE CONEJOS

Zubiri-Gaitán, A., Martínez-Álvaro, M., Casto-Rebollo, C., Blasco, A. y Hernández, P.
Instituto de Ciencia y Tecnología Animal, Universitat Politècnica de València. Apartado 22012. Valencia 46022.
ablasco@dca.upv.es

INTRODUCCIÓN

Los datos obtenidos a partir de secuenciación masiva, incluida la secuenciación del microbioma, son datos de naturaleza composicional, es decir, son partes de un total. Todas las variables detectadas en una muestra (taxones) suman un total impuesto por el instrumento que difiere del total absoluto, desconocido (Gloor et al., 2017). El espacio muestral donde se representan los datos composicionales es un espacio restringido, conocido como símplex, cuya geometría (geometría de Aitchison) es diferente a la geometría Euclídea del espacio real.

Es posible representar los datos composicionales en coordenadas de una base ortonormal a partir de la transformación log-cociente isométrica (ilr). La coordenada ilr se define como el logaritmo del cociente de las medias geométricas entre particiones de la matriz de datos, denominadas balances. A partir del algoritmo *selbal*, es posible comparar balances en un proceso similar a la regresión stepwise hasta encontrar el balance óptimo que mejor explica una variable respuesta de interés (Rivera-Pinto et al., 2018).

El objetivo de este trabajo es aplicar técnicas de análisis de datos composicionales a datos microbiómicos de ciego para discriminar entre 2 líneas de conejos seleccionadas de forma divergente por el contenido de grasa intramuscular (GIM).

MATERIAL Y MÉTODOS

Se llevó a cabo un experimento de selección divergente para GIM durante 10 generaciones en conejos. El experimento se encuentra descrito en Martínez-Álvaro et al. (2016). Para el presente estudio se sacrificaron 16 y 17 conejos de la 10ª generación de la línea alta (GA) y baja (GB) respectivamente, a las 9 semanas de edad, previo ayuno de 4 horas. Inmediatamente tras el sacrificio se tomaron muestras del contenido cecal, se homogeneizaron y se almacenaron a -80ºC. Se obtuvo el metagenoma de las muestras mediante secuenciación con Illumina NextSeq y se realizó la asignación taxonómica de las lecturas a siete niveles (dominio, filo, clase, orden, familia, género y especie) utilizando el programa Kaiju (Menzel et al., 2016).

Para este análisis se utilizaron las variables taxonómicas obtenidas a nivel de género, 2653 en total. Se eliminaron las variables que contenían 0 en más del 25% de los individuos, quedando 1435. Posteriormente se realizó una estima puntual bayesiana para sustituir los 0 restantes por valores imputados (Martín-Fernández et al, 2015). Tras la imputación, los datos se clausuraron a 1; es decir, se estimaron las abundancias relativas por animal.

Se utilizó el paquete de R *selbal* para seleccionar el balance en base logarítmica que mejor discrimina entre las líneas GA y GB. La capacidad del balance seleccionado para discriminar GA y GB se testó con un modelo de regresión logística usando el balance como variable explicativa y un vector de clasificación (GA/GB) como variable explicada. Finalmente, se estimó la distribución marginal posterior del ratio entre las líneas GA y GB para cada género microbiano que compone el balance, usando el programa Rabbit (Instituto de Ciencia y Tecnología Animal, Universitat Politècnica de València); se calculó su mediana, su $HPD_{95\%}$ y su probabilidad (P) de ser mayor o menor a 1.

RESULTADOS Y DISCUSIÓN

El balance que mejor discrimina entre las líneas GA y GB está compuesto por 4 géneros bacterianos: *Firmicutes Caldicoprobacter, Firmicutes Dendrosporobacter, Candidatus Vecturithrix* y *Acidobacter_na*, este último pertenece al filo *Acidobacter*, cuyo género no pudo ser caracterizado. El balance es:

$$\ln\left(\frac{\sqrt{(F.Caldicoprobacter * F.Dendrosporobacter)}}{\sqrt{(Acidobacter\_na * C.Vecturithrix)}}\right) \quad (1)$$

El modelo de regresión logística utilizado para predecir la variable categórica GA/GB a partir del valor del balance se representa en la figura 1. En el eje Y se representa la probabilidad de asignación a GA o GB según el valor del balance, representado en el eje X. La zona media representa la posible zona de confusión en la asignación de clases que, como se puede observar, es muy reducida. La correlación entre los valores predichos y los observados fue de 0.98, mostrando por tanto una elevada capacidad de predicción.

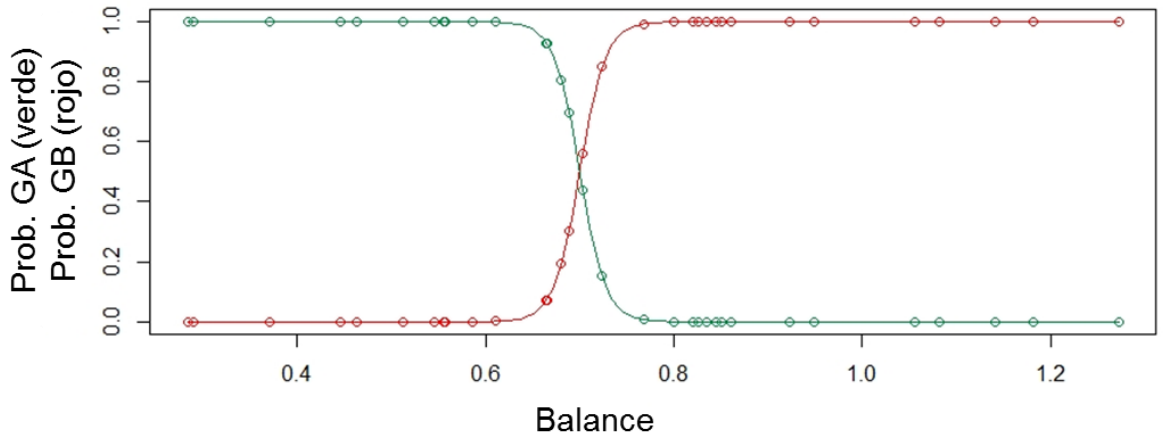

***Figura 1**. Representación de la regresión logística. En el eje X se encuentran los posibles valores del balance y en el eje Y la probabilidad de asignación a GA o GB según dicho valor.*

Además, en la figura 2, se presenta el gráfico de caja con patillas del balance seleccionado para ambaslíneas. Como se puede ver, el balance seleccionado es capaz de diferenciar las líneas GA y GB.

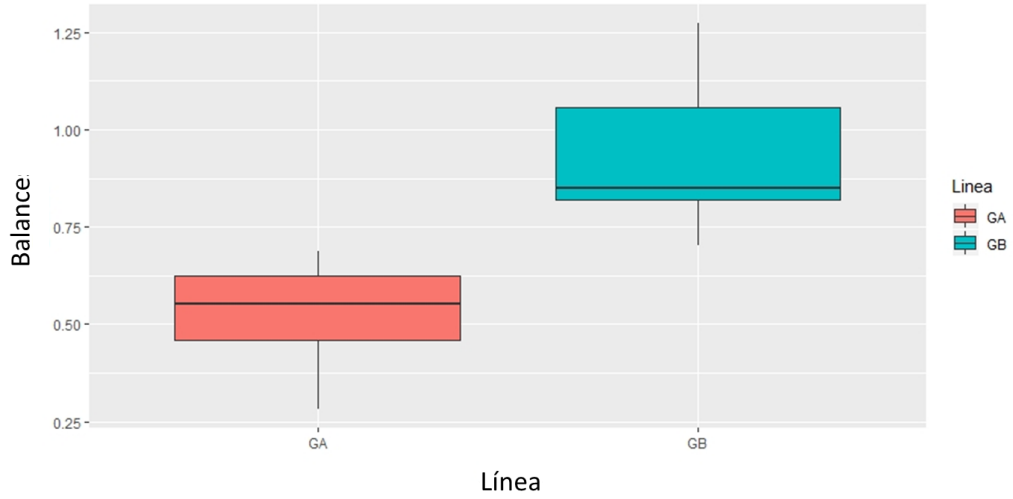

***Figura 2**. Gráfico de caja con patillas de la distribución del balance para las líneas GA (rojo) y GB (azul).*

En la tabla 1 se muestran los parámetros descriptivos (media y coeficiente de variación) y los resultados del análisis bayesiano del balance y de los géneros bacterianos que lo componen. Los resultados obtenidos evidencian que el balance conformado por estos 4 géneros, encontrado por el algoritmo *selbal*, fue capaz de discriminar con gran precisión los individuos de ambos grupos. Por otra parte, se puede ver que *Firmicutes Dendrosporobacter* tiene mayor abundancia relativa en GB, mientras que *Candidatus Vecturithrix* y *Acidobacter_na* la tienen en GA (tabla 1). Por otro lado, como se evidencia con el caso de *Firmicutes Caldicoprobacter*, es posible que algunas variables individuales no presenten diferencias relevantes entre los grupos, pero que su importancia se deba a su relación con las otras variables, puesto que los

balances se construyen comparando variables y grupos de variables entre sí. Debido a la naturaleza composicional de los datos, los valores individuales dados en porcentaje producen correlaciones espúreas, y además los valores individuales en subconjuntos de variables no respetan las relaciones mutuas originales entre variables, por lo que sólo las comparaciones relativas entre variables (y no sus valores individuales) pueden analizarse con los procedimientos habituales (Pawlowsky et al., 2015).

Tabla 1. *Parámetros descriptivos y ratios GA/GB del balance y de las abundancias relativas de los géneros bacterianos que conforman el balance seleccionado.*

Variable	Media (%)	CVx100	GA/GB	P	$HPD_{0,95}$
Balance (ecuación 1)	0,73	34	0,57	1,00	0,48 , 0,67
Firmicutes Dendrosporobacter	2,44E-03	24	0,85	0,97	0,70 , 1,00
Firmicutes Caldicoprobacter	10,14E-03	18	0,99	0,53	0,86 , 1,14
Candidatus Vecturithrix	1,65E-03	34	1,38	1,00	1,06 , 1,74
Acidobacter_na	4,08E-03	26	1,33	1,00	1,11 , 1,56

GA/GB: mediana de la distribución marginal posterior del ratio entre las líneas GA y GB; P: probabilidad del ratio de ser >1 cuando GA>GB y <1 cuando GA<GB; $HPD_{95\%}$: intervalo de máxima densidad posterior con un 95% de probabilidad.

REFERENCIAS BIBLIOGRÁFICAS

• Gloor, G.B. et al. 2017. Front. Microbiol. 8: 2224. • Martínez-Álvaro, M., Hernández, P. & Blasco, A. 2016. J. Anim. Sci. 94: 4993-5003. • Martin-Fernandez, J.A. et al. 2015. Stat. Model. 15:134-158. • Menzel, P. et al. 2016. Nat. Commun. 7:11257 • Pawlowsy, V., Egozcue, J.J. & Tolosana, R. 2015. Wiley. • Rivera-Pinto, J. et al. 2018. mSystems 3: e00053-18

Agradecimientos: Este experimento ha sido subvencionado por el proyecto AGL2017-86083-C2-1-P del Plan Nacional de Investigación.

RABBIT MICROBIOME ANALYSIS THROUGH COMPOSITIONAL DATA ANALYSIS TECHNIQUES

ABSTRACT: Compositional data are defined in a sample space called simplex with Aitchison geometry, which is different from the Euclidean geometry from the real space. For this study, compositional nature of microbiome data was considered in order to analyze the microbiota divergence between two rabbit lines selected for high (H) and low (L) intramuscular fat (IMF) content. An algorithm called *selbal* was applied to the genera microbiota dataset to find the optimal balance capable of discriminating between lines. The balance obtained included four bacterial genera, grouping *Firmicutes Caldicoprobacter* and *Firmicutes Dendrosporobacter* on one side, and *Candidatus Vecturithrix* and *Acidobacter_na* on the other. The balance was able to successfully discriminate between lines. Univariate Bayes analysis of the ratio H/L showed greater relative abundance of *Firmicutes Dendrosporobacter* in L and lower of *Candidatus Vecturithrix* and *Acidobacter_na* in H.

Keywords*:* compositional data, isometric log-ratio, balance, caecum microbiome, rabbit