Discovering the missing variation: a long read sequencing study into the structural variation in two dairy breeds

*Tuan Nguyen *, Jianghui Wang, Amanda Chamberlain, Iona MacLeod*

Structural variants (SVs) play a substantial role in the evolution of species and have an impact on Mendelian and quantitative traits in mammals. However, SVs have been challenging to accurately identify and genotype at population scale using short-read sequencing. Long-read sequencing technologies are becoming competitively priced and can address several of the disadvantages of short-read for the discovery and genotyping of SVs. In this study, we ran a pilot population scale study to catalogue the SVs in a cohort of Holstein & Jersey cattle.

Introduction

- Structural variants (SVs) are genomic variations that involve a large segment of DNA, typically larger than 50bp.
- SVs have been difficult to accurately characterise with short-read sequencing but are now becoming accessible with developments in long-read sequencing. In a long-read human population-scale study, individuals were found to have a median of ~22,000 SV [1] and a bovine pan-genome study estimated that more than 70Mb of sequence is not represented on the cattle reference genome from a single individual [2].
- However, no bovine long-read population-scale studies have been published. We therefore undertook a pilot long-read sequence study to identify SVs in 40 cattle.

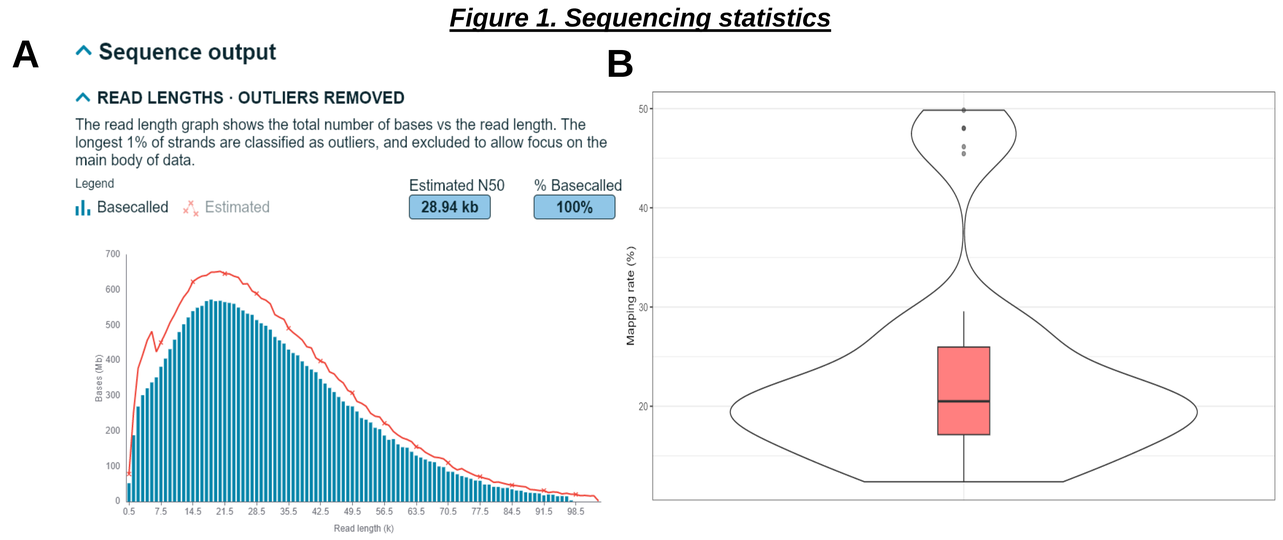
Methodology

- 20 Holstein & 20 Jersey individuals were sequenced with ONT PromethION (flow cell 9.4.1 & 10.4) with various coverages. Illustration of one sequenced individual in Fig. 1A, and boxplot illustrate reads & mapping rate of 40 individuals on Fig. 1B.
- A scalable, reproducible Nextflow pipeline for this study are now freely available at *https://github.com/tuannguyen8390/AgVic_CLRC*

Objectives

- Develop a portable & scalable solution and characterize the bovine structure variants catalogue across two breeds, Holstein & Jersey.

Figure 1. Sequencing statistics



Results

- Our Nextflow pipeline enables deployment of a reproducible, scalable workflow using on-premises or on-the-cloud systems (Figure 2)
- We discover a large number of SVs with insertions & deletions being most numerous, followed by inversion and duplication (Figure 3)
- We identify SVs that are shared between the 2 breeds & others unique to one breed (Figure 4A). We also show the relationship between allele frequency and read length (Figure 4B).
- A PCA using SV genotypes showed clear separation of breeds (Figure 5)

Figure 2. nf-VALOR Pipeline

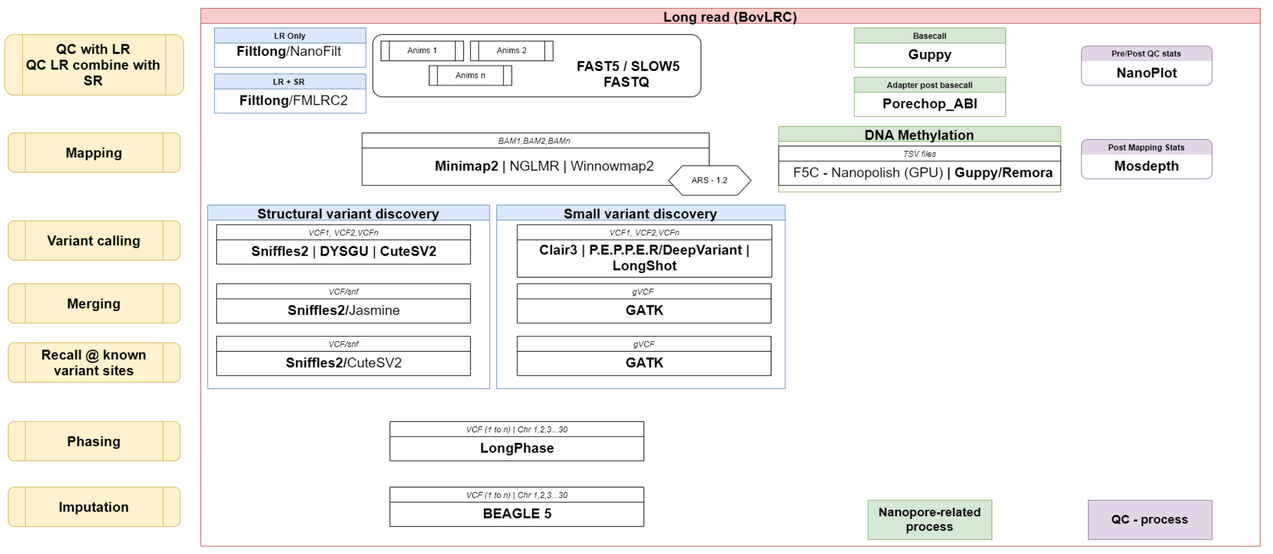


Figure 3. Summary statistics of SVs discovery

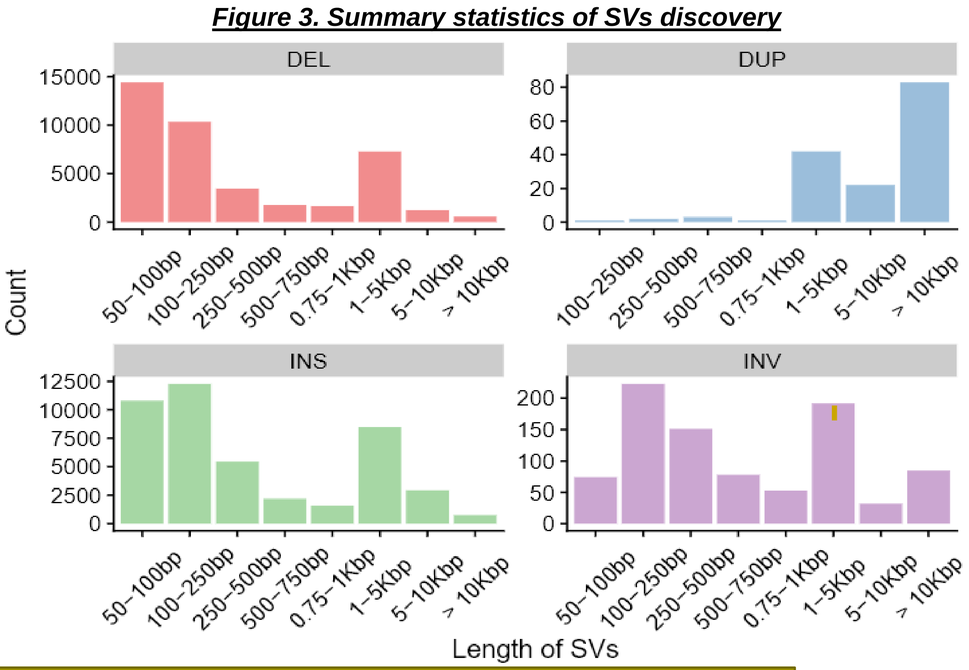
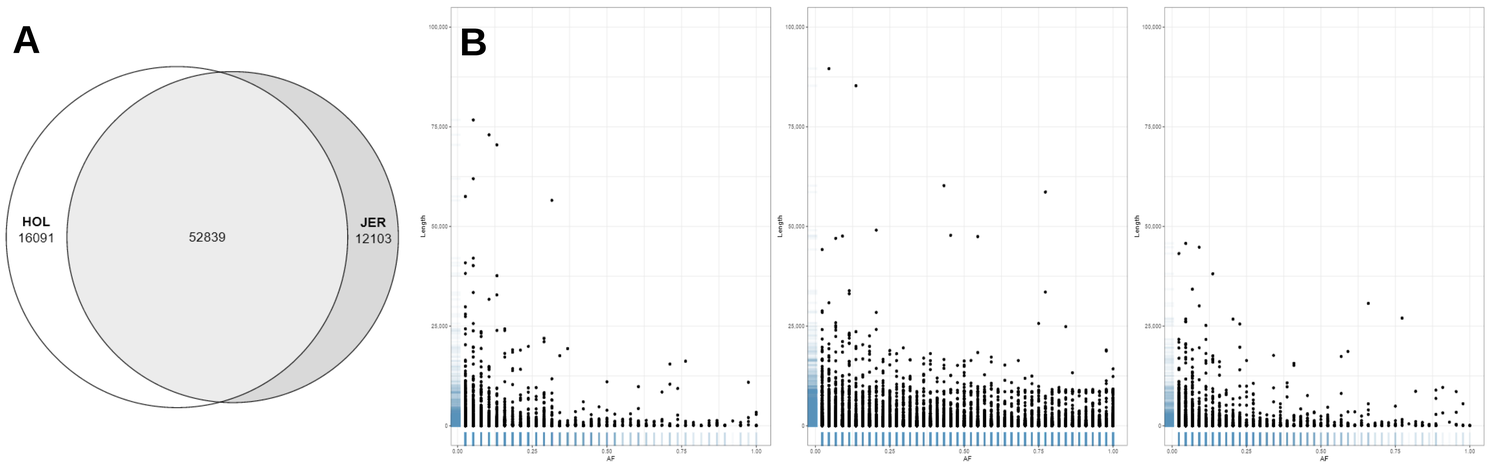


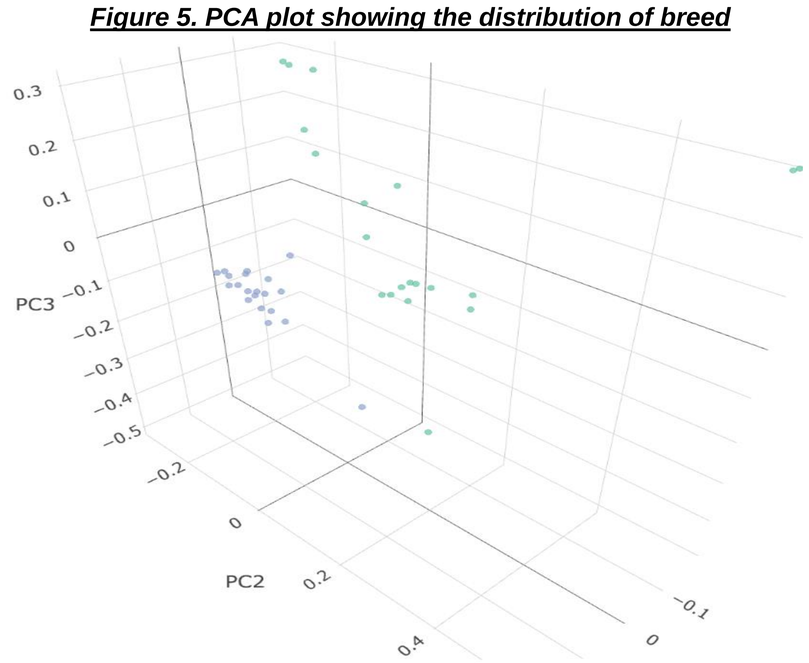
Figure 4. Shared/Unique SVs and relationship between allele frequency & SV length



Discussion

- Our results demonstrate a consistent trend of larger structural variants (SVs) with lower allele frequencies in the population, but this trend much more evident in the breed-specific SVs. We expect many of these longer & rare SV arose after breed formation and potentially have more severe effects on the phenotype.
- The target SVs that are shared across breeds, do not show this trend as strongly, suggesting that these shared SVs perhaps less likely to have a negative impact on the fitness of individuals.
- We have identified multiple SVs that span functional genes, highlighting them as priority targets for further downstream analysis.
- As more animals are included in future studies, it will be possible to identify SVs that are characteristic of specific breeds and to undertake population-scale studies on the impact of SV on important traits in livestock [3].

Figure 5. PCA plot showing the distribution of breed



Conclusions

- We have developed a pipeline for detecting structural variations (SVs) using long-read sequencing, and have established the Bovine Long Read Consortium (BovLRC) to facilitate global collaboration on population-scale studies of SVs.
- Our initial studies have explored SVs differences between Holstein and Jersey cattle.
- We are also in the process of sequencing more animals with diverse phenotype profiles.

1. Beyter, Doruk, et al. Nature Genetics 53.6 (2021): 779-786.
2. Leonard, Alexander S., et al. Nature communications 13.1 (2022): 3012.
3. Nguyen, Tuan V., et al. " Genetics Selection Evolution 55.1 (2023): 9.

Contact: tuan.nguyen@agriculture.vic.gov.au