COST STSM Reference Number: COST-STSM-FA1302-32569
Period: 2016-03-09 to 2016-03-24
COST Action: FA1302
STSM type: Regular (from Italy to United Kingdom)
STSM Applicant: Dr Gabriele Marras, PTP Science Park, Lodi (IT), gabriele.marras@ptp.it
STSM Topic: *GWAS 2.0: alternative approaches to association studies and in-depth follow-up on detected signals*
Host: Pablo Orozco-ter Wengel, Cardiff University, Cardiff (UK), Cardiff (UK), Orozco-terWengelPA@cardiff.ac.uk

Background and Aim:

Genome wide association studies (GWAS) are used to identify regions of the genome associated with the phenotypes. However, standard GWAS only identifies individual SNPs associated with traits and not directly regions of the genome or genes. Additionally, standard GWAS is prone to return a certain proportion of spurious associations.

We propose alternative and complementary approaches to association studies, that can make the detection of signals of association more robust: one is based on the power of resampling techniques, the other based on local properties of the genome (Biscarini et al. 2015). Furthermore, we take the next step, analysing the results of the association study using functional, Gene Ontology (GO) and pathway analyses, to identify linked genes and the biological processes in which they are involved.

For the identification of significant associations we used three methods: i) standard single-SNP GWAS (using the R GenABEL package); ii) resampled LASSO (Least Absolute Shrinkage and Selection Operator) penalised regression (using the R glmnet package); and iii) the frequency and distribution of runs of homozygosity along the genome (using ad hoc python code).

Our objective was to identify SNPs that were significantly associated with the phenotypes of interest with all three approaches, and use these for the functional, GO and pathway analyses.

The pathway analysis was carried out at the University of Cardiff, under the supervision of Dr. Pablo Orozco-ter Wengel.

During the STSM, we prototyped R pipelines to query relevant biological databases. We used the KEGG (www.kegg.jp) database for searching for metabolic pathways. However, there is no full coverage of the pathways in cattle (Bos taurus), and we decided to resort also to the more complete annotation of the human genome related pathways.

Dataset

For this work we used 772 dairy Holstein cows from two countries (Italy -385- and United Kingdom -387). All animals were genotyped using the GeneSeek Genomic Profiler HD v.2 (139,480 SNPs). We excluded animals with more than 10% of missing genotype, duplicate SNPs (same position), and SNP with MAF < 0.01. Only autosomal chromosomes were considered in this work. In total we used 114,511 SNPs.
For the GWAS, Lasso regression and ROH analysis, missing genotypes were imputed using the Beagle v4.0 software.
Three phenotypes of interest were selected for this work: 1) milk fat content (g/Kg of milk); 2) ruminal methane emissions in grams per day (CH4 g/d); 3) feed intake (digested dry matter intake: g/Kg).

Results

For each phenotype, the three methods for the detection of associations were run. We then selected a panel of common significant SNPs from the GWAS and LASSO analyses. This SNP panel was compared with the results obtained from the ROH analysis. In the end, we used a 25 SNPs for the functional analysis (GO and pathways).

Standard GWAS

GWAS results, reported in Figure 1, show a strong signal of association for milk FAT concentration on BTA14, at around 1,800,000 bps. Many studies have already reported this association, underlying the DGAT1 gene involved in milk fat production. We have also identified a small QTL for FAT also on BTA5.
The results for CH4 g/d and for the DMI do not show strong associations between the genome and phenotype. For ruminal methane emissions, we identified some associated SNPs on BTA4 and BTA10. These very same SNPs were also found associated with DMI.
In addition, we found possible QTLs for DMI on BTA4, BTA12 and BTA16.

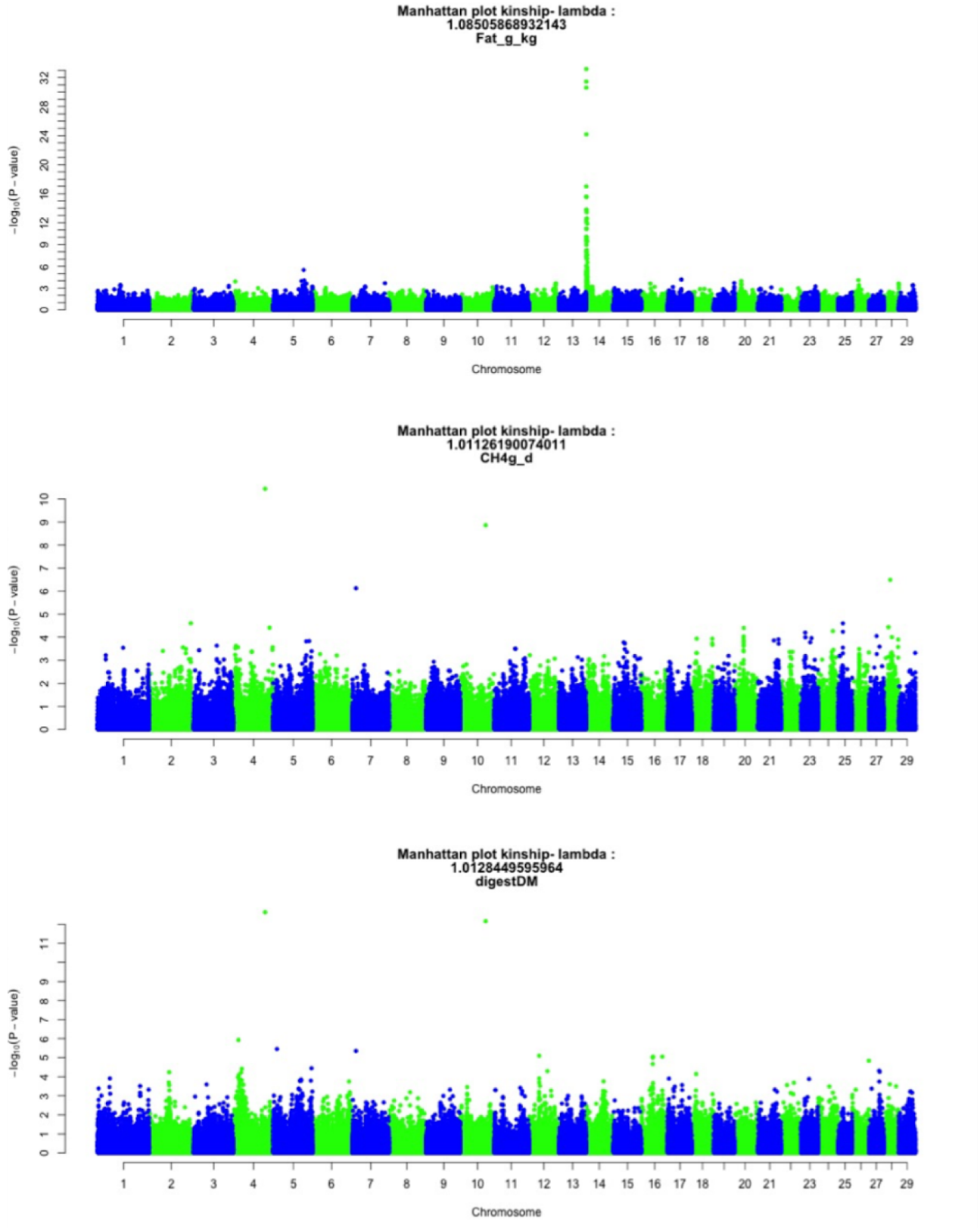

Figure 1: Manhattan plot GWAS for Fat, CH4 and digest DM

Lasso-penalised regression

The Lasso (Least Absolute Shrinkage and Selection Operator) regularization proposed by Tibshirani (1996) aims to achieve a sparse regression model. In other words, it seeks to select a subset of the variables by imposing zero as the value of some regression coefficients. Lasso regularised regression was used to predict our continuous phenotypes based on SNP genotypes. To get valid predictors for the phenotypes, the dataset was divided randomly into training (90% of the animals) and validation (10% of the animals) sets.
This step was repeated 1000 times for each phototype. In each repetition, a 10-fold cross-validation was applied to optimise the lambda hyperparameter (degree of regularization). In each of the 100 repetitions, therefore, a different subset of predictive SNPs was selected. The frequency of inclusion of each SNP in the 1000 resampled predictive models was used as measure of the relative importance of the SNPs for any given phenotype (Biffani et al. 2015). The plot of this frequency against the position on the genome helps visualize the signals of association (Figure 2).
For FAT, significant signals of association appeared on BTA14 at 1,801,116 bps (close to DGAT1), BTA5, BTA19 and BTA20.
For CH4 g/d, we found one SNP used in all tests, on BTA4 at 96,264,925 bps. The same SNP was found to be associated also to DMI.

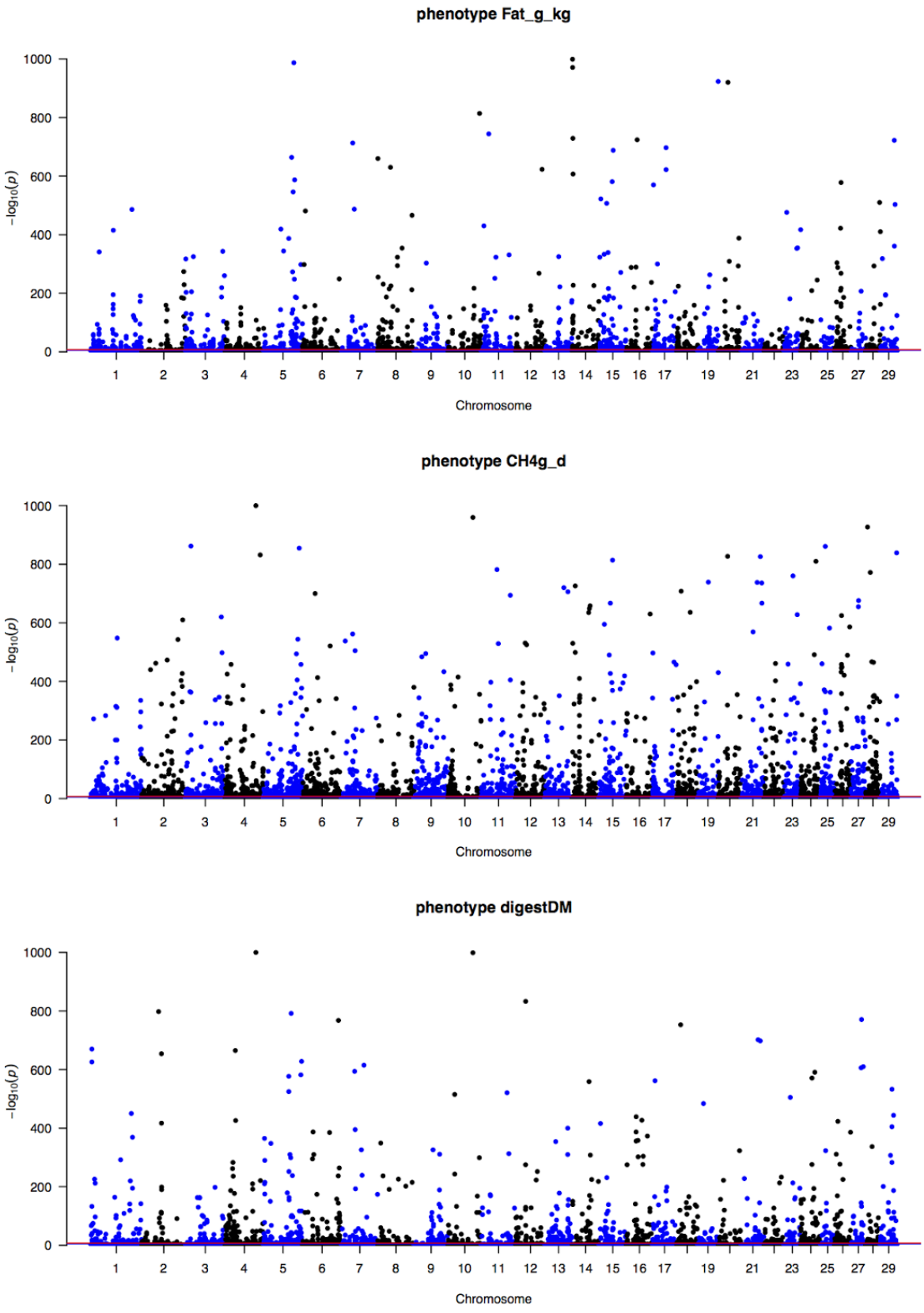

Figure 2: Plot of frequency of inclusion in resampled LASSO regression models of SNPs for Fat, CH4 and DMI

ROH analysis

ROH are defined as DNA segments that harbor uninterrupted sequences of homozygous genotypes. They are interpreted as a measure of autozygosity at genome-wide level.
The aim of this work was to describe the distribution of the ROH in two populations of Holstein cattle that were sub-sampled according to extreme phenotypes (Biscarini et al. 2014).
For each character examined, we have selected the tail of the distribution (positive and negative) and calculated the ROH in these animals. The comparison was made for both cattle populations (Italy and UK) separately.
ROH were calculated using the following parameters: i) minimum 10 SNP in a ROH;, ii) no missing SNP; iii) no heterozygous SNP allowed inside a ROH; iv) a minimum ROH length of 1 Mb (Marras et al. 2015).
In Figure 3, we have reported three different chromosomes for the three phenotypes. In general we can say that there are no very strong signals from the ROH analysis. For FAT we looked at BTA14. In Italian Holstein we can see at 2 Mbp a clear distinction between the negative (red line) and the positive (green line) tails of the phenotypic distribution. This distinction is also present in English Holsteins (though less obvious).
For CH4, BTA20 shows a difference between the extreme phenotypes in the region that goes from 41 to 44 Mbps.
We have also found a differentially autozygous region for DMI on BTA2, between 25-30 Mbps.

Figure 3: ROH for Fat, CH4 and DMI

Biological Pathway

We have selected a group of 25 SNPs that were found to be associated with the analysed phenotypes across the three approaches used. Genes associated (close) to these 25 SNPs were retrieved using the package to R BioMart. Genes were searched for in the Ensembl Bos Taurus database (UMD 3.1; www.ensembl.org). We decided to use a range of 100 kbs around the SNPs, based on the estimates of linkage disequilibrium (LD) in our population. In total, we identified 56 genes.
For each gene identified, we used the human entrezID to search the gene pathway in the human KEGG database (http://www.genome.jp/kegg/).
A pipeline in R was prototyped in collaboration with the University of Cardiff. The main objective of this step was to obtain automatically, from a list of SNPs, all related genes and associated ontologies and metabolic pathways.

Pathway analysis results

The results of the pathway analysis show a large number of pathways; however many have high p-values (non significant). Figure 4 shows the enriched pathways for our group of genes. The first two pathways are involved in the production of purines and in lipid metabolism. The pathway of purines has been identified with two genes, ALLC and PDE7A.
A thorough research suggests that the purine metabolism is involved in the folate biosynthesis, in turn involved in methane metabolism (Figure 5).
In the second pathway the gene involved is the DGAT1. This gene has been widely associated in the GWAS studies for milk production, specifically in milk fat content.

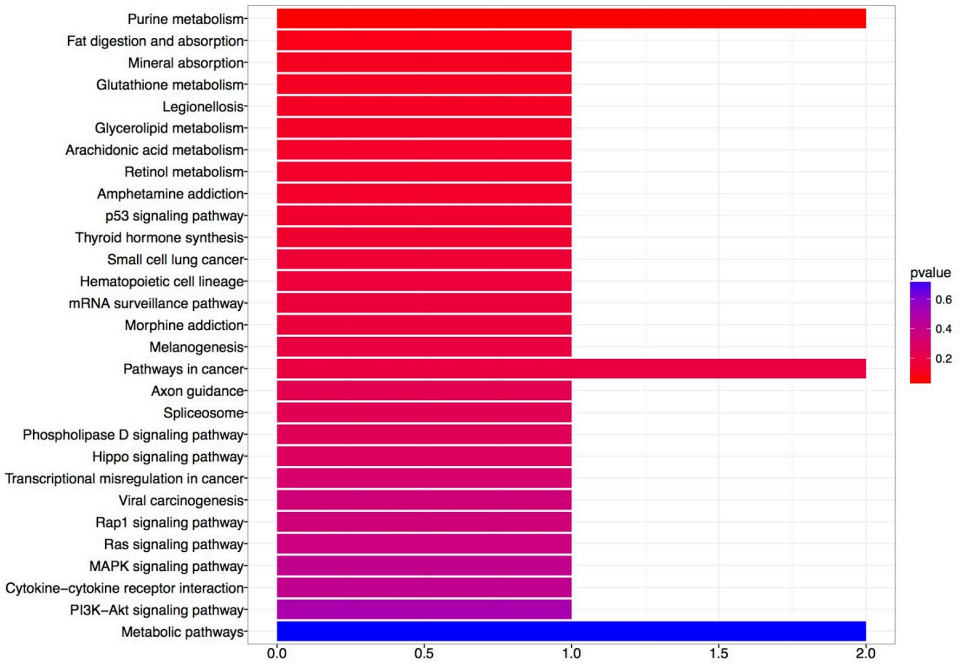

Figure 4: Enrichment genes analysis with KEGG database

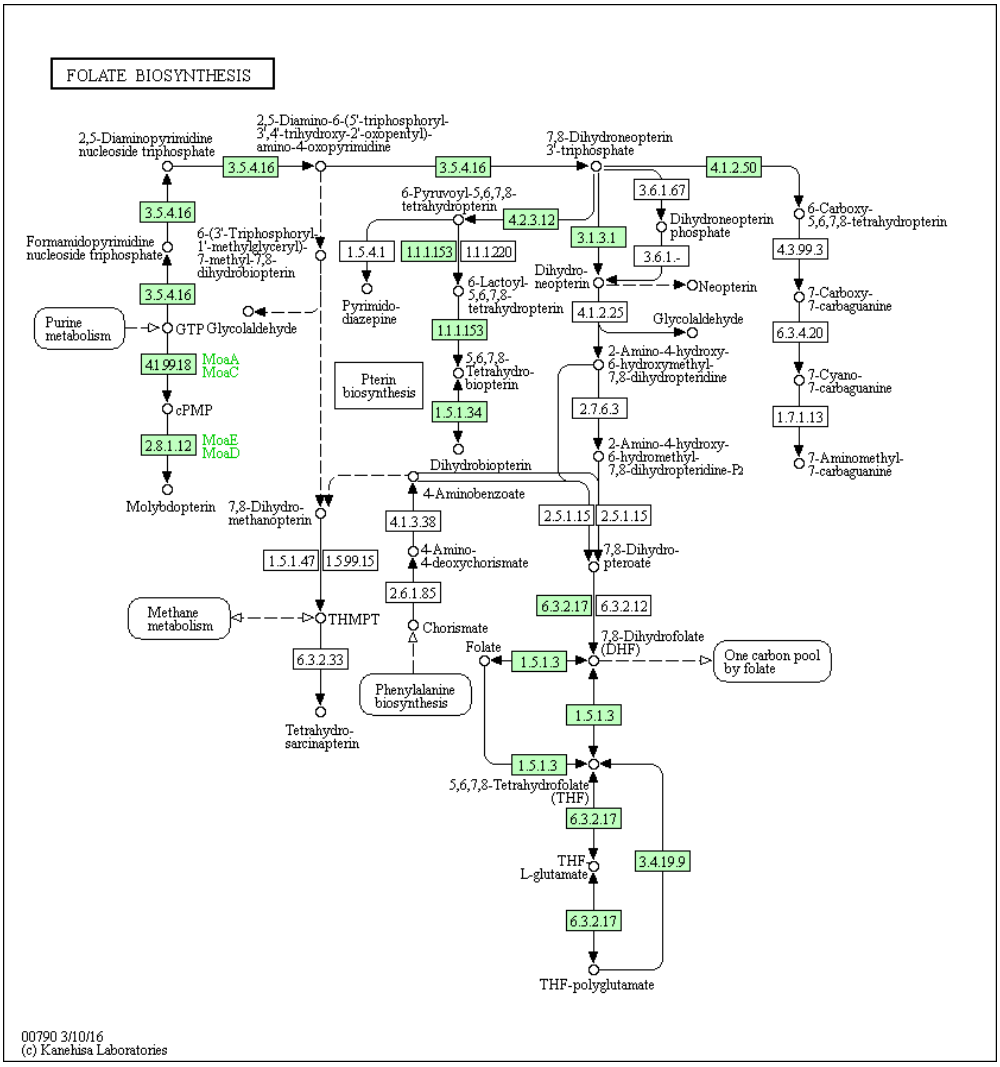

Figure 5: Folate biosynthesis

Conclusions

The results generate during my STSM at Cardiff University are preliminary, but point to promising directions in the detection of robust GWAS signals of association and in the streamlining of post-GWAS analysis. The work done in Cardiff has allowed us to prototype a pipeline to be used for pathway analysis in cattle.

However they are necessary fields of study to understand better the biological process.

In the coming months, the effort will be to refine the post-GWAS pipeline in order to expand and streamline the gene ontology and pathway analysis in Bos taurus, using all relevant public databases.

We will keep on collaborating with the University of Cardiff, in particular with Dr. Pablo Orozco-ter Wengel, for the completion of the work and for the preparation of a joint publication on GWAS 2.0 and post-GWAS analysis for methane emission, milk production and feed intake in dairy cattle.

References

Biffani, S., Dimauro, C., Macciotta, N., Rossoni, A., Stella, A., & Biscarini, F. (2015). Predicting haplotype carriers from SNP genotypes in Bos taurus through linear discriminant analysis. *Genetics selection evolution*, *47*(1), 4.

Biscarini, F., Biffani, S., & Stella, A. (2015). M\'as all\'a del GWAS: alternativas para localizar QTLs. *arXiv preprint arXiv:1504.03802*.

Biscarini, F., Biffani, S., Nicolazzi, E. L., Morandi, N., & Stella, A. (2014). Applying runs of homozygosity to the detection of associations between genotype and phenotype in farm animals. *Reproduction*, *5*(6656617), 6976839.

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone- Marsan, P., Valentini, A., Williams, J.L., Macciotta, N.P.P. (2015): Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Animal Genetics,* 46(2): 110-121.

Tibshirani R. (1996). “Regression shrinkage and selection via the Lasso”. *Journal of the Royal Statistical Society*, 58(1):267–288.