USE OF MID-INFRARED SPECTROSCOPY DATA FOR IMPROVING ACCURACY OF GENOMIC PREDICTION OF METHANE EMISSIONS

S. Bolormaa[1], P.N. Ho[1], I.M. MacLeod[1,2], M. Haile-Mariam[1,2], L.C. Marett[3], S.R.O. Williams[3], J.L. Jacobs[2,3,4], C.J. Vander Jagt[1], M.E. Goddard[1,4], R. Xiang[1], and J.E. Pryce[1,2]

[1] Agriculture Victoria Research, Agribio, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia
[3] Agriculture Victoria Research, Ellinbank Centre, Ellinbank, VIC, 3821 Australia
[4] School of Agriculture and Food, University of Melbourne, Parkville, VIC, 3010 Australia

SUMMARY

Worldwide, methane emissions (Me) from ruminants are driving a measurable impact on global warming. Breeding cattle with lower methane output is one option that could contribute to reducing total methane emissions. The aim of this paper was to explore the use of milk mid-infrared (MIR) data for improving the accuracy of Me genomic estimated breeding values (GEBV) for 436 Holstein cows with methane phenotypes and 50K SNP genotypes. Out of these 436 cows with methane phenotypes, 200 had MIR records. We selected the most informative MIR spectra using hierarchical clustering analysis, and subsequently used this information to predict Me phenotypes. The use of informative MIR spectra increased the prediction accuracy of methane phenotypes by up to 8% compared with using the entire 537 spectra and helped to improve the accuracy of GEBVs for methane emissions. However, this finding is based on a small cross-validation study and needs to be confirmed through a larger study.

INTRODUCTION

Methane is a short-lived, but potent greenhouse gas contributing substantially to global warming. Over 70% of agricultural greenhouse gases is from enteric methane emissions (Me) produced by ruminants. In cattle, methane is a heritable trait, ranging from 0.12 to 0.45 (Manzanilla-Pech *et al.* 2021). Thus, genomic selection for reduced methane is potentially a more reliable, cost-effective, cumulative and permanent tool for methane mitigation than some other strategies. In Australia, the Sustainability Index developed using breeding values of traits associated with methane was released by DataGene in 2021. While developing breeding values using direct measures of methane is the best strategy (Richardson *et al.* 2022), enteric methane is an expensive and labour-intensive trait to collect (e.g. SF6 or respiration chambers). Proxy methods such as mid-infrared spectroscopy (MIR) measured from milk samples potentially offers a cost-effective solution to estimate methane at scale. The objective of this study was to explore the potential of MIR for predicting methane, particularly if specific MIR wavenumbers (spectra) among the total of 537 spectra might be more informative than others, and to test their subsequent use to improve the accuracy of genomic breeding values (GEBV) for methane.

MATERIALS AND METHODS

Phenotype and genotype data. Direct methane production records (g/d per animal) were collected from 436 Australian lactating dairy cows as described in Pryce *et al.* (2015). Methane production primarily reflects animal feed intake and consequently milk production in Holsteins (Fresco *et al.* 2023), so we estimated the residual level of methane produced by each animal after accounting for their energy corrected milk production (ECM, Pryce *et al.* 2015) and refer to this as RMe, where RMe = methane – (mean + fixed effects + ECM) and fixed effects are described in Eq 1 below. We also calculated Methane intensity phenotypes (MeI, g/kg of ECM) calculated as methane production divided by ECM. For 200 of the 436 cows, MIR data from milk samples were processed following Ho *et al.* (2019). After processing, a total of 537 spectral wavenumbers between 928 and 1596 as well as between 1693 and 3025 cm^{-1} were retained for further analysis. The 50k SNP array genotypes used in this study included 41,276 SNP that passed quality control and were a subset of high-density genotypes used in Bolormaa *et al.* (2022). The phenotypes of all traits including raw MIR spectra were adjusted for fixed effects using Eq 1:

$$\text{trait} = \mu + \text{HYS} + \text{batch} + \text{parity} + \text{DIM} + \text{pol(age,-2)} \qquad \text{(Eq 1)},$$

where μ is the overall mean of phenotypes of each trait across the population, HYS relates to herd, year and season of calving, batch is 16 cohort groups from experiments. DIM is days in milk at the beginning of the experiment as a covariate, and parity (1, 2, 3, and 4+) were fitted as fixed effects, and poly(age,-2) is age of cows at calving fitted as a second-order orthogonal polynomial function.

MIR analysis. First, GEBV were calculated for each of the 537 MIR spectra using a linear regression model with the 50k SNP genomic relationship matrix using ASReml program (Gilmour *et al.* 2009). To identify the most related groups of MIR spectra, we applied the Hierarchical Clustering method using the Ward's criterion based on the multi-dimension variance (R program, R Core Team 2021). The clustering analysis was performed using R (R Core Team 2021) based on the pair-wise correlations of these GEBV. Number of clusters to be retained was determined by the Elbow method in K-means Clustering, and 11 specific spectra groups (grp) were identified at the Elbow point. Then, using partial least squares regression (PLS) analysis with 10-fold cross-validation, we tested how accurately predict the methane phenotypes using each group (or combined groups) of raw MIR spectra (mirMeI(grp)), compared to the phenotypes using entire 537 spectra (mirMeI). Four of our groups provided the R^2 of prediction of > 0.1 (range: 0.11 to 0.32, Table1), and were selected for the next part of the analysis. Table 1 shows the number of MIR spectra in each selected group. The R^2 using entire 537 spectra for mirMeI was lower compared with using grp5 or combinations with other (non-overlapping) spectra groups.

Table 1. Number of spectra and squared correlation (R^2) for MIR clustering groups

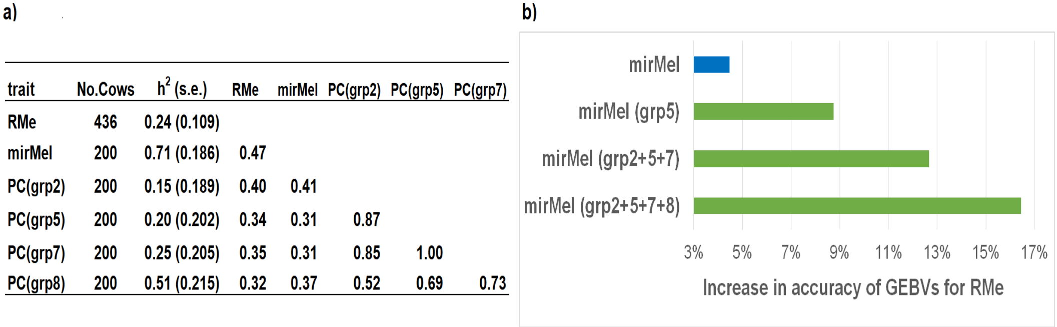
MIR clustering group	No. spectra$^\wedge$	R^{2*}
grp2	7	0.11
grp5	26	0.32
grp7	13	0.18
grp8	66	0.14
grp2+5+7$^\&$	46	0.31
grp2+5+7+8$^\&$	104	0.31
mirMeI^J	537	0.24

$^\&$represent spectra from clustering group 'grp2, grp5, grp7, and/or grp8' together; ^{J}MIR-predicted methane using full 537 spectra; ^number of spectra identified in each clustering group; $^*R^2$ is the coefficient of determination obtained from PLS regression analysis.

Genomic prediction. Next, univariate and bivariate genomic REML (GREML) analyses were performed using ASReml (Gilmour *et al.* 2009) to calculate genomic heritability (h^2) and predict Me GEBV. The genomic relationship matrix used in GREML analysis was built based on the 50k SNP genotypes using the method of Yang et al. (2010). In bivariate analyses, RMe together with MIR predicted MeI from specific MIR groups (mirMeI(grp)) were analysed by treating them as correlated traits. Genomic predictions were validated in the 200 MIR cows using a leave-one (cow)-out approach. The prediction accuracy for RMe was calculated as the correlation between GEBV and the RMe, divided by the square root of the heritability (h^2). The accuracies using bivariate models were compared to the accuracy of the univariate model. We also considered a scenario where the direct MIR spectra were included in a bivariate model. For this purpose, using those spectra allocated to each clustered group in Table 1, we performed principal component (PC) analyses, and then the 1st PC for each group was used as the second trait (MIR PC) in the bivariate model.

RESULTS AND DISCUSSION

Mean (SD) of RMe was 47.6 (11.13) g/d. Genomic heritability estimate using 436 cows in univariate analysis was 0.24 for RMe, which is within the reported range in literature (Richardson *et al*. 2021; Manzanilla-Pech *et al*. 2021). The MIR PC traits provided a range of h^2 estimates (0.15-0.51, Figure 2a), but with large standard errors (range: 0.173-0.215) due to the small dataset. The heritability for MeI predicted using all 537 MIR spectra (mirMeI) was high (0.71), probably indicating that mirMeI additionally captures the variance explained by milk production.



a)

trait	No.Cows	h^2 (s.e.)	RMe	mirMeI	PC(grp2)	PC(grp5)	PC(grp7)
RMe	436	0.24 (0.109)					
mirMeI	200	0.71 (0.186)	0.47				
PC(grp2)	200	0.15 (0.189)	0.40	0.41			
PC(grp5)	200	0.20 (0.202)	0.34	0.31	0.87		
PC(grp7)	200	0.25 (0.205)	0.35	0.31	0.85	1.00	
PC(grp8)	200	0.51 (0.215)	0.32	0.37	0.52	0.69	0.73

b)

Figure 2. a) Heritability estimates and correlations of GEBV between directly measured methane and MIR predicted methane (mirMeI) or MIR principal component (PC(grp2 to grp8)) traits; b) Increase in accuracies of GEBV for residual methane (RMe)* in bi-variate analyses

*blue bar represents using phenotypes for predicted methane intensity (mirMeI) based on entire 537 MIR spectra and green for MeI predicted based on MIR spectra specific to the clustering group(s): 'grp(2+5+7)' or 'grp(2+5+7+8)' represent spectra from grp2, grp5, grp7, and/or grp8 together.

Since our primary objective is to use more informative spectra groups that may help improve the accuracy of GEBV for methane traits, we are only interested in MIR spectra groups (MIR PC traits) that have the apparent correlations of GEBV ($r > 0.2$) with GEBV for methane traits. All four MIR PC traits shown in Figure 2a exhibited moderate relationships with directly measured and MIR predicted methane traits (ranging from 0.32 to 0.40, Figure 2a). We found that the spectra of these 4 MIR PC traits fall within the infrared spectral regions driven largely by C-O, C=O, C-H, and C-N, which are particles of or bonds to fat (volatile and fatty acids) and protein. This was further confirmed on finding that these MIR PC traits exhibited moderately strong correlations with milk fat percentages (FAT%, r = 0.44-0.53), perhaps indicating that MIR PC traits in these specific groups capture some of the genetic variance of FAT%. Manzanilla-Pech et al (2021) reported moderate genetic correlation (0.45) between direct methane and ECM. As expected, strong correlation of GEBV (r=0.5) was observed between mirMeI and FAT%, whereas the correlation of FAT% with RMe was lower (r=0.26).

To compare accuracy of the GEBV from the univariate GBLUP for RMe (independent from ECM) with the accuracies from the bivariate GBLUP analysis, in addition to RMe, we used either MIR predicted MeI traits (with R^2 of ≥0.30 from PLS) or individual MIR PC traits as the second trait. The accuracy of GEBV for RMe was 0.41 in our univariate model. The accuracy of GEBV for RMe was improved 4% by fitting all 537 MIR spectra predicted methane phenotypes (mirMeI) as a second trait in the bivariate model (Figure 2b). An even greater increase in the genomic accuracies for RMe was observed using only the MIR spectra groups rather than using all 537 spectra as a second correlated trait (e.g. up to 16% using 'grp2+5+7+8', Figure 2b). When we compared use of phenotypes based either on same number of spectra generated from PLS or PC approaches, genomic accuracy using the MIR grp5 predicted MeI from PLS was higher (+5%) compared with using the MIR PC phenotypes (results not shown). This shows an advantage of predicted methane phenotypes with acceptable R^2 of ~0.30 from PLS over the PC1 phenotypes from PC analysis. The MIR PC1 phenotypes from each spectra group also helped to increase the accuracy of directly measured RMe due to its correlation with methane. However, it is worth testing if its increase in genomic accuracy of RMe is greater than FAT% with a larger methane dataset. The findings in this study generally showed the potential of the use of MIR as a proxy for improving the genomic accuracy of RMe, but a larger dataset is required to better evaluate this.

CONCLUSION

Identifying informative MIR spectra through hierarchical clustering analysis helped increase the prediction accuracy of methane phenotypes by 8% and subsequently improved the accuracy of GEBV for methane by up to 16%. The phenotypes predicted using PLS approach provided greater genomic prediction accuracy than PC generated phenotypes. However, this study should be interpreted with caution due to the small size of reference and validation sets used in this study and requires confirmation using a larger population.

ACKNOWLEDGEMNTS

We thank DairyBio, jointly funded by Dairy Australia (Melbourne, Australia), The Gardiner Foundation (Melbourne, Australia) and Agriculture Victoria (Melbourne, Australia) for funding this project. We acknowledge use of data from the Agriculture Victoria Ellinbank SmartFarm.

REFERENCES

Bolormaa, S., MacLeod I.M., Khansefid M., Marett L.C., Wales W.J., Miglior F., Baes C.F., Schenkel F.S., Connor E.E., Manzanilla-Pech C.I.V., Stothard P., Herman E., Nieuwhof G.J., Goddard M.E. and Pryce J.E. (2022) *Gen. Sel. Evol*. **541**: 60.

Fresco, S., Boichard D., Fritz S., Lefebvre R., Barbey S., Gaborit M. and Martin P. (2023) *J. Dairy Sci.* **106**: 4147.

Gilmour, A.R., Gogel B.J., Cullis B.R. and Thompson R. (2009) ASReml User Guide Release 3.0. VSN Hemel Hempstead, UK.

Ho, P. N., Bonfatti V., Luke T.D.W. and Pryce J.E. (2019) *J. Dairy Sci.* **102**: 10460.

Manzanilla-Pech, C.I.V., Løvendahl P., Mansan Gordo D., Difford G.F., Pryce J.E., Schenkel F., Wegmann S., Miglior F., Chud T.C., Moate P.J., Williams S.R.O., Richardson C.M., Stothard P. and Lassen J. (2021) *J. Dairy Sci.* **104**: 8983.

Pryce, J.E., Gonzalez-Recio O., Nieuwhof G., Wales W.J., Coffey M.P., Hayes B.J. and Goddard M. E. (2015) *J. Dairy Sci.* **98**: 7340.

Richardson, C.M., Amer P.R., Quinton C., Crowley J., Hely F.S., van den Berg I. and Pryce J. E. (2022) *J Dairy Sci.* **105**: 4272.

Richardson, C.M., Nguyen T.T.T., Abdelsayed M., Moate P. J., Williams S.R.O., Chud T. C. S., Schenkel F.S., Goddard M.E., van den Berg I., Cocks B.G., Marett L.C., Wales W.J. and Pryce J.E. (2021) *J. Dairy Sci.* **104**: 539.

Yang J., Benyamin B., McEvoy N.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. and Visscher P.M. (2010) *Nat. Genet*. **42**: 565.