Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 395-398

USING MULTI-OMICS DATA TO IDENTIFY CANDIDATE GENES FOR MILK LACTOSE PERCENTAGE IN DAIRY CATTLE

M. Ghoreishifar[1,2] R. Xiang[1,3], T. Lopdell[4], M. Littlejohn[4], A. Chamberlain[1,2], J. Pryce[1,2] and M. Goddard[1,3]

[1] Agriculture Victoria Research, AgriBio Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia
[3] Faculty of Science, University of Melbourne, VIC, 3010 Australia
[4] Livestock Improvement Corporation, Private Bag 3016, Hamilton, 3240 New Zealand

SUMMARY,

Genome-wide association studies (GWAS) have successfully identified many quantitative trait loci (QTL) associated with complex traits in humans and livestock. However, pinpointing the causal variants and target genes through which these QTL influence phenotypes remains challenging. Transcriptome data from mammary tissue were analysed using genetic score omics regression (GSOR). This method correlates observed gene expression to genetically predicted phenotypes and is used to find associations between gene expression and genomic breeding values (GEBVs). GSOR identified 706 genes whose expression in mammary glands is significantly associated with GEBVs estimated for milk lactose percentage (FDR ≤ 10%). We observed a significant co-occurrence between the identified genes and GWAS signals for lactose percentage reported in an independent study (*P*=2.5E-10; odds-ratio=3.8). The co-occurring genes with GWAS signals enriched for ion transport Gene Ontology (GO) term included *ATP6V0A2*, *DNAH10*, *LRRC8B*, *LRRC8C*, *KCNJ2*, *P2RX4*, and *SLC34A2*. These findings introduce known and novel candidate genes related to the regulation of milk lactose, showing their involvement in the ion transport mechanism and supporting the importance of the osmoregulatory function of lactose in milk.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified many quantitative trait loci (QTL) associated with complex traits in humans and livestock. However, pinpointing the causal variants and target genes through which these QTL influence phenotypes remains challenging. This difficulty arises due to the small effect sizes of most variants, linkage disequilibrium (LD) among nearby variants, and the fact that most QTLs are located in non-coding regions of the genome. QTL in non-coding regions are believed to affect a phenotype by regulating the expression of their target genes; in such cases, they are referred to as expression QTL (eQTL).

Three types of evidence are used to identify the genes through which eQTL operates: (1) genes located near the most significant GWAS variant, (2) genes whose expression is correlated with the trait based on the GSOR method (unpublished data), and (3) genes whose physiological roles are related to the trait. Although none of these types of evidence is definitive on its own, their convergence on the same genes provides convincing evidence supporting the correct identification of genes (*unpublished data*).

The current study has three primary objectives: (1) to identify genes whose observed expression levels in the mammary gland of NZ dairy cows are significantly associated with their estimated genomic breeding values (GEBVs) for milk lactose percentage, (2) to investigate the extent of co-occurrence between the associated genes and signals obtained from an independent GWAS on lactose percentage, and (3) to determine whether the co-occurring genes (with GWAS signals) enriched in a specific Gene Ontology (GO) term.

MATERIALS AND METHODS

Phenotypic data. Phenotypic data for test-day lactose percentage was provided by DataGene Pty Ltd (Melbourne, Australia). Outliers deviating ±3 SD of the mean phenotypic value were excluded. Test-day records were included if the cow's age at calving was between 18-25 months and days in milk (DIM) between 5-315 days. The final data contained ~4.99 million test day records for ~ 477 K cows. Using ASReml (Gilmour *et al*. 2022), adjusted phenotypes were estimated and averaged for each cow (i.e. effect of Cow) using the model proposed by Khansefid *et al*. (2023):

$$\boldsymbol{y_{ijklm}} = \mu + \boldsymbol{H_i TD_j} + \boldsymbol{M_k} + \boldsymbol{pol(DIM, 8)} + \boldsymbol{pol(Age, 2)} + \boldsymbol{Cow_l} + \boldsymbol{e_{ijklm}}$$

where, $\boldsymbol{y_{ijklm}}$is the test-day record for lactose percentage, μ is the effect of overall mean, $\boldsymbol{H_i TD_j}$ is the effect of the ith herd and jth test date; $\boldsymbol{M_k}$ is the effect of the kth calving month; $\boldsymbol{pol(DIM, 8)}$ and $\boldsymbol{pol(Age, 2)}$ are the regression coefficients of Legendre polynomials of order 1–8 for DIM and of order 1–2 for age at calving in months; $\boldsymbol{Cow_l}$ and $\boldsymbol{e_{ijklm}}$ are the random effect of the lth cow and the random residual term, respectively.

Genotype data. Genotype data were available for 81,658 Australian cows, of which 79% were Holstein, 16% Jersey and 5% Aussie Red. Genotypes were imputed to whole genome sequences using Run9 of the 1000 Bull Genomes project (Daetwyler *et al*. 2014). Variants with minor allele frequency (MAF) < 0.01 and genotype frequencies departing from Hardy-Weinberg equilibrium (*P* < 1E-8) were excluded. LD pruning was performed to exclude variants that are in high LD (r^2 > 0.95). These procedures retained 1,181,628 variants for subsequent analyses. Using genotypes and phenotypes of 81,658 cows described above, we trained a BayesR3 (Breen *et al*. 2022) model to estimate prediction equations (SNP effects) for lactose percentage. The model was as follows: $\boldsymbol{y} = \boldsymbol{Xu} + \boldsymbol{Vg} + \boldsymbol{e}$, where $\boldsymbol{y}$ is vector of adjusted phenotypes; $\boldsymbol{X}$ is an incidence matrix, **u** is a vector of fixed effect including breed (three levels); **V** is the coded genotype; **g** is a vector of SNP effects; and **e** is the residual term. BayesR3 model was run using the default parameters. The estimated SNP effects were used to calculate $\widehat{\boldsymbol{GEBV}}_{\boldsymbol{cis}}$ (in the gene expression data) using nearby variants (1 Mb) to a specific gene.

Gene expression data and GSOR. Expression measures of ~12.3 K genes in mammary tissue from ~350 New Zealand (NZ) cows (including Holstein Friesian, Jersey and their crosses) were used. Imputed whole genome sequence variants were available for these cows. The processing of samples, RNA extraction, library preparation, and RNA sequencing were described in detail in Littlejohn *et al*. (2016). The same set of sequence variants as described for the Australian data were kept for the NZ animals. To perform GSOR, we used a method called genetic score omics regression, abbreviated to GSOR (unpublished data). This method estimates the significance of the correlation between the expression level of a gene and the GEBVs derived from nearby variants of that gene ($\widehat{\boldsymbol{GEBV}}_{\boldsymbol{cis}}$). The following model was used:

$$\widehat{\boldsymbol{GEBV}}_{\boldsymbol{cis}} = \boldsymbol{b}\Omega + \boldsymbol{g} + \boldsymbol{e}$$

where, $\widehat{\boldsymbol{GEBV}}_{\boldsymbol{cis}}$ is a vector calculated using the effect estimated for variants located in the 1 Mb region surrounding the gene, **b** is the regression coefficient of the $\widehat{\mathbf{GEBV}}_{\boldsymbol{cis}}$ on Ω, which is a vector representing gene expression measures; and **g** is a vector of random polygenic effects with $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$, where **G** is the genomic relationship matrix estimated from genome-wide SNPs, and σ_g^2 is the additive genetic variance explained by genome-wide SNPs; $\boldsymbol{e}$ is the vector of residuals with $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where **I** is an identity matrix and σ_e^2 is residual variance. To account for multiple testing, we applied Benjamini-Hochberg correction and genes with *FDR* ≤ 0.10 were deemed significant.

Do GSOR significant genes overlap with GWAS signals? To test this hypothesis, we investigated the overlap (co-occurrence) between the GSOR significant genes, whose expression levels are correlated with $\widehat{\mathbf{GEBV}}_{\boldsymbol{cis}}$, and a GWAS on lactose concentration previously reported in the study by (Lopdell *et al*. 2017). GWAS results were downloaded and lifted over to the new reference

genome (ARS-UCD 1.2), and SNPs with $P \leq 1E-8$ was regarded as significant (i.e., GWAS signals).

To find the degree of overlap between the GSOR significant genes and GWAS signals, we partitioned the genome into non-overlapping windows of 100 Kb. We counted the number of windows containing at least one GWAS signal, and a total number of GSOR significant genes located within these windows. These quantities were compared against a total number of windows and total number of GSOR significant genes. We used a Fisher Exact test to investigate significance of our findings at alpha ≤ 0.05.

Functional annotation analyses. We conducted functional annotation analyses for the GSOR significant genes that co-occur with GWAS signals, using the total number of GSOR significant genes as the background. We used DAVID bioinformatic tool (Sherman *et al*. 2022) for this analysis with UniProt Keyword annotation, and terms with FDR ≤ 0.05 were considered significant.

RESULTS AND DISCUSSION

The correlations between the expression of 12,237 mammary genes across cows and $\widehat{\mathbf{GEBV}}_{cis}$ for lactose percentage were tested using GSOR. Of the tested genes, 704 significant associations were found at FDR ≤ 10%. However, similar to GWAS, GSOR is also susceptible to false positives (unpublished data). Therefore, we assessed the overlap between GSOR significant genes and GWAS signals by evaluating the extent of co-occurrence within non-overlapping windows. The results are presented in Table 1 and suggest there is a meaningful biological agreement between the GWAS findings and the GSOR results. For example, using 100 Kb windows, there were 24,861 non-overlapping windows, and 706 GSOR significant genes. We found 321 windows containing GWAS signals, and 35 GSOR significant genes located within these windows (co-occurred with GWAS signals), resulting in a *p*-value of 2.5E-10 (odds ratio=3.8).

Table 1. Fisher Exact Test of co-occurrence between GSOR significant genes and GWAS signals using non-overlapping windows of different sizes

Window size (Kb)	N of Windows	N of GSOR significant genes	N of windows containing GWAS signal(s)	N of GSOR genes in windows containing GWAS signal(s)	*P*-value (odds-ratio)
100	24,861	706	321	35	2.5E-10 (3.8)
500	4,984	706	186	57	3.1E-6 (2.16)

GSOR can be confounded by LD where the correlation between gene expression and trait can be caused by SNP marker(s) instead of a causal variant (*unpublished data*). Such a correlation does not indicate a causal link between the expression of the gene and the trait. Therefore, additional sources of evidence could help prioritize the causal gene-trait associations at GWAS loci.

If GSOR significant genes co-occurring with GWAS signals include causal genes, they are expected to show enrichment for biologically relevant GO terms. We tested this hypothesis by investigating the enrichment of GO terms for the genes that show a GSOR correlation and are near GWAS signals. Our results showed the biological terms Transport, and its child term Ion transport are significantly enriched (Table 2) with 11 and seven genes involved in these GO terms, respectively.

Our findings are consistent with a previous study on milk lactose traits (Lopdell *et al*. 2017), that highlighted the role of membrane transport genes, including *LRRC8C*, *P2RX4*, KCNJ2 and *ANKH* as key modulators of milk lactose content. In addition to the first three genes, our study also identified novel genes including *ATP6V0A2*, *DNAH10*, *LRRC8B* and *SLC34A2* which are involved

in Ion transport biological process. These genes are expected to influence osmotic balance through modulation of ion concentrations in milk.

Table 2. Enrichment of GSOR significant genes co-occurred with GWAS signals compared to background genes (i.e., the total GSOR significant genes)

Category	Term	Genes	*P* value	FDR
Biological process	Transport	ATP6V0A2, RAB5C, DNAH10, LRRC8B, LRRC8C, KCNJ15, KCNJ2, P2RX4, P2RX6, SLC34A2, SLC50A1	2.64E-4	5.5E-3
Biological process	Ion transport	ATP6V0A2, DNAH10, LRRC8B, LRRC8C, KCNJ2, P2RX4, SLC34A2	7.4E-4	8.2E-3
Molecular function	Ion channel	LRRC8B, P2RX4, P2RX6, LRRC8C, KCNJ15, KCNJ2	1.5E-4	2.5E-3

CONCLUSION

Our study identified 704 genes from mammary gland, whose expression levels were associated with milk lactose percentage. Of these genes, a significant proportion showed co-occurrence with GWAS signals reported in an independent study. Enrichment analysis of these co-occurring genes highlighted transport and ion transport processes, implicating both known and novel membrane transport genes in regulating osmolality and influencing milk lactose concentration. These findings offer valuable insights into the genetic basis and biological regulation of lactose percentage.

ACKNOWLEDGEMENTS

The authors acknowledge DataGene for providing the Lactose data, as well as Dr Tuan Nguyen and Dr Iona MacLeod for imputation of sequence genotypes. We also appreciate the financial support from DairyBio, a joint venture of Dairy Australia, The Gardiner Foundation, and Agriculture Victoria (Melbourne, Australia).

REFERENCES

Breen E.J., MacLeod I.M., Ho P.N., Haile-Mariam M., Pryce J.E., Thomas C.D., Daetwyler H.D. and Goddard M.E. (2022) *Commun. Biol*. **5**: 661.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., Van Binsbergen R., Brøndum R.F., Liao X., Djari A., Rodriguez S.C., Grohs C. and Esquerré D. (2014) *Nat. Genet*. **46**: 858.

Gilmour A.R., Gogel B.J., Cullis B.R., Welham S.J., Thompson R., Butler D., Cherry M., Collins D., Dutkowski G. and Harding, S. (2014) ASReml User Guide Release 4.1. VSN International Ltd, Hemel Hempsted, UK.

Khansefid M., Pryce J.E., Shahinfar S., Axford M., Goddard M.E. and Haile-Mariam M. (2023) *Anim. Prod. Sci*. **63**: 1031.

Littlejohn M.D., Tiplady K., Fink T.A., Lehnert K., Lopdell T., Johnson T., Couldrey C., Keehan M., Sherlock R.G., Harland C. and Scott A. (2016) *Sci. Rep-UK* **6**: 25376.

Lopdell T.J., Tiplady K., Struchalin M., Johnson T.J., Keehan M., Sherlock R., Couldrey C., Davis S.R., Snell R.G., Spelman R.J. and Littlejohn M.D. (2017) *BMC Genom*. **18**: 1**.**

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T. and Chang W. (2022) *Nucleic Acids Res*. **50**: W216.