Bayesian genomic analysis of methane emissions in dairy cattle

F. Biscarini (PTP Science Park)

October 3, 2016

COST Action FA1302

Host institution: Department of Animal Breeding and Genetics - IRTA (Institut de Recerca i Tecnologia Agroalimentàries), Caldes de Montbui (Spain)

Period: 18/09/2016 to 01/10/2016

Reference code: COST-STSM-FA1302-34288

Purpose of the STSM

The objective of the STSM was to apply a Bayesian approach to the analysis of methane emission data in dairy cattle. The group of Dr. Juan Pablo Sánchez at IRTA has extensive experience in the application of Bayesian statistical methods to problems related to animal breeding and genetics/genomics.

The work was organised in three major sub-topics:

1. Estimation of variance components and genetic parameters for milk, methane and feed intake
2. Genomic predictions for milk, methane and feed intake from SNP genotypes
3. Comparison between traditional and Bayesian GWAS (genome-wide association study)

Material

The available data comprised a population of **769 Holstein cows** from 2 countries: Italy and the United Kingdom. All cows were genotyped with the *GeneSeek Genomic Profiler HD v.2* chip (139 480 SNPs), and had phenotypic data for **milk yield** (MY, kg/d), **dry matter intake** (DMI, kg/d) and **methane production** (MP, g/d). Methane was recorded either with GreenFeed® or through the milking robot.

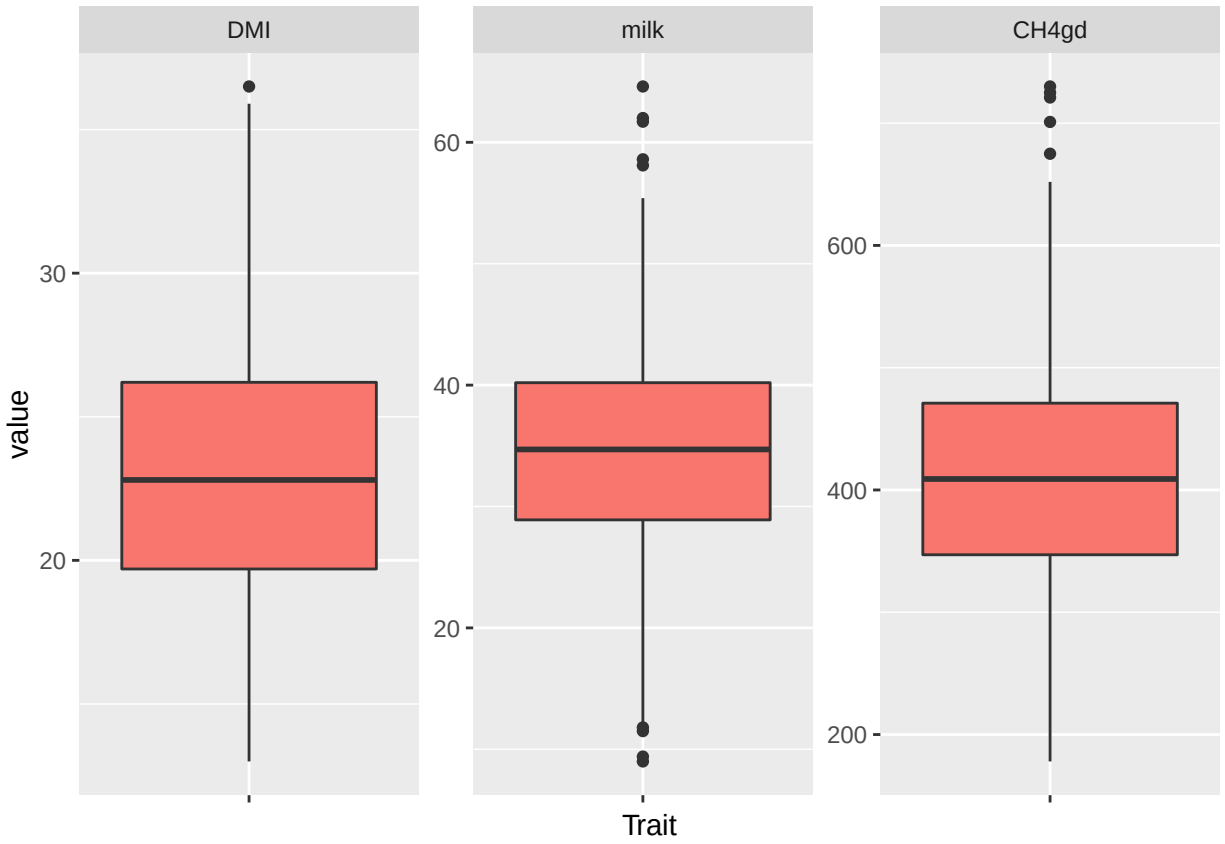
Animals and SNP loci with call-rate $< 90\%$ were removed. Remaining missing SNP genotypes were imputed using the localized haplotype clustering algorithm implementd in *Beagle v4.0*, through the open-source pipeline "Zanardi" (Marras et al. 2016).

Finally, SNP markers with MAF (minor allele frequency) $< 1\%$ were filtered out.

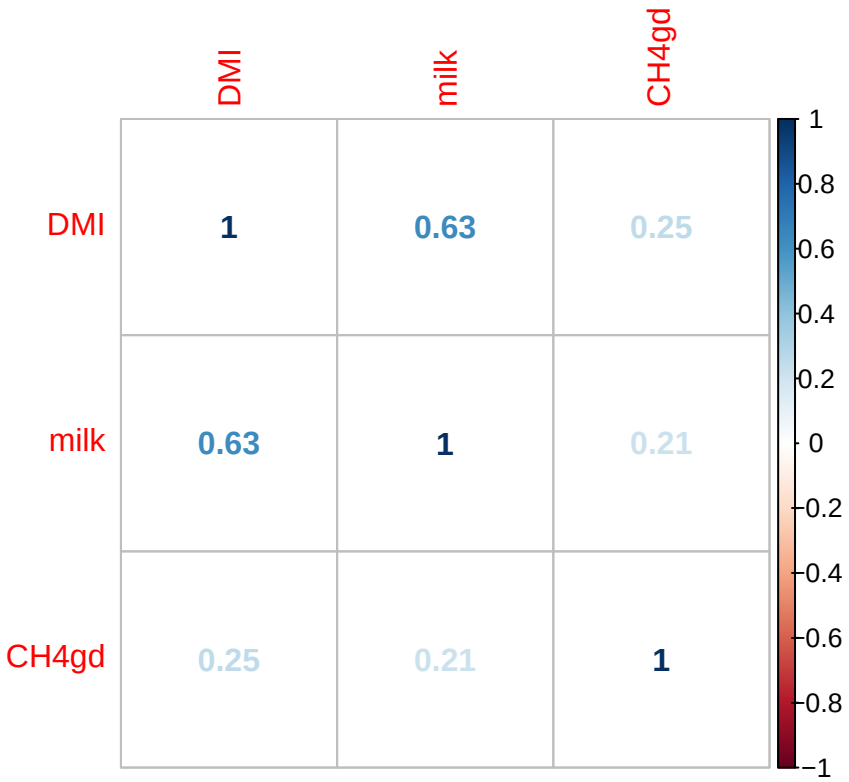
All this left 114 452 SNP to be used for subsequent analyses.

Below the summary statistics and boxplots of distributions of the three phenotypes:

variable	mean	stdev	coefVar	min	max
DMI	23.01	4.63	0.20	13	36.5
milk	34.45	8.74	0.25	9	64.6
CH4gd	410.55	91.74	0.22	178	730.0



The phenotypic correlations between traits were all positive: $r_{DMI,MY}$0.63 between **DMI** and **MY**, **0.25** between **DMI** and **MP**, and 0.21 between MY and **MP** (see Table below).



Estimation of variance components and genetic parameters

$y_{i(j)kz} = \mu + country_i + HSP_{j(i)} + DIM_k + DIM_k^2 + cow_z + e_{i(j)kz}$ (1)

where:

- $y_{i(j)kz}$ is the phenotype of animal z (DMI, Milk Yield, CH4)
- μ is the overall mean
- $country_i$ is the systematic effect of country (Italy/UK)
- $HSP_{j(i)}$ is the systematic effect of herd, season of measurement and parity, nested within country
- DIM_k, DIM_k^2 are the effects of days in milk and days in milk squared (covariables)
- cow_z is the genetic (animal) effect
- $e_{i(j)kz}$ are the residuals

The genetic and residual variances were $Var(cow) = G\sigma_a^2$ and $Var(e) = I\sigma_e^2$, where **G** is a covariance matrix of genomic relationships, and **I** is an identity matrix. Genomic relationships were estimated "à la Van Raden" (Van Raden 2008).

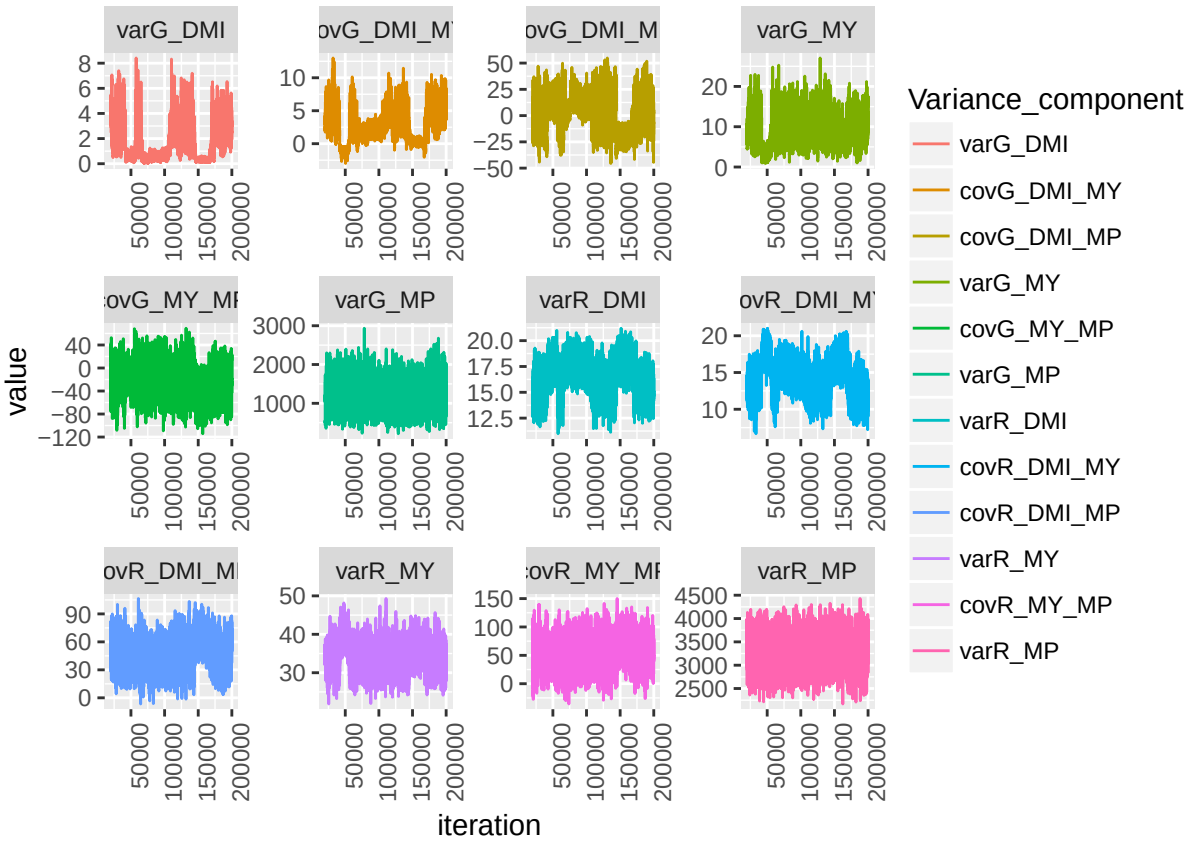
Model (1) was solved both in a **Bayesian framework** using a **Gibbs Sampling MCMC** algorithm, and with **restricted maximum likelihood** (**REML**) from a **frequentist perspective**. This was done to compare results obtained from the two different approaches.

Gibbs sampling and REML were implemented, respectively, with the Fortran computer programs **GIBBS2F90** and **REMLF90** (http://nce.ads.uga.edu/software/; Misztal 2008).

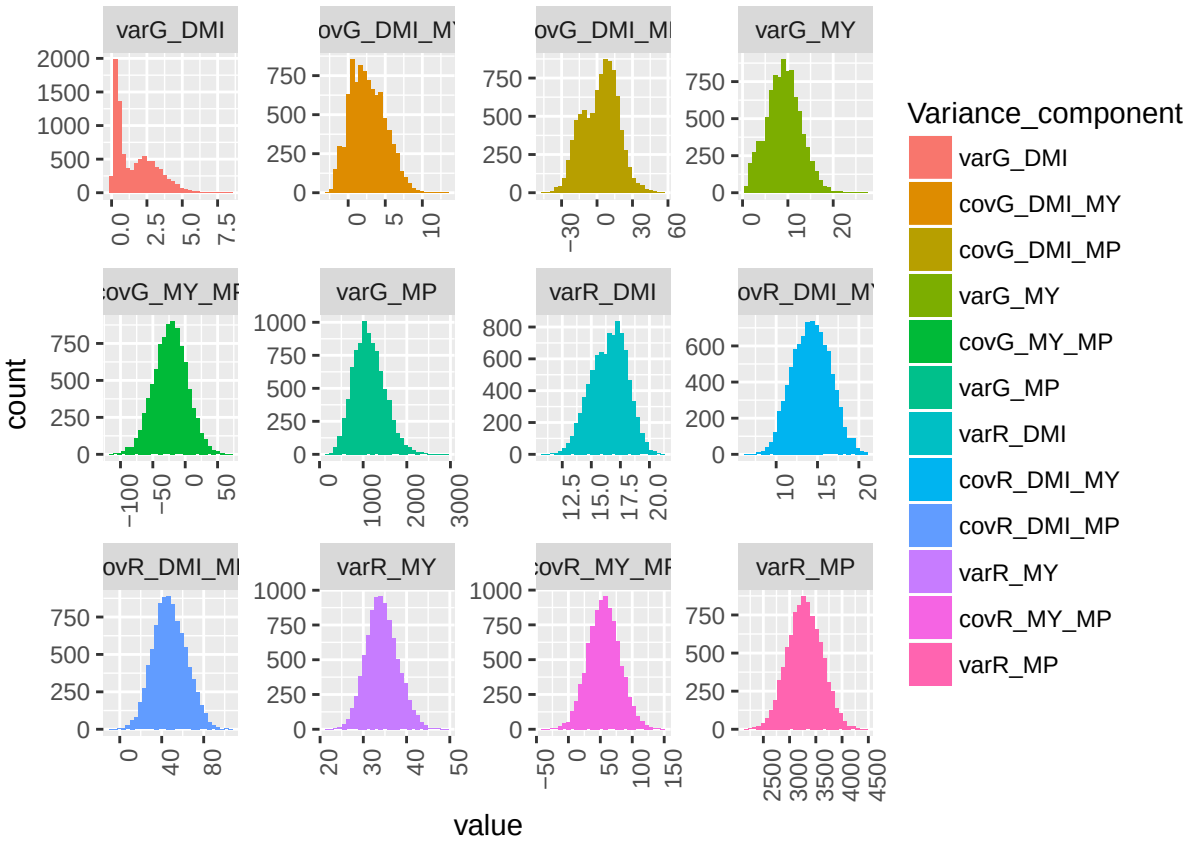
Results from Bayesian estimation

For Gibbs Sampling, 200 000 iterations were run, with a burn-in period of 20 000 iterations (discarded), and a thinning of 20 (every 20_{th} iteration was saved).

Estimates of parameters along iterations (9000 iterations after burn-in and thinning) are plotted below.



The **posterior distributions** of **genetic parameters** are also plotted:



From the the marginal posterior distributions of (co)variance components in the model, **posterior means** and **standard deviations** were obtained: these are given in the table below.

Genetic parameters (**heritabilities** and **correlations**) were computed from ratios of variance components at each iteration, in order to obtain posterior distributions of derived parameters.

```
kable(D)
```

Variance_component	media	stDev
varG_DMI	1.68	1.41
covG_DMI_MY	2.65	2.31
covG_DMI_MP	2.69	14.90
varG_MY	9.13	3.70
covG_MY_MP	-22.75	25.78
varG_MP	1135.03	352.11
varR_DMI	16.52	1.53
covR_DMI_MY	14.26	2.27
covR_DMI_MP	47.68	15.65
varR_MY	34.34	3.62
covR_MY_MP	55.77	24.61
varR_MP	3286.00	324.26
h2_DMI	0.09	0.08
h2_MY	0.21	0.08
h2_MP	0.26	0.07
corrG_DMI_MY	0.58	0.51
corrG_DMI_MP	0.04	0.50
corrG_MY_MP	-0.24	0.27

Variance_component	media	stDev
corrR_DMI_MY	0.60	0.06
corrR_DMI_MP	0.20	0.07
corrR_MY_MP	0.17	0.07

A 3x3 matrix with h^2 (diagonal), **genetic correlations** (above diagonal) and **residual correlations** (below diagonal) is also reported:

	DMI	MY	MP
DMI	0.09	0.58	0.04
MY	0.60	0.21	-0.24
MP	0.20	0.17	0.26

Results from REML estimation

Genetic parameters were estimated also with REML. **Heritabilities** (diagonal), **genetic correlations** (above diagonal) and **residual correlations** (below diagonal) are reported in the following table:

	DMI	MY	MP
DMI	0.14	0.68	-0.01
MY	0.59	0.22	-0.22
MP	0.21	0.16	0.20

Summary

A low-to-moderate h^2 was estimated for DMI ($9-14\%$); moderate heritabilities were estimated for MY and MP (0.22, 0.20--0.26). Genetic correlation was high between DMI and MY ($0.58-0.68$), moderately negative between MY and MP ($-0.22--0.24$) and absent (0.04 – -0.01) between DMI and MP. The negative genetic correlation between MY and MP (with moderately positive residual correlation) can be interpreted as follows: the more milk a cow produces, the more ruminal methane she releases. This (moderately) positive phenotypic correlation has one residual/environmental component which is also positive (0.16 – 0.17), indicating that "environmental" factors (e.g. feeding) increase both MY and MP. However, the genetic correlation is negative, which means that top-producing cows are metabolically more efficient and direct most of the energy towards milk production rather than methane production.

Genomic predictions for milk, methane and feed intake

For genomic predictions of MY, DMI and MP, the same model as for variance components estimation (model (1)) was used. The model was run in a **Bayesian MCMC setting** using the software GIBBS2F90. A 10-fold cross-validation scheme was applied to measure the accuracy (predictive ability) of genomic predictions. Random sampling of observations into training and testing sets was performed within *Country* and *HSP* groups, in order to ensure estimability of effects in the training set.

For each effect, mean values from posterior distributions were obtained and used as estimates of effects. Estimated effects were then used to **predict phenotypes** -for each trait- in the **testing set**:

$$\hat{y}_{i(j)kz} = \mu + country_i + HSP_{j(i)} + DIM_k + DIM_k^2 + cow_z$$

Predicted phenotypes were compared with observed phenotypes to estimate the accuracy of genomic predictions.

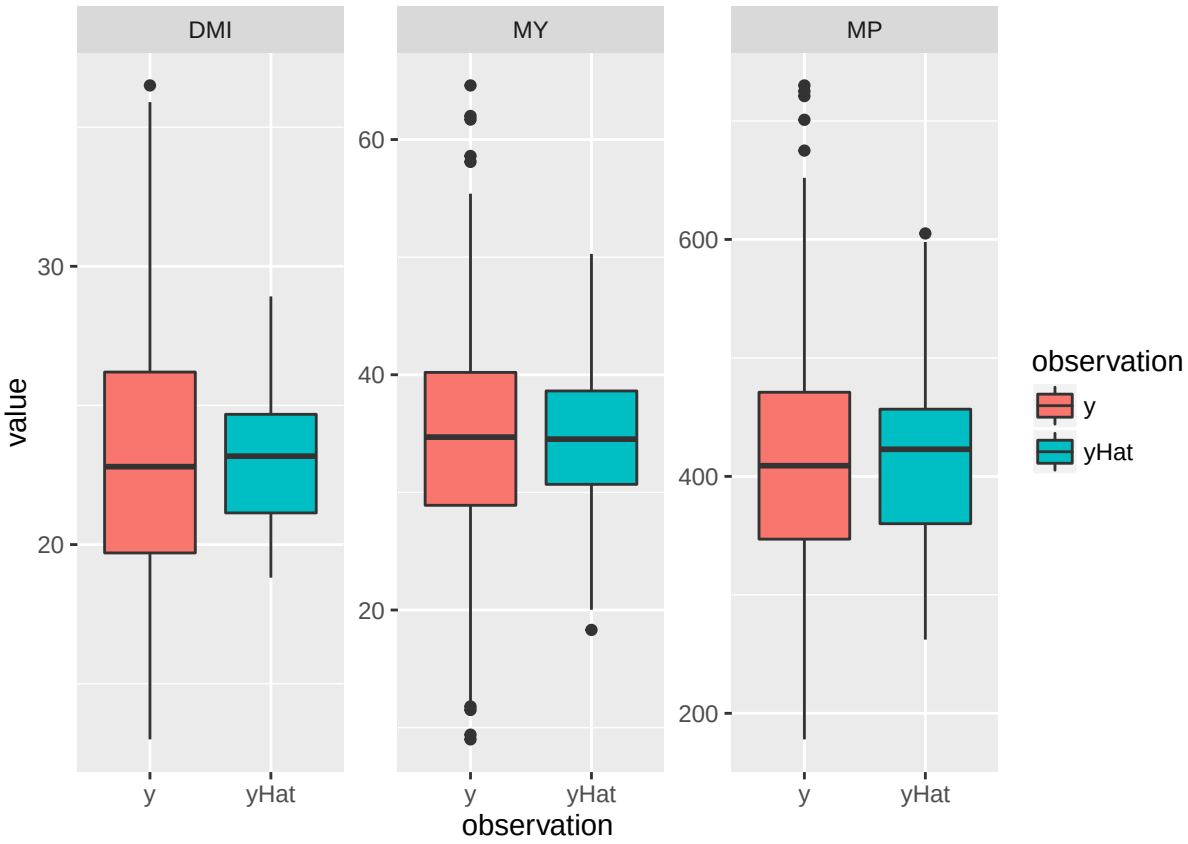
Here the Pearson correlations per trait for each of the 10 cross-validation folds.

trait	fold	correlation
DMI	fold1	0.3212877
DMI	fold10	0.3604204
DMI	fold2	0.4821891
DMI	fold3	0.3452766
DMI	fold4	0.5321532
DMI	fold5	0.2831555
DMI	fold6	0.2794844
DMI	fold7	0.4494551
DMI	fold8	0.2848989
DMI	fold9	0.2858069
MY	fold1	0.8019948
MY	fold10	0.6578909
MY	fold2	0.6326623
MY	fold3	0.6762054
MY	fold4	0.7440760
MY	fold5	0.6005653
MY	fold6	0.6487713
MY	fold7	0.7032273
MY	fold8	0.6496951
MY	fold9	0.5071688
MP	fold1	0.5363619
MP	fold10	0.6678241
MP	fold2	0.7561295
MP	fold3	0.7572692
MP	fold4	0.7239985
MP	fold5	0.7219364
MP	fold6	0.7548469
MP	fold7	0.6520170
MP	fold8	0.6974573
MP	fold9	0.6747242

For the three traits, **correlations** (both **Pearson** and **Spearman**) between predicted and observed phenotypes, **Root Mean Squared Error** and **normalized Root Mean Squared Error** are reported below:

trait	r_pearson	r_spearman	RMSE	normalizedRMSE
DMI	0.357	0.371	4.345	0.939
MY	0.660	0.649	6.567	0.751
MP	0.692	0.671	66.331	0.723

The distributions of predictions are plotted against the distributions of observed values:



Bayesian GWAS

Traditional GWAS analyses are very popular for identifying genetic polymorphisms associated to specific phenotypes. Traditional GWAS is usually carried out by running a model for each SNP separately and, although it is a quite effective approach, is known to suffer from some limitations. On one hand, there is the issue of multiple-testing: several thousands of SNPs are tested, and this increases the probability of false positive results. Additionally, except for clear signals of associations, results are often of ambiguous interpretation. Therefore, alternative or complemetary approaches to GWAS may be useful to obtain more robust and convicing results (see for instance Biscarini et al. 2015).

Bayesian models may offer another way to perform association studies (e.g. Stephens and Balding, 2009). From a **Bayesian solution of a G-BLUP model** (like that in Equation (1)), SNP effects can be obtained as functions of the estimated animal genetic values at each iteration of the Gibbs Sampling MCMC algorithm, by employing the equivalence between the **G-BLUP** and **SNP(RR)-BLUP** models:

$$\beta = X'(XX')^{-1}\hat{g}(2)$$

From the SNP effects thus obtained, in a Bayesian context, different metrics (e.g. variance explained by the SNP, probability that such variance is greater than a threshold etc ...) can be used to detect phenotype-genotype associations.

Bayesian models can potentially avoid problems related to multiple-testing (all SNP are analysed together, and marginal posterior distributions of SNP effects are obtained, conditional on all other effects in the model). Additionally, Bayesian models offer greater flexibility since a large array of metrics of interest can be derived from posterior distributions.

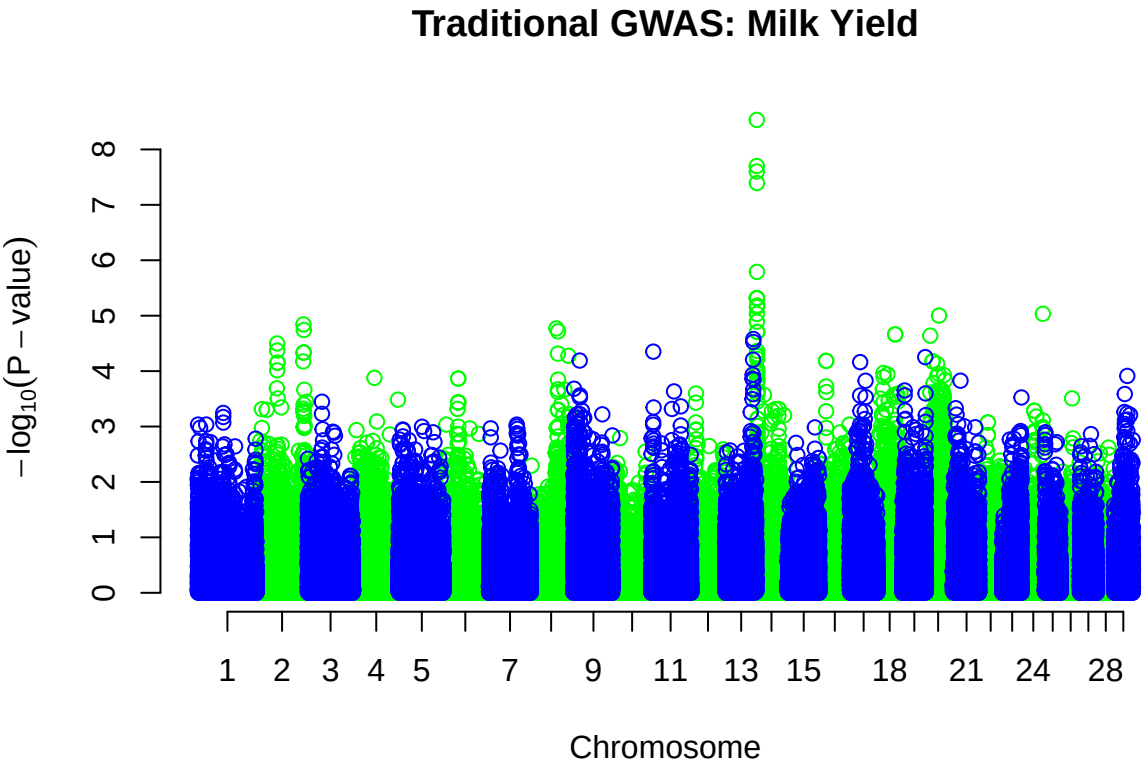
The Bayesian GWAS is compared to a traditional GWAS, by applying both approaches to Milk Yield, for which a known association on BTA14 is expected (the *DGAT1* gene).

Traditional GWAS

A linear model of the following form was run:

$$\mathbf{y} = \mathbf{1}\mu + X\beta + Z \cdot u + W \cdot SNP + e$$

- $\mathbf{y}$ was a vector of milk yield kg (day of sampling) or methane g/d (methane production)
- μ is the overall mean
- $X\beta$ is the term related to the systematic effects
- $Z \cdot u$ is the term related to the polygenic effect
- $W \cdot SNP$ is the term related to the SNP effect



Bayesian GWAS

From Bayesian G-BLUP, SNP effects were derived from animal genetic effects as in Equation (2). SNP effects were grouped in sliding windows of 10 SNPs, and from their distribution over the n MCMC iterations two metrics were chosen to detect potential associations:

1. the proportion of the variance explained by the SNPs in the window;
2. the probability that such proportion was larger than 0.1%

Results are shown in the plots below.

Conclusions

All three objectives of the STSM were carried out successfully. Variance components and genetic parameters for milk yield (MY), methane emissions (MP) and dry matter intake (DMI) were estimated using a G-BLUP

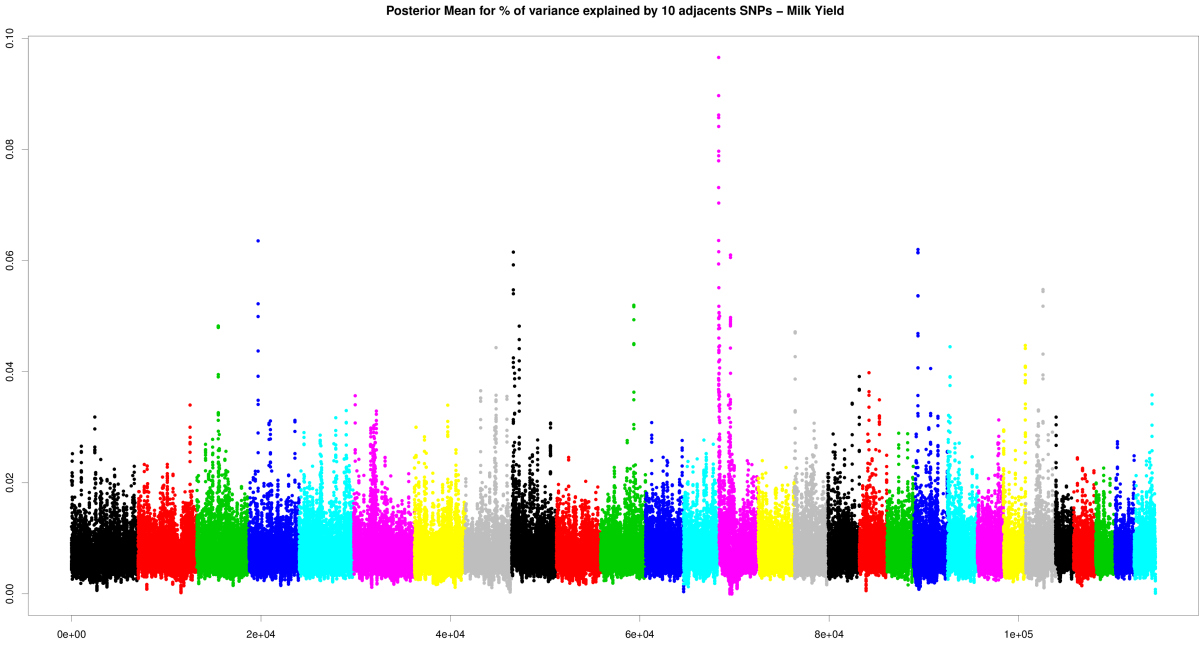


Figure 1:

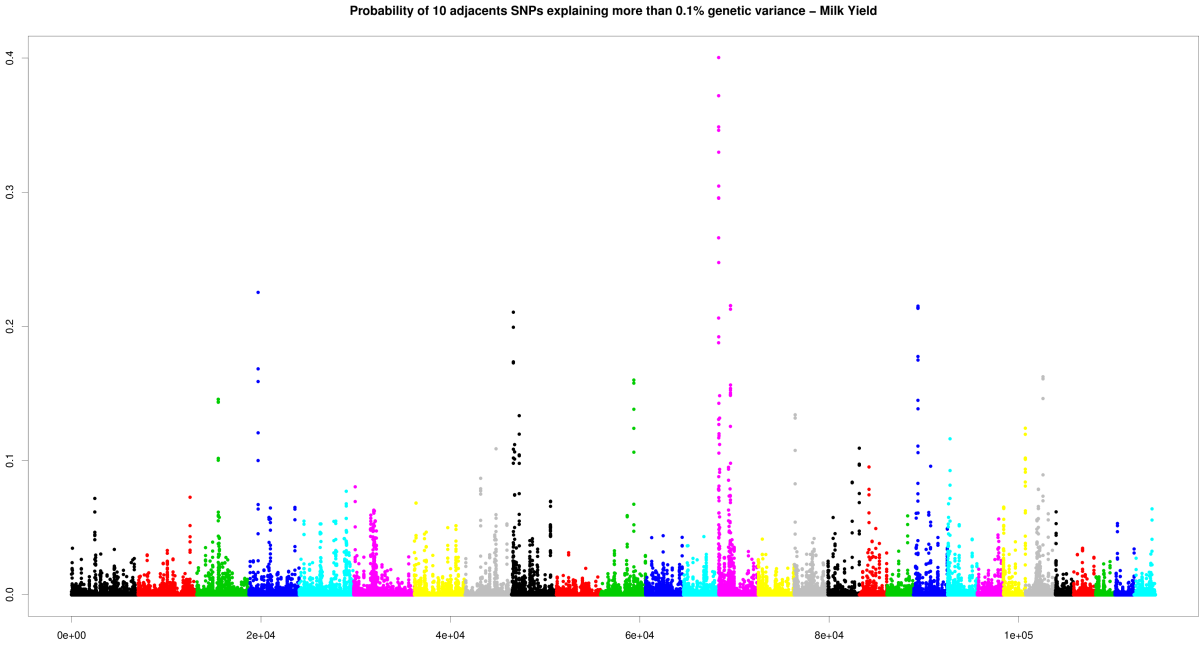


Figure 2:

model solved through a Gibbs Sampling MCMC algorithm. The Bayesian approach has the nice property of natively providing the marginal posterior distribution of parameters, thus allowing for conveniently assessing the variability of estimates. A moderate heritability was estimated for MY and MP; a low heritability was estimated for DMI. The estimated genetic correlation between MY and DMI was relatively strong and positive; the genetic correlation between MY and MP was moderatley negative; no genetic correlation was estimated between DMI and MP. Results from the Bayesian approach were confirmed also by a REML algorithm.

Genomic predictions for DMI, MY and MP were moderately effective: the correlation between observed and predicted phenotypes ranged between 0.35 and 0.69.

The Bayesian methodology for GWAS was developed and tested on milk yield, by comapring results with traditioanl GWAS. Both approaches detected a clear signal of association on BTA14 (DGAT1). Further investigations are needed to clarify the properties of the proposed Bayesian GWAS, before applying it to other traits like methane emissions.

Benefits from the STSM to the METHAGENE network;

The results produced during the STSM at IRTA, CAldes de Montbui, will be very useful for the Methagene consortium. From the methodological perspective, the Bayesian approach to the estimation of genetic parameters, to genomic predictions and GWAS will offer an additional angle from which to analyse methane emission data in dairy cattle and their relationships with milk production and feed intake.

The results themeselves offer interesting insights on the variability and credibility of estimates, and add interesting data to the scientific literature on heritability, genetic correlations and accuracy of genomic predictions from methane emissions and related traits.

Future collaboration with the host institution

The analysis of the data and the further development of the Bayesian GWAS approach will be continued to be carried out jointly by the the two institutions (IRTA and PTP). Given the very good relationships established between the two institutions, it is very likely that further collaborations on this and other projects will be initiated.

Foreseen publications

Two potential publications will emerge from the work carried out during this STSM. The first is the work on the estimation of genetic parameters and on genomic predictons for methane emissions, milk yield and feed intake. This work needs to be completed by adding pedigree data and comparing the relative efficiency of models with genomic data, pedigree data and both sources of information. The second potential publication can stem from the work on a Bayesian approach to GWAS. Some further methodological and theoretical work is needed to better understand its properties; subsequently, sufficient testing of the method will be obtained by applying the Bayesian GWAS to multiple traits.

Confirmation of the host institution of the successful execution of the STSM

See the attached letter from the host institution.

REFERENCES

- Biscarini F, Biffani S, Stella A. Mas alla del GWAS: alternativas para localizar QTLs. arXiv:1504.03802. 2015.

- Marras G, Rossoni A, Schwarzenbacher H, Biffani S, Biscarini F, Nicolazzi EL. Zanardi: an open-source pipeline for multiple-species genomic analysis of SNP array data. Animal Genetics. 2016.
- Misztal I. Reliable computing in estimation of variance components. Journal of animal breeding and genetics. 2008 1;125(6):363-70.
- Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009, 10(10):681-90.
- VanRaden PM. Efficient methods to compute genomic predictions. Journal of dairy science. 2008 30;91(11):4414-23.