Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 225-228

METAGENOMIC PREDICTIONS FOR ENTERIC METHANE EMISSIONS IN SHEEP USING LONG-READ SEQUENCING OF RUMEN FLUID SAMPLES

Y. Li[1], L.T. Nguyen[1], C.T. Ong[1], S. Yadav[1], M. Aldridge[2,3], P. Fitzgerald[2], J. van der Werf [2] and E. M. Ross[1]

[1] The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, St Lucia, QLD, 4072 Australia
[2] University of New England, School of Environmental and Rural Science, Armidale, NSW, 2350 Australia
[3] Animal Genetics Breeding Unit*, University of New England, Armidale, NSW, 2350 Australia

SUMMARY

Metagenomic data were generated using Oxford Nanopore Technologies (ONT) platform for 396 sheep rumen fluid samples. Taxonomic and gene functional abundance matrices were constructed to develop predictive models for estimating methane emission phenotypes based on methane concentration from portable accumulation chambers (PAC). The greatest microbiability (m^2; proportion of phenotypic variation explained by the microbiome relationship matrices) was generated using Clusters of Orthologous Groups (COGs) functional annotations of the reads (m^2 = 0.92). Correlations between the predicted phenotypes and the PAC phenotypes in a 5-fold cross-validation ranged from 0.48 to 0.51. These results indicate that metagenomic predictions could be implemented as a proxy methane phenotype in sheep and could be used to increase the feasible reference population size in genomic predictions for methane using multi-trait models.

INTRODUCTION

Methane emission from ruminant livestock is one of the main contributors to greenhouse gas production from the agricultural sector. Enteric methane emission levels can be predicted from the variation in rumen microbiota composition and metabolic activities (Ross and Hayes 2022). Recent advances in long-read sequencing technology have enhanced recognition of genetic elements, taxonomic identification and functional annotation in metagenomics (Kim *et al*. 2024). We applied long-read metagenomic sequencing of rumen samples to identify which microbial taxonomic or functional classification approach best captures methane phenotypic variation and compared prediction accuracies across these different classification methods.

MATERIALS AND METHODS

Collection of methane emission data and associated parameters. Methane emissions were measured using portable accumulation chambers (PAC) for 50 minutes per animal in 396 lambs aged 5.75 ± 0.12 months. The lambs were sourced from the 2023 lambing season of MLA resource flock at Kirby (Armidale, NSW). Methane concentrations were measured in parts per million (ppm) using an Eagle2 Gas Monitor (RKI Instruments). The recorded gas concentrations were corrected using CH_4 ppm × standard temperature and pressure. Associated fixed effects were measured simultaneously with CH_4 phenotype measurements, including body weight (kg), sampling date, and time off feed, which accounts for key potential confounding effects.

Metagenomic library preparation, long-read sequencing and basecalling. For metagenomic sequencing, libraries of 396 samples were prepared using the Oxford Nanopore Native Barcoding 11Kit (SQK-NBD114.96). Following end preparation and clean-up, DNA fragments were ligated

* A joint venture of NSW Department of Primary Industries and Regional Development and the University of New England

with unique barcodes and sequencing adapters, then pooled for sequencing on the PromethION platform (ONT). Dorado basecaller (v0.7.0) with the "Super Accuracy" model v4.3.0 generated a total of 4TB of long-read sequencing data from the raw dataset. Reads with an average Phred quality score below 10 and length < 200 bp were removed from the dataset.

Taxonomy and functional identification. SqueezeMeta v1.6.3 (Tamames and Puente-Sánchez 2018), was used to assign individual reads to taxa at the levels of Phylum to Species, as well as classify the function of the genes included in each read using both the EggNOG (Huerta-Cepas et al., 2019) and KEGG (Kanehisa and Goto 2000) databases. Within SqueezeMeta, Prodigal v2.6.3 was used for protein-coding gene prediction, DIAMOND v2.1 for protein alignment against NCBI non-redundant protein databases, and EggNOG (v5.0) and KEGG (v110) databases for metabolic pathway reconstruction. Taxa assigned to the kingdoms of *Bacteria*, *Archaea*, and *Eukarya* were retained for microbially abundant matrices, with an average of 65.7% of sequence counts not assigned to these taxonomic classifications being removed. Eight distinct abundance matrices were constructed for microbial features composition: six for taxonomic levels (*Phylum* to *Species*) and two for functional annotations marked as KEGG and COGs (Clusters of Orthologous Groups).

Statistical model definition. Low-abundance microbial taxa and pathways (count < 10) were removed. Counts were normalised to relative abundances (percentages) per sample, summing to 100%. Both methane emissions (ppm) and microbial abundances were scaled to z-scores for modeling. ASReml-R 4.2 (Butler and Cullis 2023) was used to fit a BLUP-based mixed linear model:

$$y = Xb + Zu + e$$

where, y was the vector of phenotypic values (CH_4 emissions). X and Z were design matrices linking observations to fixed and random effects, respectively. The fixed effects, b included linear cofactors such as weight and time off feed prior to entering the PAC. The microbiome random effects u were assumed to follow a multivariate normal distribution as $u \sim N(0, G\sigma_m^2)$. The G was a variance-covariance (microbial relationship) matrix derived from the standardised taxonomic or functional abundance matrix *M*,

$$G = MM'/n$$

where M was the dimensions $i \times n$ standardised taxonomic or functional abundance matrix, n was the number of features (taxa or functional pathways) and i was the number of samples (Ross and Hayes 2022). This matrix represents pairwise covariances among samples based on their microbial abundance profiles. The variance component σ_m^2 associated with the G quantifies the proportion of total variation explained by composition of microbiome features. The microbiome variance component σ_m^2 (V_m), estimated to be using Residual Maximum Likelihood (REML), represented the proportion of CH_4 emission variation attributed to the microbiome abundance matrix. The residuals were assumed to be normally distributed with variance σ_e^2. Environmental variance σ_e^2 (V_e), obtained from the BLUP model, represented the random error effects after accounting for weight, batch-level and microbial abundance effects. The proportion of the phenotypic variance explained by the microbial component (microbiability m^2), was calculated as the ratio of linear microbiome variance to total variance: $m^2 = V_m / (V_m + V_e)$, indicating the predictive ability of microbial characteristics by using abundances matrixes for CH_4 emissions.

Three distinct cross-validation strategies were performed to assess predictive microbial ability. These included 2-fold and 5-fold random partitioning with 20 iterations each, and a leave-one-batch-out approach testing generalisability across different sampling conditions. For the latter, one validation batch from 42 total batches (structured across 6 days with 7 different time off feed intervals) served as the testing set. The accuracy (r) was calculated as the average Pearson's correlation between predicted and observed methane phenotypes across all folds: $r = \frac{1}{k}\sum cor(y_i, \hat{y}_i)$. y_i represented the vector of methane phenotypes, and $\hat{y}_i$ were the corresponding

predicted phenotype value. The factor k represented the number of folds in cross-validation. To account for microbiability (m^2), the corrected accuracy was also reported as $\frac{r}{\sqrt{m^2}}$. All accuracy values were reported as a means with standard errors.

RESULTS AND DISCUSSION

The CH_4 concentrations, which were corrected for standard temperature and pressure, ranged from 123.21 to 1415.87 ppm. The mean (±SD) concentration was 647.69 (±238.67) ppm, with a median of 624.92 ppm. The methane yield coefficient of variation between animals was 36.85%.

Filtering removed 27.27% (±5.73%) of low-quality reads. The number of sequencing reads was 1.68×10^6 ($\pm1.08\times10^6$), with an average read length of 642.79 (±221.63) bp, a read length N50 of 1013.74 (±1012.71) bp, and an average sequencing accuracy of 98.06% (±0.25%).

The taxonomic and functional abundance matrices derived from long-read metagenomic analysis were quantified at multiple classification levels, with the feature counts summarised in Table 1. Notably, all taxonomic levels from long-read metagenomic explained more of the variance than short-read approaches in other studies (Xie *et al*. 2021). The microbiability (m^2) (Table 1) increased from phylum (0.27) to genus/species level (0.65), with stable standard errors (± 0.2) as the number of features (i.e. columns in the count matrix) increased. Functional abundance matrices had higher m^2 values compared to taxonomic abundance matrices. The microbiability for methane emission in this study showed more variance explained than previous approaches. Earlier Bayesian modeling studies reported lower microbial effects (comparable to m^2): average of 0.07 based on OTUs (Operational Taxonomic Units) taxonomic groups (Zhang *et al*. 2020), and <0.33 based on taxonomic abundance from long-read metagenomic data (Marcos *et al*. 2024).

Table 1. Microbiability (m^2) and variance components of microbiome data

Level	n	m^2	V_m ± SD	V_e ± SD	MSE
Phylum	242	0.27	0.13 ± 0.04	0.35 ± 0.03	0.73
Class	655	0.45	0.23 ± 0.05	0.28 ± 0.03	0.60
Order	1174	0.51	0.25 ± 0.06	0.24 ± 0.04	0.54
Family	2008	0.55	0.26 ± 0.06	0.21 ± 0.04	0.52
Genus	3042	0.65	0.30 ± 0.06	0.16 ± 0.04	0.45
Species	3042	0.65	0.30 ± 0.06	0.16 ± 0.04	0.45
KEGG	9591	0.80	0.48 ± 0.10	0.12 ± 0.04	0.39
COGs	22873	0.92	0.49 ± 0.08	0.04 ± 0.05	0.34

MSE (Mean Squared Error) = $\frac{1}{k}\sum(\hat{y}_i - y_i)^2$; SD: Standard Deviation.

Cross-validation showed robustness of corrected accuracy at 2-fold (range: 0.45-0.51) and 5-fold (range: 0.48-0.51) across all taxonomic and functional levels (Table 2). A previous study in sheep using covariance matrices based on metabolomic matrices and rumen microbial OTUs matrices in sheep revealed average accuracy (r) below 0.4 (Ross *et al*. 2020), although their sample size was much smaller. However, the leave-one-batch-out validation accuracy was lower (range: 0.18-0.33), potentially attributed to either large variation in the fixed effects (weight and time off feed) with unknown environmental conditions or the reduced test sample size which introduced unknown predictive errors. After correcting accuracies by microbiability (m^2), leave-one-batch-out (42-Fold) accuracy ranged from 0.32 to 0.39. The correlation between microbiome-predicted methane and PAC methane was comparable to the correlation between repeated PAC measurements, which ranged from 0.51 to 0.90 in sheep across 7 days (O'Connor *et al.* 2021).

Table 2. Accuracy estimation of cross-validation across taxonomic and functional levels

	2-Fold		5-Fold		42-Fold	
	r	r/m	r	r/m	r	r/m
Phylum	0.42 ± 0.05	0.49 ± 0.06	0.44 ± 0.03	0.49 ± 0.03	0.18 ± 0.06	0.34 ± 0.11
Class	0.41 ± 0.05	0.46 ± 0.04	0.43 ± 0.03	0.48 ± 0.03	0.23 ± 0.05	0.35 ± 0.08
Order	0.44 ± 0.04	0.51 ± 0.03	0.45 ± 0.05	0.51 ± 0.05	0.27 ± 0.06	0.39 ± 0.08
Family	0.40 ± 0.01	0.45 ± 0.05	0.44 ± 0.03	0.50 ± 0.03	0.26 ± 0.06	0.35 ± 0.08
Genus	0.41 ± 0.00	0.46 ± 0.02	0.44 ± 0.04	0.49 ± 0.04	0.26 ± 0.06	0.32 ± 0.08
Species	0.43 ± 0.01	0.49 ± 0.02	0.45 ± 0.04	0.51 ± 0.05	0.26 ± 0.06	0.32 ± 0.08
KEGG	0.42 ± 0.04	0.47 ± 0.04	0.45 ± 0.03	0.51 ± 0.04	0.32 ± 0.05	0.36 ± 0.06
COGs	0.42 ± 0.00	0.48 ± 0.04	0.42 ± 0.02	0.48 ± 0.03	0.33 ± 0.06	0.35 ± 0.06

r: Accuracy; r/m: Corrected Accuracy; All values are mean ± standard error.

CONCLUSIONS

Long-read metagenomic data analysis and appropriate model selection improved the prediction accuracy of methane emission phenotypes in this study. Variation explained by the microbiome relationship matrix was highest for the COGs functional categorisation. Higher prediction accuracy was achieved by including within cohort methane measurements in the reference set.

ACKNOWLEDGEMENTS

This research is funded by Meat and Livestock Australia and MLA Donor Company (P.PSH. 2010 and PSH.2011). The author acknowledges support through the Australian Government and The University of Queensland Research Training Program (RTP) Scholarship. The author gratefully acknowledges Collins Asiamah for essential DNA extraction work and thank Daniel Brown and Ed Clayton for their contribution to this project scope.

REFERENCES

Butler D.G., Cullis B.R., Gilmour A.R., Gogel B.G. and Thompson R. (2023) In 'ASReml-R Reference Manual Version 4.2' pp. 30-105, VSN International Ltd., HP2 4TP, UK.
Huerta-Cepas J., Szklarczyk D., Heller D., *et.al*. (2019) *Nucleic Acids Res.* **47**: D309.
Kanehisa M., and Goto S. (2000). *Nucleic Acids Res*. **28:** 27.
Kim C., Pongpanich M. and Porntaveetus T. (2024) *J. Transl. Med.* **22**: 111.
Marcos C.N., Carro M.D., Gutiérrez-Rivas M., *et al*. (2024) *J. Dairy. Sci.* **107**: 7064.
O'Connor E., McGovern F.M., Byrne D.T., *et al*. (2021) *J. Anim. Sci.* **99**: skab132.
Ross E.M. and Hayes B.J. (2022) *Front. Genet.* **13**: 865765.
Ross E.M., Hayes B.J., Tucker D., *et al*. (2020) *J. Anim. Sci.* **98**: 10.
Tamames J. and Puente-Sánchez F. (2018) *Front. Microbiol.* **9**: 3349.
Xie F., Jin W., Si H., *et al*. (2021) *Microbiome* **9**: 137.
Zhang Q., Difford G., Sahana G., *et al*. (2020) *ISME J*. **14**: 2019.