Developing DNA to function AI models to improve cattle traits

Tanmay Debnath[1], Chumeng Zhu[1], Rachel Owen[1], Jess Powell[1], Rongrong Zhao[1], Lindsey Plenderleith[1], Musa Hassan[1], Liam Morrison[1], Tim Connelley[1], James Prendergast[1,]*

[1] Roslin Institute, The Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh, UK
* Corresponding author: James.Prendergast@ed.ac.uk

Resumen

Cattle productivity has increased dramatically through genomic selection. However, the rate of gains can be constrained by incomplete knowledge of functional variation in the cattle genome. Without precise identification of the causal variants underlying important traits, genomic estimated breeding values can suffer from low signal-to-noise ratios, reducing their accuracy and portability across populations and generations. Similarly, although combining selective breeding with targeted genome editing has long been recognised as a route to accelerate genetic gain, its practical application remains limited by the scarcity of confidently identified functional variants.

This talk presents our work integrating large-scale functional genomics with AI-based sequence-to-function models in an attempt to address this challenge. More than 80% of functional variants influencing complex traits are thought to act by altering gene expression. Sequence-to-function models, that predict the regulatory consequences of DNA sequence variation, therefore offer a powerful potential route to prioritising functional regulatory variants in livestock genomes. However, training such models requires high-resolution, large-scale datasets linking DNA sequence to transcriptional activity. Such resources are abundant in humans but remain limited in livestock species.

To help fill this gap, we have generated two complementary cattle datasets. First, using massively parallel reporter assays, we assayed the regulatory potential of nearly 1.5 billion cattle DNA fragments and tested more than 15 million variants. This identified over 150,000 variants associated with changes in transcriptional activity. These expression-modulating variants enable fine-mapping of eQTL and GWAS loci to candidate causal variants, including the resolution of loci in perfect linkage disequilibrium and the identification of functional structural variants. We demonstrate that these data can be used to train an effective deep-learning sequence-to-function model capable of predicting regulatory activity directly from DNA sequence.

Second, with a focus on infectious disease traits, we generated novel PRO-cap data across key bovine immune cell types, including B cells, CD4+ and CD8+ T cells, γδ T cells, and monocytes. These data capture transcription initiation at base-pair resolution. Using the ProCapNet architecture, we trained a cattle-specific model that predicts both activity and read profiles at transcription start sites, with performance broadly comparable to equivalent human models. Model predictions show strong concordance with the regulatory activity of more than 8,000 experimentally validated variants.

Together, these datasets and models demonstrate the potential of bespoke AI approaches to improve functional variant discovery in cattle. By refining the identification of causal variants underlying productivity and health traits, this work provides a foundation for more accurate genomic selection and future genome-editing strategies to accelerate genetic gain.

Keywords: Cattle genomics, AI models, functional regulatory variation, MPRA, PROCap

Agradecimientos: This work was funded by the BBSRC, Gates Foundation and Roslin Foundation.