

Development and Deployment of a Computation Tool for Efficient Whole-Genome Sequence Association and Prediction Analysis

Institution: University of Alberta

Primary Investigator: Graham Plastow

Term: 2020 - 2021

Funding: \$120,000 from Genome Alberta

Background: The latest run of the 1000 bull genomes project has identified around 151 million genetic variants across 171 cattle breeds. Variants have been annotated for their effects and impacts on known genes. This vast amount of information can potentially be used to facilitate the discovery of causal mutations and to greatly improve the accuracy of genomic prediction for economically important traits in beef cattle. The main challenge is to develop statistically powerful and computationally efficient methods for large-scale whole-genome-sequence-based analyses. Recently, Livestock Gentec has imputed its legacy genotypes on about 25,000 beef individuals to whole-genome sequence data.

Goal: to improve the efficiency and accuracy of whole-genome sequence analysis and prediction

Objectives: to develop a powerful and efficient computing algorithm for whole-genome sequence association and prediction analyses

- 1) Development of a sampling algorithm that integrates whole-genome sequence variants and annotations for genome-wide association and prediction analyses
- 2) Evaluation of the novel algorithm for accuracy and efficiency using data simulations
- 3) Development of a cloud-based platform that automatically runs the algorithm for data processing, analysis and summary report

Benefit: this project will provide the Alberta beef industry and research institutions with a powerful tool for fast integration of sequence information into genomic research and applications. It is also expected to improve the accuracy of genomic prediction compared to the current array-based methods from using the whole genome sequence variants and their genome annotation information in the prediction.