INSIGHTS FROM POPULATION-SCALE LONG-READ SEQUENCING: STRUCTURAL VARIANT CHARACTERISATION AND ANNOTATION

T. Nguyen[1], J. Wang[1], A. Chamberlain[1,2] and I. Macleod[1,2]

[1] Agriculture Victoria, Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia

SUMMARY

This study presents an analysis of structural variants (SVs) in 108 dairy cattle (50 Holstein and 58 Jersey) using high quality long-read Oxford Nanopore Technologies sequence data. Analysis of SV size distribution across allele frequencies revealed distinct evolutionary patterns, with larger SVs predominantly occurring at lower frequencies. A larger proportion of the longer insertions show low allele frequency compared to deletions, suggesting possible differential selective pressures and potentially because long insertions are harder to identify than deletions. The study validated large breed-specific known SV including HH0, HH5, and POLL, demonstrating the reliability of our approach. In investigating the relationship between the severity of SV impact, SV allele frequency and length, we found that the larger the variant the more likely the mutation (alternate allele) had a predicted high impact. Furthermore, among deletions, the longer and less common variants showed a larger proportion of predicted high impact effects compared to more common SV. This population-scale long-read reference dataset provides valuable new insights into the genomic architecture of dairy cattle and establishes a foundation for understanding the functional impact of SVs in cattle breeding.

INTRODUCTION

Although there have been limited studies of structural variants (SVs) in livestock, it is likely that they play a crucial role in the genetic diversity and adaptability of livestock (Nguyen *et al.* 2023). SV are defined as variants longer than 50 bp of DNA sequence: this includes deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations. Several recent cattle studies have identified SV that cause embryonic lethality (Schütz *et al.* 2016), influence physical traits (Rothammer *et al.* 2014), and disease resistance (Lee *et al.* 2021). Therefore, gaining a better understanding of SVs at population scale is essential for breeding programs aimed at improving livestock because it is likely that on average SV will have a higher impact on gene function and regulation compared to single nucleotide polymorphisms (SNPs) (Lee *et al.* 2023). Long-read sequencing technologies have emerged as a powerful tool for population-scale studies of SVs compared to traditional short-read sequencing that struggles to accurately resolve complex genomic regions due to the limited read length. Long-read sequencing provides longer contiguous sequences of DNA, which enhances the ability to identify and characterize SVs with greater precision (Amarasinghe *et al.* 2020). A long-read reference population is crucial to study the SV diversity within and between cattle breeds, and to enable research to identify SVs that may impact economically important traits.

MATERIALS AND METHODS

DNA sequencing. 50 Holstein and 58 Jersey animals were selected (total = 108), avoiding full and half sib relationships to maximise diversity. High molecular weight DNA was extracted from semen, liver tissue or whole blood using Gentra Puregene DNA extraction kit (Qiagen). Sequencing libraries were prepared using ligation sequencing kit V9 or V10 or 14 (Oxford Nanopore Technologies) according to manufacturer's instructions and sequenced on R9.4.1 or R10.4.1

flowcells on PromethION P24 (Oxford Nanopore Technologies). Super high accuracy basecalling was performed with either Guppy v6.1.7 or Dorado v0.7.0.

Data analysis. Reads were quality trimmed using FiltLong (https://github.com/rrwick/Filtlong accessed December 2022) with default settings. Filtered reads were then mapped to ARS-UCD2.0 reference genome (Rosen *et al*. 2020) using Minimap2 (Li 2018). For discovery and genotyping of SV, Sniffles2 v.2.5.3 (Sedlazeck *et al.* 2018) was used to detect SVs for each sample and subsequently to merge and re-genotype SVs from all individuals. We only considered SVs on chromosome 1-29 of less than 3 Mb length, and a minor allele count > 0. Genotypes with quality score < 5 were set to missing and variants with > 20% missing genotypes removed. Variant Effect Predictor (VEP) software (McLaren *et al,* 2016) was deployed to annotate the SVs using Ensembl cattle annotation release 112 (exported from ARS-UCD1.3), having identical autosomal coordinates to ARS-UCD2.0).

RESULTS AND DISCUSSION

Across all Holstein animals, the mean and median raw sequence read coverage was 25.2X and 24.9X, respectively, with read length N50 values of 26,184bp (mean) and 23,666bp (median). The Jersey samples yielded comparable sequencing metrics, with mean and median coverage of 22.7X and 24.6X, respectively, and read N50 values of 22,366bp (mean) and 20,830bp (median). These sequencing metrics indicate high-quality data suitable for downstream genomic analyses.

Four distinct types of SVs were identified across the joint genotyping of all 108 sequences in Figure 1A. INS were the most abundant of the total SVs (55.5%), followed closely by DEL (45.3%). INV were much less frequent while DUP were the least common. This relative abundance of SV types is in keeping with population scale studies in humans (Collins *et al*. 2020). Here we focus mainly on results for INS and DEL. The length distribution of INS and DEL variants (Figure 1B) revealed that both were most common in the shorter length ranges (<10kb), while SV > 20kb were much less common. Additionally, a peak for both DEL and INS was observed between 100 – 200 bp, and a strong INS peak between 7,000 to 9,000 bp. These peaks likely represent different types of transposable elements and have been reported in cattle, human and mouse studies (Collins *et al*. 2020; Ferraj *et al*. 2023).

Using this dataset, we confirmed the presence of several known SVs, including two Holstein specific embryonic lethal DEL, Brachyspina/HH0 (Charlier *et al*. 2012) and HH5 (Schütz *et al*. 2016), as well as a known DUP which causes the so called “Friesian” polled (POLL) phenotype (Rothammer *et al*. 2014). The aligned sequence reads across the HH0 deletion region are visualized in Figure 1C, for one animal genotyped as 0/1 (carrier) and one free of the deletion (0/0). No other Holstein or Jersey samples were found to carry the deletion. In addition to HH0, two Holstein carriers of the 138 kb HH5 DEL were identified. As seen in Figure 1D, the HH5 DEL is more difficult to directly visualise because it is much longer than most of the sequenced reads. Finally, in Figure 1E, we confirmed a known POLL carrier (Jersey bull).

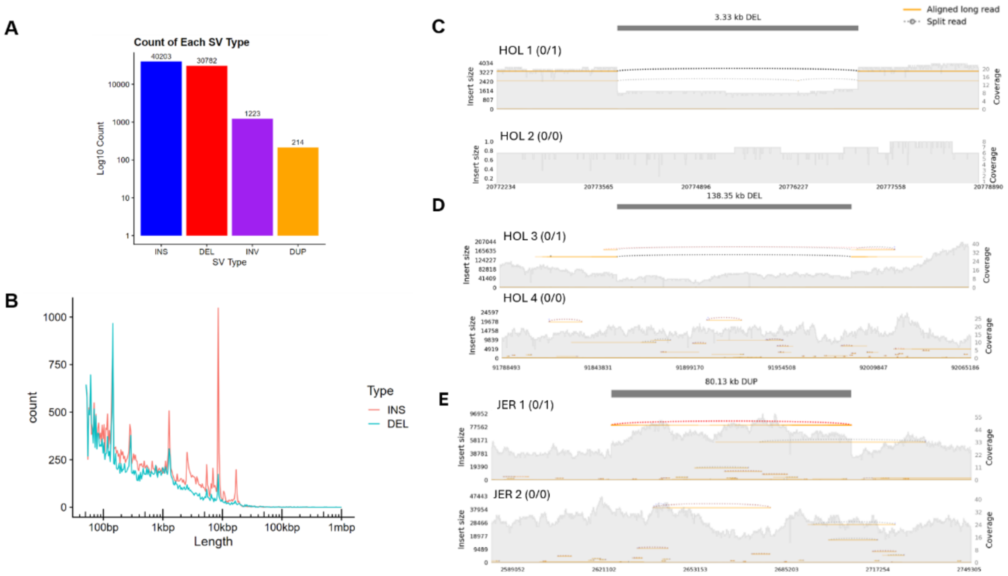

Figure 1. Structural variant discovery in 108 cattle: (A) distribution of types (INS: insertions; DEL: deletions; INV: inversions; DUP: duplications) and (B) length distribution of INS and DEL. Samplot coverage profiles visualising aligned reads for individual animals across the breakpoints for known recessive lethal DEL HH0 (C), HH5 (D) and POLL DUP (E) (carrier genotypes labelled 0/1 and homozygous reference as 0/0). The Y-axes show estimated length of split reads "insert size" (left) and the read coverage (right). The x axis is the position along the chromosome.

Using all discovered DEL and INS, the relationship between SV size, alternate allele frequency (AF) distribution, and VEP predicted "HIGH" functional impact of SV on genes was investigated to gain insights into possible selection pressures (Figure 2). Within each length category in Fig 2, the SV with low AF (<0.15) were more common than higher allele frequencies (AF>0.15) and overall ~40% of DEL and INS fell in the <1 kb with AF < 0.15 category. Additionally, as SV length increased, the proportion of low frequency variants of each length category increased e.g. 57% of all INS < 1kb length were low AF, while 97% of all INS > 10kb length were low AF, and similarly for DEL the respective proportions were 52% and 75%. This potentially indicates that there has been negative selection pressure operating on longer SV because they are more likely to be deleterious while the smaller variants may be better tolerated and less likely to negatively affect fitness of the animal. Among DEL, there was further support for this hypothesis because we observed strong enrichment for "HIGH" impact variants in the longer and low AF categories, particularly for SV > 10kb with AF < 0.15 (Fig 2: red colouring). Although this pattern was not observed for INS, it is important to note that several technical and biological factors could explain this: (i) VEP software may have inherent bias in predicting functional impacts of different SV types (e.g. predicting the result of a gene deletion is easier than predicting the impact of inserting a new piece of DNA into a gene), (ii) technical challenges in precise alignment of insertion events due to low coverage and/or bias of the reference genome, and (iii) the sequence content of insertions may require more complex interpretation frameworks for specific types of insertion, such as retrotransposons).

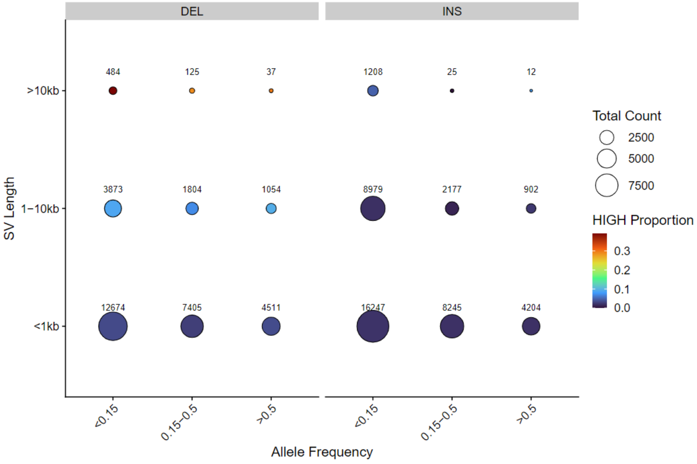

Figure 2. Proportion of Variant Effect Predictor HIGH impact deletions (DEL) and insertions (INS) by variant length and alternate allele frequency categories. Circle size indicates total variant count in each category and the number above each circle is the exact total. Colour gradient indicates the proportion of variants (0.0 to 0.4) in each category that were predicted HIGH functional impact. The Y-axis shows SV length in three bins and the X-axis displays three allele frequency bins.

CONCLUSION

This population-scale analysis of dairy cattle SVs using long-read sequencing revealed several key findings: 1) SV length distributions differed between INS and DEL , 2) a high proportion of long DEL predicted to have a HIGH functional impact were at low AF , 3) evolutionary constraints may be operating on larger SVs as evidenced by their low allele frequencies, and 4) known breed-specific variants were successfully detected. This reference dataset establishes a foundation for investigating the role of structural variation in livestock genetics and breeding applications.

ACKNOWLEDGEMENTS

The authors thank DairyBio, a joint venture project between Agriculture Victoria, Dairy Australia and The Gardiner Foundation, for funding.

REFERENCES

Amarasinghe S.L., Su S., Dong X., *et al.* (2020) *Genome Biol.* **21**: 30.
Charlier C., Agerholm J.S., Coppieters W., *et al.* (2012) *PLoS One* **7**: e4308.
Collins R.L., Brand H., Karczewski K.J., *et al.* (2020) *Nature* **581**: 444.
Ferraj A., Audano P.A., Balachandran P., *et al.* (2023) *Cell Genomics* **3**: 100291.
Lee Y.L., Bosse M., Takeda H., *et al.* (2023) *BMC Genomics* **24**: 225.
Lee Y.L., Takeda H., Costa Monteiro Moreira G., *et al.* (2021) *PLoS Genet.* **17**: e1009331.
Li H. (2018) *Bioinformatics* **34**: 3094.
McLaren W., Gil L., Hunt S.E., *et al.* (2016) *Genome Biol.* **17**: 122.
Nguyen T.V., Vander Jagt C.J., Wang J., *et al.* (2023) *Genet. Sel. Evol.* **55**: 9.
Rosen B.D., Bickhart D.M., Schnabel R.D., *et al.* (2020) *GigaScience* **9**: 1.
Rothammer S., Capitan A., Mullaart E., *et al.* (2014) *Genet. Sel. Evol.* **46**: 44.
Schütz E., Wehrhahn C., Wanjek M., *et al.* (2016) *PLoS ONE* **11**: e0154602.
Sedlazeck F.J., Rescheneder P., Smolka M., *et al.* (2018) *Nat. Methods* **15**: 461.