THE METAGENOME CAN PREDICT FEED EFFICIENCY: A VALIDATION STUDY

González-Recio[1,2],O., Delgado[1], B., Guasch[3], I., González[1], C., Elcoso[3], G., Pryce[4], J.E., Bach[5], A.

[1] Departamento de Mejora Genética Animal. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria O.A., M.P. 28040 Madrid, Spain. [2]Departamento de Producción Agraria. Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas. UPM. Ciudad Universitaria s/n 28040 Madrid, Spain. [3] Blanca from the Pyreness, 25795 Hostalets the Tost, Spain. [4] Bioscience Research Division, ECODEV, Bundoora 3038, Australia. [5] Department of Ruminant Production, ICREA-IRTA, 08140 Caldes de Montbui, Spain.

correo-electrónico: gonzalez.oscar@inia.es

INTRODUCCIÓN

Recent research has proposed the microbiota as a proxy or phenotype to predict complex traits, such as body mass index in humans or feed efficiency in livestock animals (Ross et al., 2013; Beamount et al., 2016). Furthermore, links have been observed between the host genotype and the gastrointestinal microbial composition (Roehe et al., 2016; Camarinha-Silva et al. 2017; González-Recio et al., 2013), proving that the microbial communities that populate the individual digestive niches are not only dependent on environment and diet, but also on the host genotype. Microbiome research is gaining attention in livestock species, as it assists on understanding diseases and efficiency processes that occur in animals. Previous studies have related well-known taxonomical groups or community composition with feed efficiency or residual feed intake (RFI) (Roehe et al., 2016). Most of those studies have used 16S rRNA sequencing as a description of the microbiota. Few studies have used whole metagenome sequencing, and their results have not yet been validated (Ross et al., 2013). Recording individual feed efficiency is extremely cumbersome and expensive, it has become an important limitation to improve feed efficiency from genetic selection. The metagenomics era offers new opportunities to use microbiome composition to assess feed intake of an individual and its relationship with metabolic processes involved in the digestion, absorption and utilization of nutrients. The objective of this study was i) to unveil potential associations between the rumen microbiota and feed efficiency related traits in dairy cattle, 2) to investigate the possibilities to use the metagenome as a proxy for these traits across individuals and different environments.

MATERIAL Y MÉTODOS

Eighty Holstein cows from the experimental farm *BLANCA from the Pyrenees* were monitored during 2 weeks. All animals were under the same management routines, eat the same diet based on fescue, ryegrass and concentrate and were in the same lactation stage (between 60-120 days post-partum). Daily milk yield, fat and protein contents, dry matter intake, and body weight during the study period were averaged to obtain a single record per cow. Daily feed efficiency was calculated as the ratio between milk yield (kg/d) and dry matter intake (kg/d), and the average along the study was used as FE phenotype. Residual feed intake was calculated as the difference between observed and expected dry matter intake.

Rumen content (approximately 50 mL) from each cow was sampled at day 7 using a stomach tube. All samples were frozen immediately after the extraction and then stored at -80ºC until analysis. DNA extraction was performed using 250 µl from the homogenized samples with the "DNeasy Power Soil Kit" (QIAGEN, Valencia, CA, USA), and following the manufacturer's instructions. The 30 cows with extreme feed efficiency phenotype (15 larger and 15 lowest feed efficiency record) were selected to sequence the whole metagenome of their rumen digesta sample. Illumina libraries were prepared from the extracted DNA and sequenced on Illumina MiSeq v3 systems (2x300) by Fisabio (Valencia, Spain). De novo assemble of the metagenome was carried out using MEGAHIT. Then, microbial functional genes encoding for proteins (contigs) were identified using the KEGG genes database with PROKKA. Quantification of contigs in each sample was performed with SALMON. The normalized number of contigs per million (CPM) was used in downward analyses.

metaGenome Wide Association Analyses

Pre-selection of contigs was performed using the information gain or entropy reduction criterion (Ewens y Grant, 2005; Long et al., 2007). Information gain is the difference in entropy of a probability distribution before and after observing a variable contig, i.e., it measures how much uncertainty is reduced by observation of CPM. The information gain for each contig *k* (*k*= 1,2, …, 174,247) was the change in entropy after observing the CPM, calculated as:

$$IG(contig_k) = H(\Pr(\mathbf{Y})) - \frac{1}{N_k^{High} + N_k^{Low}} \sum_{s=High,Low} \left(N_k^s \log\left(s_{y|s} \sqrt{2\pi g} \right) \right),$$

The contigs with largest information gain in the top 95 percentile were pre-selected for subsequent analyses. Association analysis was performed using logistic regression of CPM observations from each contig on the observed response ('HIGH' or 'LOW', codified as 1 or 0, respectively). Hierarchical clustering was then performed using the average distance matrix method. Bootstrap (BP) and approximately unbiased (AU) values were computed.

Independent validation

Data used in the validation set were 16 and 15 cows from batches 1 and 2 (Macdonald et al., 2014) from the Victorian Department of Primary Industries Ellinbank Centre near Warragul, Victoria Australia (latitude 38 14`S, longitude 145 56`E). Cows received feed ad libitum, and were monitored for 32d and 37d, respectively. Individual intakes were determined using electronic monitoring of load cells under feed bins (Gallagher Animal Management Systems, Hamilton, New Zealand) and electronic identification of individual animals.

Rumen fluid was collected via stomach pump. DNA was extracted using the PowerMax Soil DNA Isolation kit (MoBio) and sequenced on the HiSeq2000 (Illumina). Using the CPM of selected contigs as predictors, we compute an estimator $\hat{\boldsymbol{\beta}}$ of the linear effects of CPM on the phenotypes as follows:

$$y_i = y_0 + \mathbf{x}_i \boldsymbol{\beta} + e_i$$

The vector of linear effects of contigs, $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, was estimated using L1-penalized regression (LASSO). This corresponds to minimizing the objective function

$$\hat{\beta} = \arg\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} O_\lambda(\mathbf{y}, \mathbf{X}; \beta) \ , \ O_l(\mathbf{y}, \mathbf{X}; b) = \tfrac{1}{2} \left\| \mathbf{y} - \mathbf{X}b \right\|^2 + n\lambda \left\| b \right\|_1,$$

Then, FE and DMI were estimated in the Australian population as $\hat{y}_i^* = \mathbf{x}_i^* \hat{\boldsymbol{\beta}}$, where $\hat{y}_i^*$ is the estimated phenotype for the *i*th individual in the testing set (*i*=1,…,31), and $\mathbf{x}_i^*$ is the corresponding *i*th-row of the contig CPM design for the validation set ($\mathbf{X}^*$).

RESULTS AND DISCUSSION

The metagenome assembly resulted in 941,823 contigs (summary statistics). Filtering on information gain left 8713 contigs for the metaGWAS. Four hundred and forty eight contigs were statistically associated with feed efficiency. Average Spearman correlation between these contigs and FE was above 0.50 (in absolute value). Large correlation was also observed with RFI and DMI. Milk yield, milk solids, and body weight had weaker associations with these contigs. Larger (absolute) values were found for RFI and FE, averaging 0.40 and 0.36, respectively. Correlation with DMI was also relevant, averaging 0.39. Productive traits showed weaker correlations (0.09-0.13) with the selected contigs. A weak association was found with body weight. The classification accuracy into either high or low efficiency was 0.93. The uncertainty of these two clusters was computed via multiscale bootstrap resampling, resulting in p-values <0.05. Hence, this is strong support against the null hypothesis, and we conclude that these contigs are associated with FE. The metaGWAS for DMI selected 523 contigs. Hierarchical clustering also differentiated between cows with high and low DMI. Interestingly, classification within FE group was highly accurate (Figures 1a and 1b). Multiscale bootstrap resampling in the low FE group, resulted in 2 most probable clusters for high and low DMI, with p-values <0.05 and classification accuracy of 93% (Figure 1c). The same uncertainty analyses

in the high FE group resulted in two clusters (P<0.05) (Figure 1d) and classification accuracy of 73%.

The validation in the Australian population resulting in a low, but positive predictive accuracy was obtained for both traits (0.25 and 0.36 for FE and DMI, respectively). In order to randomize these analyses, 1,000 replicates were obtained with random sampling of the contigs at each iteration. The accuracies obtained after selection of contigs from metaGWAS were in the 75th and 70th percentile of the distribution from random selection. These accuracies were above average, but they demonstrate limited statistical power from metaGWAS. There is room for improved predictive accuracy involving larger data sets and possibly more sophisticated statistical methods.

REFERENCES

•Ross, EM, et al. PLoS One 2013. 8:e73056. •urnbaugh, PJ, et al. 2006. Nature 444:1027–131. •Wang, J, & Jia, H. 2016. Nat Rev Microbiol 14:508–22. •Roehe, R, et al. 2016. PLoS Genet 12:e1005846. •Beaumont, M, et al. 2016. Genome Biol 17:189. •Camarinha-Silva, A, et al. 2017. Genetics 206:1637–44. •Bonder, MJ, et al. 2016. Nat Genet. 48. •Imhann, F, et al. 2016. Gut. gutjnl-2016-312135. •Gonzalez-Recio, O, et al. 2018. J Dairy Sci. 101:2285–92. •Pryce, JE, et al. 2015. J Dairy Sci 98:7340–50. •Long, N, et al. 2007. Mach Learn. 124:377–89. •Ewens, WJ & Grant, G. 2005. Statistical Methods in Bioinformatics. New York, NY: Springer New York; •Macdonald, KA, et al. 2014. J Dairy Sci. 97:1427–35. (Texto justificado) (1 línea en blanco)

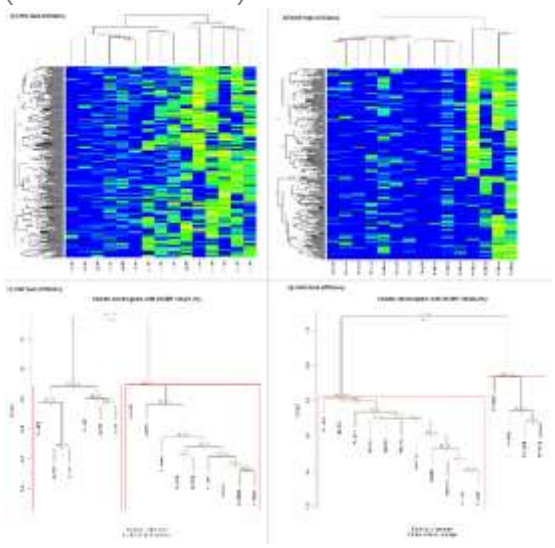

***Figure 1**. Hierarchical clustering for feed intake levels (top) within feed efficiency group, low (left) or high (right), from metaGWAS analyses for dry matter intake.*

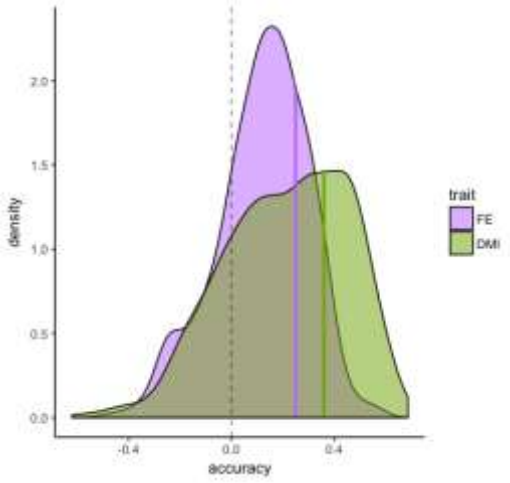

***Figure 2**. Randomized predictive accuracy in the Australian population. Purple and green vertical lines indicated accuracies obtained with the selected contigs.*

THE METAGENOME CAN PREDICT FEED EFFICIENCY: VALIDATION STUDY

ABSTRACT: Several microbial genes (*e.g.*, RPOA, FABF, LACZ, METH, FABG, AROA) involved in fatty acids and cellulose degradation pathways showed enrichment (P<0.05) in high efficient cows. Pre-selection of microbial contigs allowed high classification accuracy for feed efficiency and intake levels using hierarchical classification. These microbial contigs were also able to predict feed efficiency and intake levels with accuracy of 0.25 and 0.36 for FE and DMI, respectively, in an independent population. Nonetheless, a larger potential accuracy of up to 0.69 was foreseen in this study for datasets with a larger statistical power. The findings indicated that there are differences between the microbiota compositions of high and low efficient animals both at the taxonomical and gene levels. Some of these differences remain

even between populations under different diets and environments and can provide information on the feed utilization performance without having intake level information.

Keywords*:* feed efficiency, metagenome, metaGWAS, prediction.