STSM report

Grantee: Gareth Difford

COST STSM Reference Number: COST-STSM-FA1302-33935

Home institution: Centre for Quantitative Genetics and Genomics, Aarhus University, Denmark

Host Institution: Livestock Research, Wageningen University, The Netherlands

Grant Period: 2016-05-09 to 2016-05-20

1. Purpose of STSM

Currently, there is a need to provide a proof of concept regarding genetic variation of methane emissions in ruminants and the associated genetic correlations with existing traits. Recording methane emission in dairy cattle is not without some controversy, as respiration chambers (RC) widely regarded as the benchmark are costly and alter the underlying biology of the cow. Short-term methods (of which there are many) are cheaper and commercially viable but suffer from lack of precision. Furthermore, validating short term methods is not possible without genetic correlations to RC records exceeding (0.8) (Robertson, 1980), which in themselves are not possible as RC records are too costly to accumulate on the scale necessary for meaningful genetic correlations. Thus validating short-term methane emission recording techniques is a circular argument. This poses a bottle-neck to the investment and development of the sorely needed genetic improvement of methane emission in dairy cattle.

However, combining records from multiple countries across multiple instruments and production environments with the aim of genetic/genomic estimated breeding values poses an opportunity to validate methods in a different manner. If short-term estimation methods from multiple instruments and countries are capturing the same underlying genetic variation, this in itself is preliminary proof of genetic control of methane emission in dairy cattle. Furthermore, identifying a subset of individuals with the highest and lowest breeding values for methane emission could prove useful in validating methods through re-recording in RC and testing for significant differences under identical environments. Furthermore, gene mapping and genomic prediction will also provide an indirect validation method, if SNP's with a large effect are identified, genes in LD with SNPs can be identified and if their known function associates with the underlying biology of methane emission, this would supply further evidence.

To this end data from 5 countries and 3 methods namely, Australia, Republic of Ireland (SF6 method), Denmark, The Netherlands (Sniffer) and United Kingdom (hand held laser - LMD); were collated within the gCH4 consortium. As part of the specific objectives of the consortium investigations into the genetic variability, genetic correlations and genomic tools for predicting breeding values are to be conducted. The primary purpose of this STSM was to conduct preliminary studies into methods of best modelling such data (fixed effect models) and constructing Variance-Covariance matrices for partitioning variance into genetic components.

2. Activities during STSM

Specific activities of this STSM were to obtain consensus on an optimum fixed effect model (without over parameterisation) to be used in achieving the objectives of the aforementioned consortium. Examining if there is population admixture or cryptic familial structure within populations, which must be taken account of for genomic association studies. Furthermore, since multiple methods are used to obtain phenotypes the question then arose "Is it more informative to evaluate the records of each method in the actual units of each respective method?"

3. Main results

Modelling of effects

Since data from all of the countries except Denmark was from research herds, it is not surprising to find numerous feeding trials, contemporary groups or genetic lines within each population. Fixed effects included in models were done so on the basis of exceeding the critical value of Wald's f-tests conditional on their respective numerator and denominator degrees of freedom. Nested models were evaluated for goodness of fit by means of log likelihood ratio tests and non-nested models were evaluated by means of minimising the Bayesian information criterion (BIC) as described in equation 1 below:

$$BIC = -2 * loglikelihood + t * \log(n - p) \dots eq(1)$$

Where the log likelihood is estimated by the REML function of ASReml 3, t is the number of variance parameters, n is the number of observations and p is the combined number of levels of all fixed regressors and classes, i.e n –p is the residual degrees of freedom. The models evaluated are as follows, with random effects denoted in italics and nested effects within parentheses:

$$Model\ 1:$$

$$y = \mu + \text{age at calving} + \text{age at calving}^2 + \text{parity} + \%\text{Holstein} + \text{Month Milk} + \text{Month recording} + \text{Country year herd} + \text{experimental level} + \text{Country year herd} * \text{experimental level} + Animal + e$$

$$Model\ 2:$$

$$y = \mu + \text{age at calving} + \text{age at calving}^2 + \text{parity} + \%\text{Holstein} + \text{Month Milk} + \text{Month recording} + \text{Country year herd (experimental level)} + Animal + e$$

$$Model\ 3:$$

$$y = \mu + \text{age at calving} + \text{age at calving}^2 + \text{parity} + \%\text{Holstein} + \text{Month Milk} + \text{Month recording} + \text{Country year herd} + Country\ year\ herd\ (experimental\ level) + Animal + e$$

Methane in grams per day is denoted by y. The fixed effects : country year herd (27 classes), experimental level (41 classes), the experimental level nested within country-year-herd (32 classes), age at calving as a

second order polynomial, fixed regression of month in milk (1-12), fixed regression of month of recording (Jan-Dec), and fixed regression on percentage of Holstein (available for Irish cows only).

Table 1 Results of model evaluation for by means of log likelihood ratio tests or BIC

Models	Residual D.F	Phenotypic variance	Log likelihood	Bayes Information Criterion	Log likelihood ratio test (*p value*)
1	2737	0.8667321	-1353.00	2716.31	N/A
2	2769	0.785863	-1201.47	2413.27	N/A
3	2827	1.079038	-1351.84	N/A	0.128

From table 1 above it can be seen that model 2 as compared to model 1 by means of BIC is a better fit and model 3 as compared to model 1 by means of log likelihood ratio tests is not a significant improvement of fit. It is pertinent to note in model 1 the residual degrees of freedom do not reflect a full interaction term for country-year-herd by experimental level even though this is the format supplied to ASReml 3 and described for model 1 above. This is because even though the nested structure is hidden from ASReml 3, it still detects experimental level is nested within country-year-herd and still produces valid conditional Wald statistics thus the deviations in residual degrees of freedom are due including the main effect of experimental level.

Genomic Data

Of the 2857 cows in the combined dataset, 2543 were genotyped with a mixture of commercial and customized SNP chips. The genotypes of all countries were combined by the matching of SNP names. This resulted in ~44 000 SNPs in common after filtering for deviations from Hardy-Weinberg equilibrium proportions ($p<0.001$) and minor allele frequencies (MAF) ($p<0.005$). Although multiple Irish animals were of mixed breed origin these animals were not included in the genotyped population. A principle component analysis of the resulting SNP marker matrix and a plot of the first two principle components can be found in figure 1.

As can be seen in figure, there is no isolated groupings by country. However, it can also be seen that there is cryptic familial structure within and between populations. Thus it is necessary to include the principle components in any mapping endeavours which may follow.

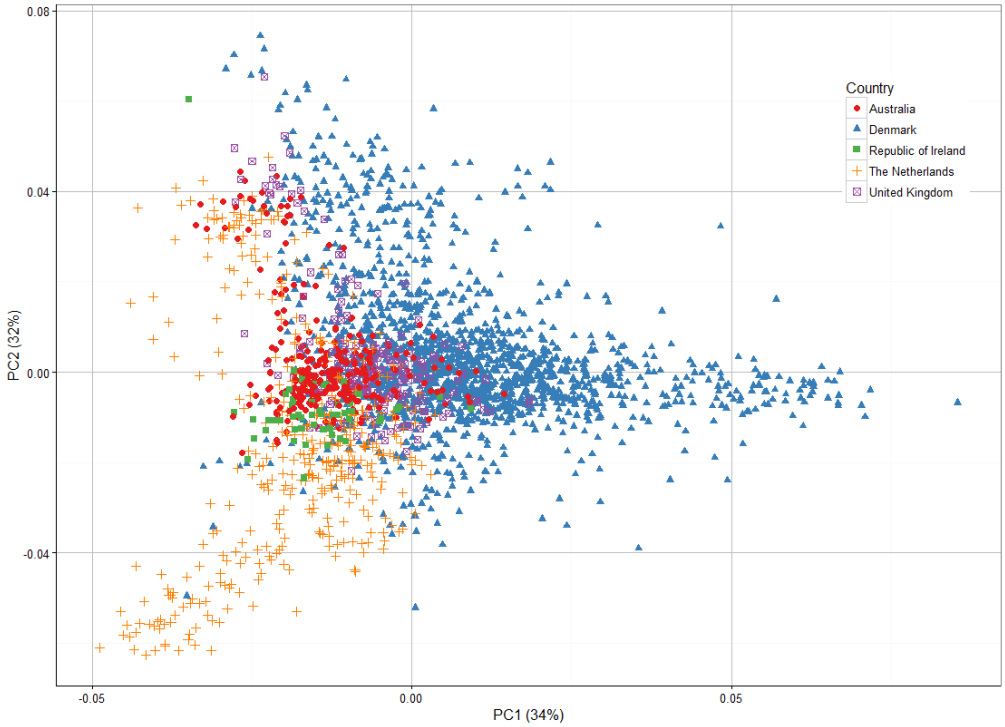

Figure 1 Plot of first two principle components of SNP matrix with countries denoted by different shapes and colours.

Methane Phenotypes and Original Units

In order to evaluate whether it is possible or feasible to evaluate the different methane phenotypes in the units in which they are originally recorded, pairwise bivariate models in the form of model 2 above were run in ASReml 3 for the entire dataset with the pedigree relationship matrix (A^{-1}) for the following phenotypes:

Methane in grams per day (CH4g/day), methane in ppm (CH4ppm), ratio of methane in ppm to carbon dioxide in ppm (Ratio), methane yield per kg milk (CH4Yield), live weight in kg (LW), fat and protein corrected milk yield (FPCM) and methane in the original units they are recorded (ppm/m UK, ppm Denmark & The Netherlands and grams/day Republic of Irelands and Australia) (CH4mix).

Singular value decomposition was performed on the resulting 7 x 7 genetic correlation matrix and the vector trajectories of each of the traits plotted within principle component space in figure 2 below. Note the cosine of the angle between two vectors reflects the genetic correlation between traits. For example an angle = 90° reflects a correlation of 0 and an angle of 0° reflects a genetic correlation of 1.

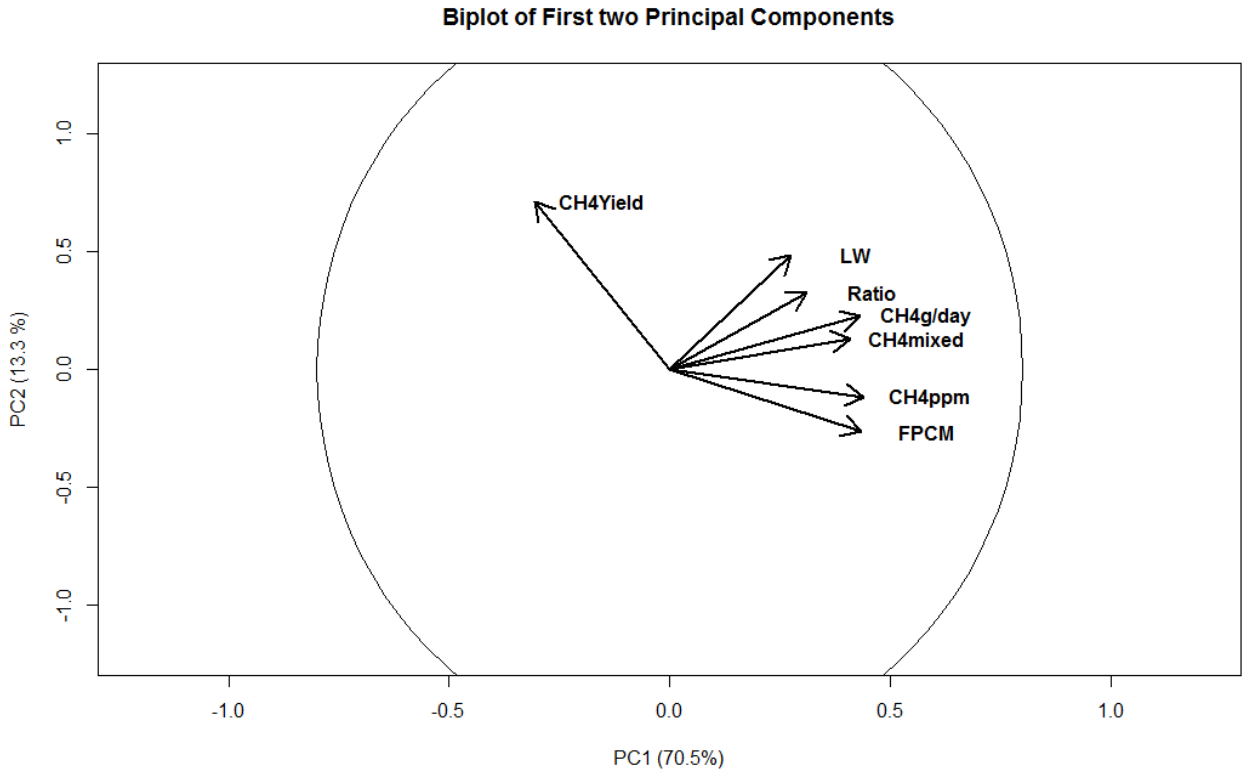

Figure 2 A biplot of vector trajectories for traits of interest in principal component 1 and 2 space corresponding to 83.8% of genetic variation.

Of interest in the plot above is that CH4 mixed falls between CH4g/day and CH4ppm which is expected since CH4 mix is comprised of the two aforementioned variables. It would appear evaluating populations within the original units they are recorded in does not alter the genetic correlations with production traits FPCM and LW included here. However, the correlations above did contain standard errors between 0.07 – 0.21. Thus it would be advisable to re-evaluate when more data becomes available. The genetic correlations between CH4mixed and other traits of interest are tabulated in table 2 below:

Table 2 Pairwise genetic correlations and standard errors thereof between methane traits, fat- and- protein corrected milk and body weight

	CH4g/day	CH4Yield	CH4ppm	Ratio	FPCM	LW
CH4 mix	0.89 (0.11)	-0.42 (0.22)	0.90 (0.07)	0.49 (0.22)	0.79 (0.12)	0.60 (0.11)

4. Conclusion

The results presented here are preliminary and should be treated as such. Of the models evaluated here model 2 which nests experimental level within country-herd-year appears to be the most suitable model and the least over parameterised. From a genotypic perspective the populations included here are not suitably differentiated to preclude the use of genomic relationship matrices (G^{-1}) in further analyses. However, association studies must take cognisance of cryptic familial structure to prevent spurious

associations. It may be possible to evaluate data across countries with the methane recorded in original units of each instrument/technique although whether it is advisable to do so requires further investigations.

5. Future Collaboration with Host Institution

Ongoing collaborations will persist between all institutions who have contributed to the dataset analysed herein. Regular skype meetings have been planned and work partitioned between researchers.

6. Foreseen publications

Investigations into genetic parameters, genomic associations and genomic prediction will proceed, all of which may results in publications.

7. Comments

Please note the results contained herein are preliminary analyses conducted during a two week period and have not been subjected yet to peer review.

8. Literature

Robertson, A., 1959. The sampling variance of the genetic correlation coefficient. *Biometrics*, *15*(3), pp.469-485.