Using expression data to detect small QTL in dairy cattle

I. van den Berg[1], B.J. Hayes[2,3] & M. E. Goddard[1,3]

[1] Faculty of Veterinary & Agricultural Science, University of Melbourne, Parkville 3052, Victoria, Australia
irene.vandenberg@unimelb.edu.au (Corresponding Author)
[2] Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, St Lucia 4067, Queensland, Australia
[3]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

Summary

QTL of small effect account for a large part of the genetic variance in economic traits in dairy cattle but are difficult to be detected. Gene expression data may help to identify such QTL. The objective of this study was to combine results from a Bayesian prediction model, GWAS, expression QTL (eQTL) and expression data to detect variants that are associated both with QTL and gene expression. We identified a number of regions that show a strong correlation between gene expression and local GEBVs for milk traits and fertility, and are therefore likely to harbour regulatory mutations affecting both gene expression and the milk traits. While these regions often explained a large proportion of the variance in expression, the amount of variation in the milk traits explained was small to modest. The results suggest that gene expression levels can be useful in identifying QTL for complex traits, provided the mutation affecting the complex trait is regulatory.

Keywords: gene expression, GWAS, dairy cattle, QTL

Introduction

One way to benefit from whole-genome sequence data for genomic prediction is to use the data to identify quantitative trait loci (QTL) and then include them in the genomic prediction model. While there are few QTL of large effect (e.g. DGAT1, Grisart *et al.,* 2004) for complex traits in cattle, a large part of the genetic variance is likely to be caused by a large number of QTL of small effects. Detecting QTL with small effects is challenging, as the large number of tests in a GWAS requires stringent the detection thresholds to avoid a large number of false positives, maybe too stringent to detect QTL with small effects. Additional information such as gene expression data may be useful to identify of QTL of small effect. However, it is difficult to determine if a complex trait QTL is caused by the same polymorphism as an expression QTL (eQTL). The objective of this study was to combine results from a Bayesian prediction model, GWAS, eQTL study and expression data to detect variants that are associated both with QTL and gene expression.

Material and methods

Calculation of local GEBVs

The dataset used to estimate SNP effects for all SNPs on the Illumina 800K BovineHD bead

chip (HD) consisted of 35,775 Holstein, Jersey and crossbred bulls and cows for milk, fat and protein, 32,923 Holstein and Jersey bulls and cows for fat% and prot%, and 32,819 Holstein, Jersey and crossbred bulls and cows for fertility (calving interval). The analyses were performed using daughter trait deviations (DTD) and trait deviations (TD) for bulls and cows, respectively. All individuals were either genotyped with the Illumina BovineSNP50 chip and imputed to, or directly genotyped for the HD chip. Bayes R hybrid (Wang *et al.*, 2016) was used to estimate marker effects for the HD SNP for all traits. Subsequently, local GEBVs were calculated by summing up the effects of the SNP alleles carried by each animal in sliding intervals of 250 kb for 131 Holstein and Jersey individuals with expression data. Intervals overlapped by 50 kb.

GWAS

For all individuals described in the previous section, HD genotypes were imputed to full sequence using sequences of Holstein, Jersey and Australian Red bulls and cows from Run 5 of the 1,000 Bulls Genome Project. After removing all variants with a minor allele frequency lower than 0.002 and removing variants in high LD ($r^2 > 0.9$), the sequence dataset contained 4,812,745 variants (SEQ). Genome-wide association studies (GWAS) were performed for the SEQ variants for all traits, using GCTA software (Yang *et al.*, 2011). Because GCTA does not allow weighted phenotypes and DTD in bulls are more accurate than TD in cows, GWAS were performed separately in bulls and cows, and combined in a meta-analysis, using the weighted z-scores model with Metal (Willer *et al.,* 2010).

eQTL analysis

Expression data was obtained from milk samples of 131 Holstein and Jersey individuals, and blood samples of 105 Holstein individuals. Only genes that were expressed in at least 25 individuals were analysed, resulting in 12,772 and 11,577 genes in milk and blood, respectively. All individuals with expression data had imputed HD and sequence genotypes. After filtering out sequence variants with a MAF less than 0.05, 10,904,750 and 10,469,612 sequence variants were used for the eQTL detection for milk and blood, respectively (Chamberlain et al. 2018). For each of these genes, the association of expression level was tested with all variants on the chromosome the gene was located on.

Selection of QTL intervals correlated with gene expression

First, intervals were selected where the local GEBV variance was at least 1/10,000 * the total additive genetic variance. This arbitrary threshold was used to select intervals that have at least some association with the trait that was analysed, but was not very stringent as we focussed on small QTL. For each of these intervals, the correlation between the local GEBVs for milk production traits and gene expression levels (log transformed read counts) of genes within 1 Mb of the interval was calculated. Intervals were considered further if the p-value of the correlation was lower than 10^{-5}.

Results and discussion

When cells from milk were used for the expression analysis, there were 1, 3, 0, 2, 4 and 0 genome intervals selected for milk, fat, protein, fat percentage (fat%), protein percentage

(prot%) and fertility, respectively, because the correlation between local GEBV and expression level was $P < 1x10^{-5}$. Using blood tissue this was larger, with 3, 13, 7, 8, 11 and 1 intervals for milk, fat, protein, fat%, prot% and fertility, respectively, as shown per chromosome in Figure 1. From the 10 intervals selected using milk tissue, 5, 8 and 3 had at least one variant with a p-value below 10^{-5} in the eQTL analysis, GWAS and both the eQTL analysis and the GWAS, respectively. From the 43 intervals selected using blood tissue, there were 34, 24 and 20 with at least one variant with a p-value below 10^{-5} in the eQTL analysis, GWAS and both the eQTL analysis and the GWAS, respectively.

Chromosome 18 had the largest number of significant intervals. Most of these intervals were associated with the expression of fucokinase (*FUK*), a gene reported by Ibeagha-Awemu *et al.* (2016) as a potential candidate gene for milk traits, and associated with butyric acid levels (C4:0). Figure 2 shows local GEBV variances in this region and the p-values for the correlations between the local GEBVs and *FUK* expression. The strongest correlation with *FUK* expression was 0.90, with a p-value of 5.4×10^{-39}, for the interval located from 1,496,152 to 1,746,152 bp.

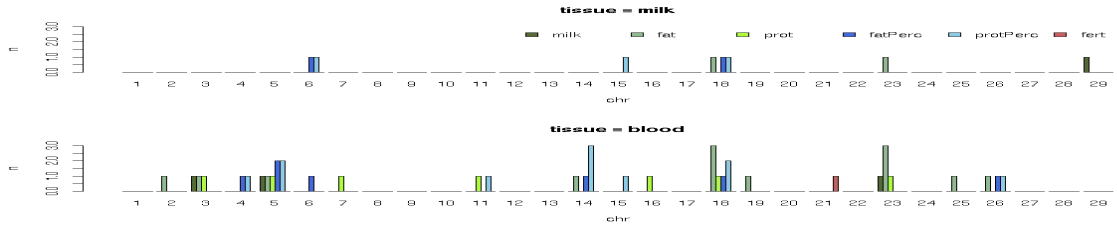

Figure 1. The number of significant intervals ($P<1x10^{-5}$) per trait and chromosome, correlating local GEBVs to gene expression levels in blood tissue.

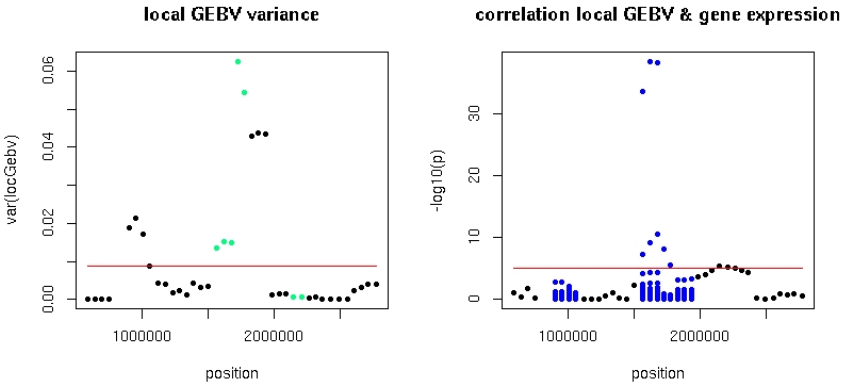

Figure 2. Intervals selected for fat yield on chromosome 18. Left = local GEBV variance with the intervals with a p-value below 10^{-5} in green, right = -log10(p-values) of correlations between local GEBVs and *FUK* expression with in blue the intervals with a local GEBV variance of at least 1/10,000 * the total additive variance.

Figure 3 compares the association of sequence variants in the region around the *FUK* gene on chromosome 18 with fat yield and the expression of FUK. The most significant GWAS hit for fat yield is in an intron variant in the DEAD-box helicase 19B (*DDX19B*) gene, located at 1,804,006 bp, outside the interval that showed the strongest correlation with *FUK* expression. While there were stronger correlations between local GEBV for fat yield and

FUK expression, there were also intervals with strong correlations between local fat GEBVs and the expression of *DDX19B*. The strongest detected correlation with *DDX19B* was 0.88, with a p-value of 2.4×10^{-30}, for an interval between 1,550,757 and 1,800,757 bp. The most significant variant detected in the eQTL analyses associated with *FUK* expression is in an intron located in the splicing factor 3b subunit 3 (*SF3B3*) gene, located at 1,634,115 bp, within the GEBV interval that showed the strongest association with the expression of *FUK*. The strongest correlation with *SF3B3* expression was -0.23, with a p-value of 0.02, for a GEBV interval between 628,064 and 878,064 bp.

The GEBV interval that showed the strongest correlation with *FUK* expression contains a large number of sequence variants. The peak observed in the GWAS is located slightly beside the peak in local GEBV variance, while the eQTL peak is located within the interval. The difference in location of the GWAS and eQTL peaks, in combination with the presence of multiple genes in the region showing a correlation with gene expression and local GEBVs, makes it difficult to pinpoint specific sequence variants that are associated with both gene expression and fat yield.

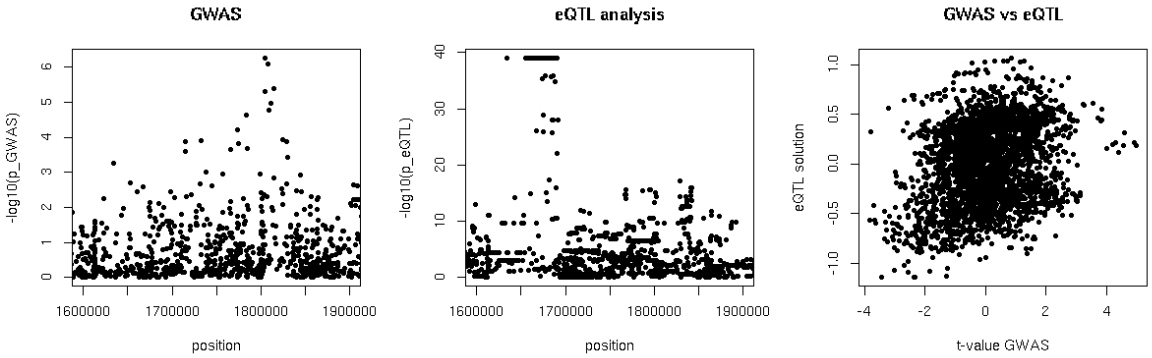

Figure 3. Comparison of eQTL and GWAS results for fat yield on chromosome 18. Left: association between sequence variants and fat yield, middle = association with sequence variants and expression of *FUK*, right = comparison of t-values for the GWAS and eQTL effects.

Conclusions

We identified a number of regions that show a strong correlation between gene expression and local GEBVs. While these regions did not explain a large part of the total genetic variance in milk production traits, they often contained significant GWAS and eQTL hits.

Acknowledgements

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research. We acknowledge DataGene for providing access to data used in this study. We acknowledge our partners in the 1000 Bull Genomes Project for access to the reference genomes.

List of References

Chamberlain AJ, BJ Hayes BJ, R Xiang, CJ Vander Jagt, CM Reich & IM Macleod, et al.

Identification of regulatory variation in dairy cattle with RNA sequence data. 11th World Congress on Genetics Applied to Livestock Production (WCGALP); Auckland, New Zealand2018.

de los Campos, G.G., A.I. Vazquez, R. Fernando R, Y.C. Klimentidis & D. Sorensen, 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS genetics 9(7):e1003608.

Grisart B., F. Farnir F, L. Karim, N. Cambisano, J.J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frere, W. Coppieters & M. Georges, 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA. 101(8):2398-403.

Ibeagha-Awemu E.M., S.O. Peters, K.A. Akwanji, I.G. Imumorin & X. Zhao, 2016. High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. Sci. Rep. 6:31109.

Wang, T., Y.-PP. Chen, P.J. Bowman, M.E. Goddard & B.J. Hayes, 2016. A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping. BMC Genomics 17:744.

Willer, C. J., Y. Li, and G. R. Abecasis, 2010. METAL: Fast and efficient meta-analysis of genome wide association scans. Bioinformatics 26:2190–2191.

Yang, J., S.H. Lee, M.E. Goddard & P.M. Visscher, 2011. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 88(1): 76-82.