Contributed paper

IDENTIFICATION OF GENETIC VARIANTS LINKING DAIRY FERTILITY AND MILK PRODUCTION TRAITS

E. Ooi[1,2], J.E. Pryce[2,3] and M.E. Goddard[1,2]

[1] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC, 3052 Australia
[2] AgriBio, Department of Environment & Primary Industries, Bundoora, VIC, 3083 Australia
[3] Biosciences Research Centre, La Trobe University, Bundoora, VIC, 3083 Australia

SUMMARY

Fertility in dairy cattle has declined as an unintended consequence of selection for high milk yield. The negative genetic correlation between milk yield and fertility is now well-documented, however, the underlying biological causes are still uncertain. The objective of this study was to examine this problem from a genomic perspective by first identifying the variants that link dairy fertility and milk production traits, and then using an archetypal clustering method to group variants with similar patterns of effects. Each cluster was finally subjected to over-representation analysis to identify the biological processes underpinning variants with similar effects. Nine groups with distinct effects on production, fertility and conformation traits were identified. Initial results from over-representation analysis suggest that the clusters formed are consistent with prior knowledge about the associated genes, but also suggest new areas of interest for further research.

INTRODUCTION

Fertility in dairy cattle has declined over the last 50 years as an unintended consequence of selection for high milk yield. Lactation is obviously contingent on parturition, making fertility a key driver of profitability, particularly on pasture-based dairy farms. The ideal cow does not only conceive – she does it at the right time, on the first attempt, and achieves and maintains pregnancy despite producing 60+ litres of milk per day.

The exact physiological mechanisms linking fertility and milk production are still uncertain, despite significant research investment. Results from observational studies and in vivo experimentation have been equivocal – largely because nutrition, health, management interventions and environmental factors all combine to confound analysis of herd reproductive performance.

Advances in genomics allow a direct approach to testing hypotheses. However, from a genetic perspective, fertility is a complex trait composed of successive biological events, with phenotypes that are difficult to measure. In this study, the use of a genome-wide association study incorporating large multi-breed reference population and a subset of variants which have been pre-selected for significance gives us significantly more power to identify variants of interest. It also allows us to identify variant clusters that have similar effects on multiple traits possibly indicating a common physiological pathway.

This study aims to uncover the physiological mechanisms underlying milk production and fertility, which may assist herd managers in uncoupling these traits to breed cattle that are both productive and highly fertile.

MATERIALS AND METHODS

Data preparation. Genotype and phenotype data for a total of 5,123 bulls and 29,081 cows from DataGene, Australia were used for this study. This data included a mix of Holstein-Friesians (4,061 bulls/22,899 cows) and Jerseys (1,062 bulls/6,174 cows).

Genotypes included a total of 46,771 sequence variants, which were selected from a total of 17,669,372 imputed variants prepared according to a multi-phase method which includes regression

involving FAETH scores, variant clustering and pruning, and Bayesian approaches (Xiang *et al*. 2021). Two hundred and forty-seven variants thought to be informative for milk fat and protein percentage from an analysis performed by van den Berg *et al*. (2020) were also included.

Phenotype data included trait deviations and daughter trait deviations for cows and bulls, which were calculated using a model that corrects for fixed effects including herd, season, and year. Twelve traits were selected which are thought to have effects on production and/or fertility, including protein yield, fat yield, protein percentage, fat percentage, milk yield, fertility, direct survival, stature, angularity, bone quality, udder texture and body condition score.

Single-trait GWAS. Each trait was analysed one at a time in each sex with linear mixed models using GCTA (Yang *et al*. 2011). Results for both genders were then combined using a weighted meta-analysis based on a method described in (Xiang *et al*. 2018). This allowed us to fully utilize GWAS summary data and thereby expand the power of our analysis.

Although most of the initial 46,771 variants were the result of LD pruning in the set of 1.7 million variants identified by (Xiang *et al*. 2021), we found that for known QTL with large effects such as DGAT1, some variants remained in high LD. To remove these, further post-processing was undertaken using the snp_clumping function within R package bigsnpr (Privé *et al*. 2018). This function is analogous to the –clump function implemented in PLINK 1.9 but has been adapted for memory-efficient usage within the R environment. For our study, as we were most interested in the relationship between milk production and fertility traits, we used fertility t-values as our ranking statistic. This reduced the starting set of 46,771 variants to 15,220 variants.

Archetype-based clustering. We then clustered the sequence variants according to their pattern of effects on the 12 traits of interest. This was done by first ranking the variants in descending order according to the magnitude of their effect size on these traits, and then completing iterative pairwise comparisons of their cosine similarity. Whenever a variant was identified which had < 0.8 cosine similarity with the index variant, it was considered a new archetype. Subsequent variants were considered to represent new archetypes only if this held true for all preceding archetypal variants.

Using this method, we identified 9 archetypal sequence variants with unique patterns of effects on the traits of interest. The remaining 15,211 variants were then assigned to the archetype with which they had the highest measure of cosine similarity, forming 9 variant clusters. The direction of effects was standardised across variants.

Enrichment analysis. To better understand the underlying biology for each of the 9 clusters, pathway analysis was performed on each cluster using the over-representation analysis (ORA) function provided by a gene-set analysis toolkit, WebGestalt (Liao *et al*. 2019).

RESULTS AND DISCUSSION

It is important to note in Figure 1 that, as the fertility trait is measured by calving interval, positive effects represent infertility. With this in mind, we can distinguish 4 broad groups amongst the 9 variant clusters. One primarily affects fertility (i.e., clusters 3 and 8), one affects production traits with a negative effect on fertility (i.e., cluster 9), and one affects production traits without impacting fertility (i.e., clusters 1, 5). Another group could be considered to include clusters which have varying effects on conformation, particularly in clusters 4 and 8.

Cluster 1 includes genes such as DGAT, FASN and MGST1, which have all be implicated in fat synthesis. The pattern of effects is consistent with this, with fat, fat percentage and protein percentage traits in the opposite direction to milk yield and protein. There is little impact on other traits. The most represented GO terms reported by ORA include fat cell differentiation, carbohydrate derivative biosynthetic process, response to toxic substance, lipid biosynthetic process and endocrine system development.

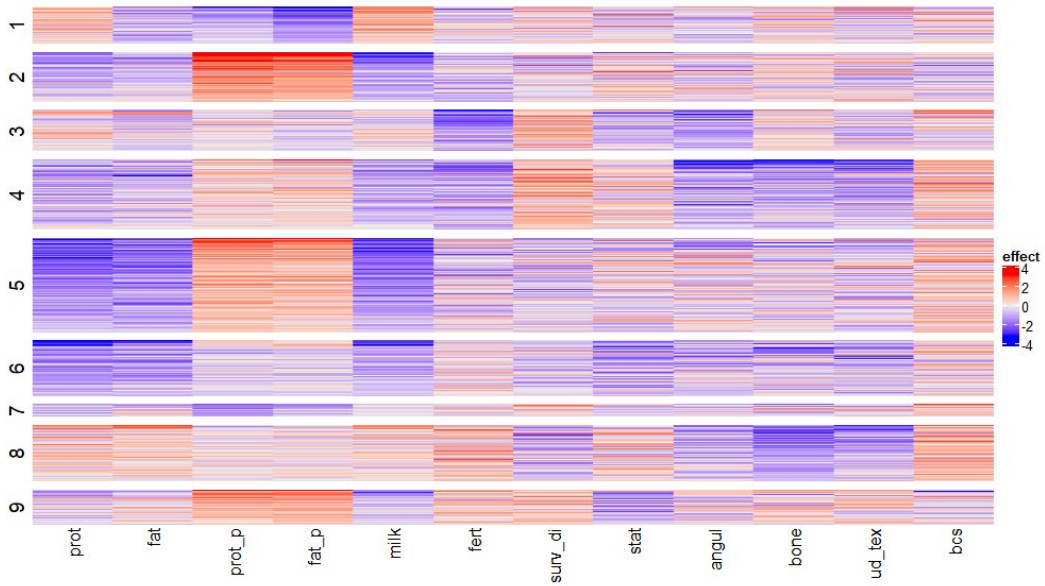

Figure 1. Nine clusters exhibiting shared patterns of effect for 15,220 variants on 12 traits

Cluster 2 has strong effects on protein and fat percentage, which is likely due to an antagonistic effect on milk volume. Notable genes include RORA, LAMA4 and PROX1. The most represented GO terms included cardiovascular system development, tube morphogenesis, regulation of cellular response to stress, and cell fate commitment.

Cluster 3 displays strong effects on fertility, direct survival, and angularity. Notable genes include NOG, ASCL1 and GDNF. The most represented GO terms included neuron death, neuron development, appendage development, regulation of system process, and regulation of cell development.

Cluster 4 also has strong effects on fertility and direct survival, with some interaction with conformation traits and weaker but consistent effects on production traits. Notable genes include LRRK2, DHX36 and BMP7. The most represented GO terms included regulation of nervous system development, regulation of cell development, regulation of cell projection organisation, regulation of secretion, and response to inorganic substance.

Cluster 5 represents very strong production effects, without impacting conformation or fertility. Notable genes include ADCYAP1, EDN1, and TGFBR1. The most represented GO terms included carbohydrate derivative transport, multicellular organismal response to stress, circulatory system process, anion transport, and response to growth factor.

Cluster 6 comprises variants with effects on fat, protein and milk yield that do not affect fat and protein percentage. Notable genes include BMP4, TP63 and WNT5A. The most represented GO terms included negative regulation of developmental process, signal transduction by p53 class mediator, cranial skeletal system development, positive regulation of cell proliferation, and epithelial cell proliferation.

Cluster 7 affects protein percentage and not much else. Notable genes include BCL2, IL6 and ISL1. The most represented GO terms included peptidyl-threonine modification, peptidyl-serine modification, tricarboxylic acid metabolic process, negative regulation of transcription, and regulation of ion transport.

Cluster 8 represents conformation traits, along with body condition score and possibly fertility.

Notable genes include BMP7, MEF2C and SYK. The most represented GO terms included connective tissue development, cardiovascular system development, appendage development, tube morphogenesis, and integrin-mediated signaling pathway.

Cluster 9 is similar to cluster 2 in that it primarily affects protein and fat percentage. However, unlike cluster 2 it also has effects on fertility, direct survival, and stature. Notable genes include ARRDC3, LGR4 and CIB1. The most represented GO terms included second-messenger-mediated signaling, G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger, secretion by cell, cellular component disassembly, and cell-cell adhesion.

Care must be taken when interpreting these preliminary results, particularly when pathway analysis has been performed. Pathway analysis is still a developing area in computational biology, with no current consensus as to the best tool, method, or annotation database to utilise. Pathway analysis also requires a gene to be linked to each variant, which is a complex problem. Although GWAS can identify genetic loci associated with complex traits, the causal gene associated with each locus is often difficult to determine. This is because firstly, LD between loci can mask the identity of the causal variant and secondly, the causal variants at most associated loci are not coding, instead acting through gene regulatory mechanisms which are difficult to determine (Weeks et al. 2020). Validation of our results is still ongoing, through the development of new statistical methods and the cross-validation of our findings against experimental datasets comprising expression QTL results.

CONCLUSIONS

This study shows that clustering variants by their patterns of effects and combining the results with pathway analysis may help to elucidate the underlying genes and biological processes which link genetically associated traits.

ACKNOWLEDGEMENTS

The ideas and assistance of colleagues, including Ruidong Xiang and Ed Breen, have been invaluable for this project. We would also like to thank DataGene and the herd managers who have provided data for our work. This project was supported through an Australian Government Research Training Program Scholarship, as well as financial contributions from the University of Melbourne and DairyBio (a joint initiative of Agriculture Victoria, Gardiner, and Dairy Australia.

REFERENCES

van den Berg, I., Xiang, R., Jenko, J., Pausch, H., Boussaha, M., Schrooten, C., Tribout, T., Gjuvsland, A.B., Boichard, D., Nordbø, Ø., Sanchez, M.-P., Goddard, M.E. (2020) *Genet. Sel. Evol*. **52**, 37.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., Zhang, B. (2019) *Nucleic Acids Research* **47**, 199.

Privé, F., Aschard, H., Ziyatdinov, A., Blum, M.G.B. (2018) *Bioinformatics* **34**, 2781.

Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, *et al*. (2020) medRxiv. doi:10.1101/2020.09.08.20190561

Xiang, R., Hayes, B.J., Vander Jagt, C.J., MacLeod, I.M., Khansefid, M., Bowman, P.J., Yuan, Z., Prowse-Wilkins, C.P., Reich, C.M., Mason, B.A., Garner, J.B., Marett, L.C., Chen, Y., Bolormaa, S., Daetwyler, H.D., Chamberlain, A.J., Goddard, M.E. (2018) *BMC Genomics* **19**, 521.

Xiang, R., MacLeod, I.M., Daetwyler, H.D., de Jong, G., O'Connor, E., Schrooten, C., Chamberlain, A.J., Goddard, M.E. (2021) *Nat Commun* **12**, 860.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M. (2011) *The American Journal of Human Genetics* **88**, 76.