THE MOLECULAR CATTLE FOR PRECISION BREEDING

R. Xiang, M.E. Goddard, A.J. Chamberlain and J.E. Pryce

Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, VIC, 3083 Australia

SUMMARY

Genomic selection has revolutionised livestock breeding. The success of genomic selection is attributed to the polygenic architecture of most economically important traits and the small effective population size in livestock. The latter results in high linkage equilibrium (LD) between SNP markers and causal mutations, which allows genomic selection to work without knowing the causal mutations or biological mechanisms behind phenotypes. However, emerging studies have shown that the knowledge of causal mutations can be used to improve genomic selection. Also, genomic selection cannot directly remove some large-effect mutations with undesired phenotypic effects. With a potential increasing number of gene-edited animal products entering the food chain, the knowledge of causal mutations may warrant future interventions. Causal mutations are challenging to identify. However, with increasingly advanced statistical methods, molecular techniques and the availability of unique or large populations/samples, identifying causal mutations has become possible. For example, causal mutations may have significant effects on transcriptomics, epigenomics and proteomics, before they eventually affect phenotypes. Therefore, fine-mapping the effects of sequence variants on these biological cascades can break LD to inform causal mutations. We propose the concept of Molecular Cattle, describing the effort in developing a deep phenotyped cattle cohort to enhance causal variant discovery which can improve both polygenic genomic selection and monogenic interventions. We propose this concept in this paper and invite collaborations.

INTRODUCTION

The genetic or breeding value of animals can be predicted using genome-wide SNP markers, and this approach is called genomic prediction or selection (Meuwissen *et al*. 2001). Genomic selection has been successful in improving animal breeding. Genomic selection works without the need to know the causal mutations. This is due to the low effective population size in most livestock species where many SNP markers well tag causal variants via LD. However, the predictions performed in such a way have accuracies far away from perfect. Further, as LD structure differs dramatically between populations, genomic prediction accuracy further declines when the training and validation populations are genetically distant. Therefore, if we know the causal mutations and use them in the genomic prediction, the accuracy of genomic prediction is expected to increase and be better maintained across populations.

While livestock genomic research has a long history, few causal mutations underlying quantitative trait loci (QTLs) have been mapped and confirmed. One of the most famous QTL in dairy cattle for milk traits is *DGAT1* which explains more than 30% of genetic variations in milk fat. However, the causal mutations underlying this QTL are still being debated (Grisart *et al*. 2004; Fink *et al*. 2020). Another more recently identified QTL with undesirable effects on milk yield, fertility and survival is on chromosome 18 (e.g. *CTU1* (Xiang *et al*. 2017)). The identification of causal mutations could be used for intervention, such as gene editing, a technique that has been heavily regulated in animals (Solomon 2020). However, in recent years, there has been an increasing amount of research on genetically edited animal products entering the food chain worldwide (Ledesma 2024). Ideally, before the wet-lab experiments, one needs to know which candidate mutations among 10s of millions of sequence variants are to be edited.

Causal mutations affect phenotypes via their biological pathways related to gene expression, regulation, epigenetic modification and protein translation. Therefore, by profiling multi-omics and phenotypes in a large cohort, one can portray the effects of causal mutations. As there are 10s of millions of sequence variants in the animal genome, fine-mapping, i.e., joint analysis of all variants, should be conducted. For example, a Bayesian genome-wide fine-mapping using prior information based on multi-omics data has identified a set of potentially causal mutations in dairy cattle (Xiang *et al*. 2021; Xiang *et al*. 2023). The main challenge of genome-wide fine mapping is the requirement of computing power for ever-growing numbers of animals and sequence variants. Therefore, we propose the concept of Molecular Cattle, the proposal of developing a deep phenotyped dairy cattle cohort that can be used to fine-map causal variants. This choice of dairy cattle is due to our existing access to large datasets and experienced research farms, but this concept can be applied to any cattle breed. This effort will allow us to identify causal mutations that can be used for both improving genomic prediction and genetic intervention.

PROPOSED POPULATION AND ASSAYS

We propose to start to build this cohort using the existing dairy cow population, with a starting sample size of 1,000 (ideally 10,000). This considers good phenotypic records in dairy cattle, statistical power, and sequencing experiments' expenses. We expected this cohort to have conventional phenotypes already, such as milk production, fertility, mid-infrared spectroscopy (MIR) and if possible, methane emissions. All environmental and management exposures should be recorded as precisely as possible. We also expect this cohort to have genotype data which will be imputed to full sequence. We then deep phenotype this cohort with transcriptomics (RNA-seq from blood), epigenomics (DNA methylation from blood), proteomics (mass-spectrum from blood), rumen microbiome (metagenome-sequencing from rumen), metabolomics (LCMS from blood), and clinical biochemistry markers (e.g., Albumin, glucose and etc). As a collaboration effort, we also propose that partners join the effort with all data from multi-omics experiments from collaborators that can be shared openly. Depending on the privacy restrictions of different parties, one can decide to share either raw or summary data. One key hypothesis of our study is that variants with consistent effects on phenotypes across different environmental conditions are more likely to be causal. We propose to conduct a multi-trait Bayesian genome-wide fine-mapping on this dataset to prioritise causative mutation based on BayesR3 (Breen *et al*. 2022), with a few additional functions to be developed to suit this particular dataset. We will use massively parallel reporter assays (MPRA) (Cooper *et al*. 2022) to confirm these statistically finely mapped variants.

DISCUSSION

We use the following graph to illustrate the concept of Molecular Cattle (Figure 1), where we aim to identify causal mutations that can inform both genomic selection and intervention (e.g., gene editing). One key difference of our proposal from previously proposed phenomics is the inclusion of the larger number of molecular phenotypes (hence the molecular cattle). As described above, the effect of variants on different multi-omics data can be fine-mapped using advanced statistical models. The fine mapping result of all variants analysed can be treated as a variant ranking. Such ranking of variants can be used in a different dataset as weights to improve genomic prediction. This task can be done via either GBLUP (Meuwissen *et al*. 2024) or BayesRC (MacLeod *et al*. 2016). One challenge of this analysis is the strong LD between variants that prevents precise mapping of the causal variants. We expect this challenge to be resolved using the data from diverse multi-omics data because these data are independent of LD and can be used to inform biological pathways where causal mutations act. A starting point for intervention could be finely mapped causal variants behind known loci, as summarised in Table 1.

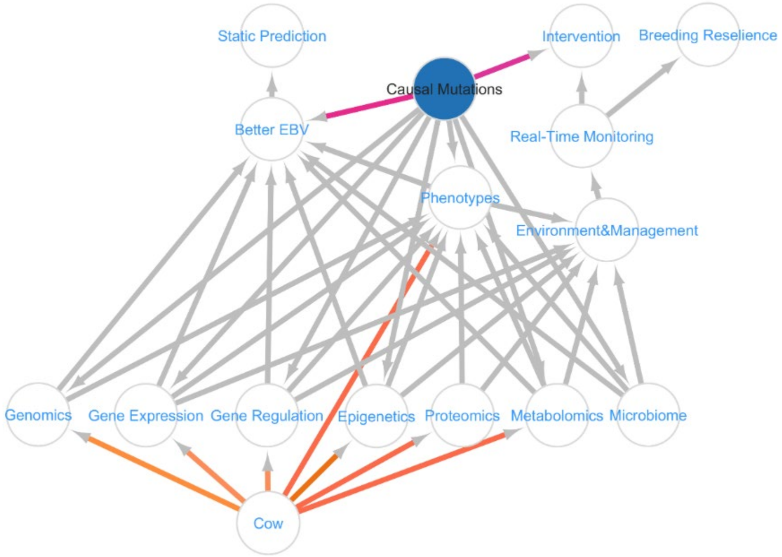

Figure 1. Illustration of the concept of Molecular Cattle. Arrows indicate the direction of data flow. Orange arrows indicate the sequencing experiments; grey arrow indicates the data analysis while the purple arrows indicate the use of causal mutations

Table 1. Examples of some known candidate mutations behind cattle QTLs to be intervened

Loci	Candidate mutation (ARS-UCD1.2)	Phenotypic effects	Molecular phenotype
DGAT1	Chr14:611019-20	Fat	eQTL, sQTL
GHR	Chr20:31888449	Milk production	Conserved across 100 species
CTU1	Chr18:57062518	Survival and fertility	Missense, conserved and regulatory
MGST1	Chr5:93516066	Milk fat yield	eQTL
GC	Chr6:86949653(starting position)	Mastitis	12-kb structure variant

Before taking causal variants to *in vivo* experiments, we expect to validate these regulatory variants using MPRA. MPRAs offer a flexible high-throughput framework to study elements regulating transcription and posttranscriptional events. MPRAs are widely used in humans to verify causal variants identified from GWASs (Cooper *et al*. 2022). This technique is rarely used in animals, likely due to its high costs. However, if wet-lab interventions are considered in the downstream work, it would be necessary to use MPRA to confirm identified causal variants.

Apart from improving genomic selection and providing targets for intervention, the results of this Molecular Cattle can be used to improve animal breeding in several areas that require further development, including real-time management of animals, and quantifying effects of GxE. Because DNA does not change over the lifespan of animals, the breeding values estimated from them represent a static prediction based on the genetic merit of animals. While EBV helps farmers make decisions at an early stage, real-time monitoring of cattle based on informative biomarkers could be more useful when intervention is needed. In this case, the creation of the Molecular Cattle dataset can be used to identify useful biomarkers for real-time monitoring of cattle health (e.g. mastitis onset). One future challenge is the expense of life-time assay of such biomarkers in herds. One potential solution is to take samples from milk cells instead of from blood, as they have shown similar properties in terms of the identification of regulatory variants (Xiang *et al*. 2018). Another solution is to create predicted biomarkers using more easily obtained phenotypes such as MIR.

The lack of high-throughput measurements of environmental exposures, partly due to multiple measurements in non-dairy livestock or on traits that are not automatically recorded, has prevented the proper model of GxE in animal breeding. While the host DNA does not change, how, when and where this DNA information should be expressed depends on environmental factors. Therefore, the multi-omics data, including DNA methylation (Clarke *et al*. 2021) in well-designed experiments, are important for inferring environmental variations that could be useful to inform GxE analysis.

CONCLUSION

We propose the concept of Molecular Cattle as a future research direction, where we could discover causal variants to improve both genomic selection and genetic intervention. Other benefits included better real-time management of herds and a better understanding of GxE factors. We also propose this effort as a national and international collaboration so we can achieve a larger sample size and better breed and environmental diversity for this experiment. We see this process starting from the Oceania regions with the potential to expand across the globe.

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) *Comm. Biol*. **5**: 661.

Clarke S., Caulton A., McRae K., Brauning R., Couldrey C. and Dodds K. (2021). *Anim. Front*. **11**: 90.

Cooper Y.A., Teyssier N., Dräger N.M., Guo Q., Davis J.E., Sattler S.M., Yang Z., Patel A., Wu S. and Kosuri S. (2022) *Science* **377**: eabi8654.

Fink T., Lopdell T.J., Tiplady K., Handley R., Johnson T.J., Spelman R.J., Davis S.R., Snell R.G. and Littlejohn M.D. (2020) *BMC Genomics* **21**: 1.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J.-J., Kvasz A., Mni M., Simon P., Frere J.-M. and Coppieters W. (2004) *Proc. Nat. Acad. Sci.* **101**: 2398.

Ledesma A.V. and Van Eenennaam A.L. (2024) Vet. J. **305**: 106142.

MacLeod I.M., Bowman P., Vander Jagt C., Haile-Mariam M., Kemper K., Chamberlain A., Schrooten C., Hayes B.J. and Goddard M. (2016) *BMC Genomics* **17**: 1.

Meuwissen T., Eikje L.S. and Gjuvsland A.B. (2024) *Genet. Sel. Evol*. **56**: 17.

Meuwissen, T.H., Hayes B.J. and Goddard M. (2001) *Genetics* **157**: 1819.

Solomon S.M. (2020) *Nat. Biotechnol*. **38**: 142.

Xiang R., Fang L., Liu S., Macleod I.M., *et al.* (2023) *Cell Genomics* **3**.

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. and Mason B.A. (2018) *BMC Genomics* **19**: 1.

Xiang R., MacLeod I.M., Bolormaa S. and Goddard M.E. (2017) *Sci. Rep-UK* **7**: 9248.

Xiang R., MacLeod I.M., Daetwyler H.D., de Jong G., O'Connor E., Schrooten C., Chamberlain A.J. and Goddard M.E. (2021) *Nat. Comms.* **12**: 1.