DISCOVERING THE MISSING VARIATION IN THE BOVINE GENOME; A LONG READ SEQUENCING PILOT STUDY INTO THE STRUCTURAL VARIATION IN TWO DAIRY BREEDS.

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

Structural variation has been posited to contribute equal or greater diversity at the nucleotide level than any other form of genetic variation. Short read sequencing technologies are limited in their ability to characterise structural variants (SVs), however long read sequencing, which is now cost effective, poses as a solution to this problem. The Bovine Long Read Consortium (BovineLRC) aims to use long read sequencing technologies to sequence cattle at population scale to characterise the structural variation of the bovine genome for downstream applications. This pilot study sequenced 41 animals from two breeds in an effort to understand how much SV variability exists within and across breeds. A total of 76,572 SVs were detected across all samples, one third of which were segregating in only one breed. Insertions and deletions tended to be smaller and duplications larger. Insertions and deletions more often segregated across both breeds, while inversions were more often breed specific. Few duplications were detected but they tended to be slightly more likely to be breed specific. The results highlight that it would be beneficial to have a dataset with large numbers of animals and breeds to understand the structural variation that exists and explore the impact of SVs on traits of interest.

INTRODUCTION

The 1,000 bull genomes project has had a massive impact on cattle genomics worldwide cataloguing single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs) in more than 6,000 cattle genomes (Daetwyler *et al.* 2014; Hayes, Daetwyler 2019). However, limitations of short read sequencing technologies mean that SVs are not easy to characterise. SVs can be large INDELs (>50 basepairs), inversions, translocations, copy number variations or segmental duplications and studies in human estimate that SVs together occupy a proportion of the genome that is equal to or greater than that of SNPs and small INDELs (Feuk *et al*. 2006; Ho *et al*. 2020) and contribute greater diversity at the nucleotide level than any other form of genetic variation (Chaisson *et al*. 2019). Multiple studies in cattle have demonstrated that SVs impact classic mendelian traits, quantitative traits and gene expression (Kadri *et al*. 2014; Rothammer *et al*. 2014; Lee *et al*. 2021).

Long read sequencing, such as nanopore sequencing from Oxford Nanopore Technologies (ONT) and single molecule real time sequencing from Pacific Biosciences (PacBio), have recently become cost-effective. Both claim costs of <$1,000US per genome at 30x coverage and have the advantage of being able to sequence across large SVs and therefore better characterise them compared to short read technology (Chaisson *et al.* 2019).

To date genome wide SV detection in cattle at population scale has largely used short read sequence data (Boussaha *et al*. 2015; Mesbah-Uddin *et al*. 2017; Mielczarek *et al*. 2018; Hu *et al*. 2020; Mei *et al.* 2020; Chen *et al.* 2021; Upadhyay *et al*. 2021) or limited long read sequence data (Low *et al*. 2020; Crysnanto *et al*. 2021) or a combination of the two (Couldrey *et al*. 2017). Like the Human Genome Structural Variation Consortium (Chaisson *et al*. 2019) the Bovine Long Read Consortium (BovineLRC) (Nguyen *et al*. 2023) aims to use long read sequencing technologies to sequence cattle at population scale to characterise the structural variation of the bovine genome.

Such a reference dataset will empower imputation of SVs into larger populations to examine their impact on quantitative traits as well as better resolve segmental duplication regions with copy number variants, understand the evolution of SVs and identify deleterious causal variants.

As a pilot study we have sequenced 41 animals from two breeds with ONT in an effort to understand how much variability exists within and across breeds in SV.

MATERIALS AND METHODS

DNA sequencing. 19 Holstein and 22 Jersey animals were selected, 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 kit (Qiagen). Sequencing libraries were prepared using ligation sequencing kit v9 or v10 (ONT) according to manufacturer's instructions and sequenced on R9.4.1 flowcells on a MinION or PromethION (ONT). Super high accuracy basecalling was undertaken with Guppy v6.1.7 and reads with q-score greater than 7 retained for analysis.

Data analysis. Reads were quality trimmed using FiltLong (https://github.com/rrwick/Filtlong accessed December 2022) with default settings and samples with short reads (6 Holstein and 6 Jersey, 150 cycle paired reads) polished. Filtered reads were then mapped to ARS-UCD1.2 (Rosen *et al*. 2020) with additional Btau5.0.1 Y (Bellott *et al*. 2014) using Minimap2 (Li 2018). Sniffles2 (Sedlazeck *et al.* 2018) was used to detect SVs for each sample and subsequently merge SVs from multiple individuals and re-genotype. SVs larger than 3Mb or with a genotype quality score less than 20 were excluded.

RESULTS AND DISCUSSION

A mean of 26x and 20x read coverage was achieved with mean read length N50 of 30kb and 26kb for Holstein and Jersey samples respectively. On average 20,770 deletions, 19,620 insertions, 234 inversions and 38 duplications were detected for each Holstein and 19,815, 18,458, 177 and 39 respectively for each Jersey. After merging and filtering data from all animals a total 76,572 SVs were detected. This is more than studies using short read data with similar sample numbers (Boussaha *et al*. 2015; Couldrey *et al*. 2017; Mesbah-Uddin *et al*. 2017; Mielczarek *et al.* 2018; Hu *et al*. 2020; Upadhyay *et al*. 2021) and similar to small studies using long read data in a pangenome approach (Crysnanto *et al*. 2021) but less than the largest pangenome approach with short read data and almost 900 samples (Zhou *et al*. 2022) which detected greater than 3.6 million SVs. 14,526 SVs were segregating in Holstein only and 11,264 only in Jersey (Figure 1A). 50,782 (66%) were detected in both breeds, therefore one third of all SVs were breed specific.

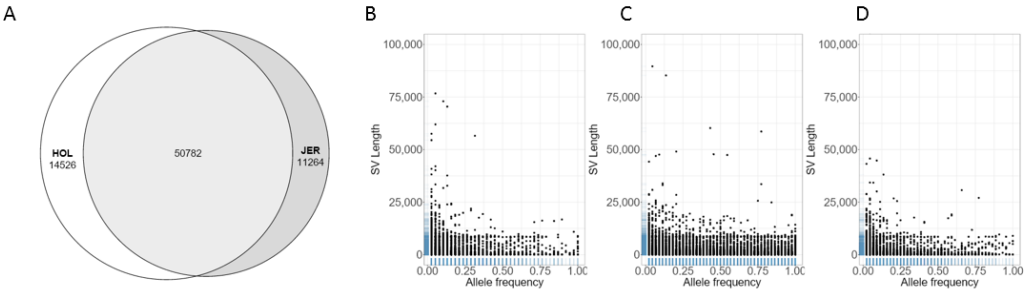

Figure 1. A Venn diagram showing SVs detected across or within Holstein (HOL) or Jersey (JER) breed (A). The relationship between allele frequency and length of SVs for Holstein specific SVs (B), those that occurred in both breeds (C) and Jersey specific SVs (D)

Figures 1B-1D show a trend of longer SVs with lower allele frequencies in the population, for both breed specific as well as across breed SVs. As expected, high allele frequency SVs were more likely across breeds. Other studies have estimated the proportion of breed specific SVs at 66% (Boussaha *et al* 2015) when comparing 3 breeds, 15% (Mielczarek *et al*. 2018) in 13 breeds, 48% (Hu *et al*. 2020) in 10 breeds, 54% (Mei *et al*. 2020) in 8 breeds and 76% (Low *et al*. 2020) in 3 breeds. While others found different allele frequencies of the same SV in different populations of taurine, indicus and zebu cattle (Upadhyay *et al.* 2021). This variation reflects the variable power of the different studies, driven by the numbers of samples, breeds included and breed definitions. Large numbers of samples and large numbers of breeds are likely required before we can be certain of the proportion of SVs that are breed specific.

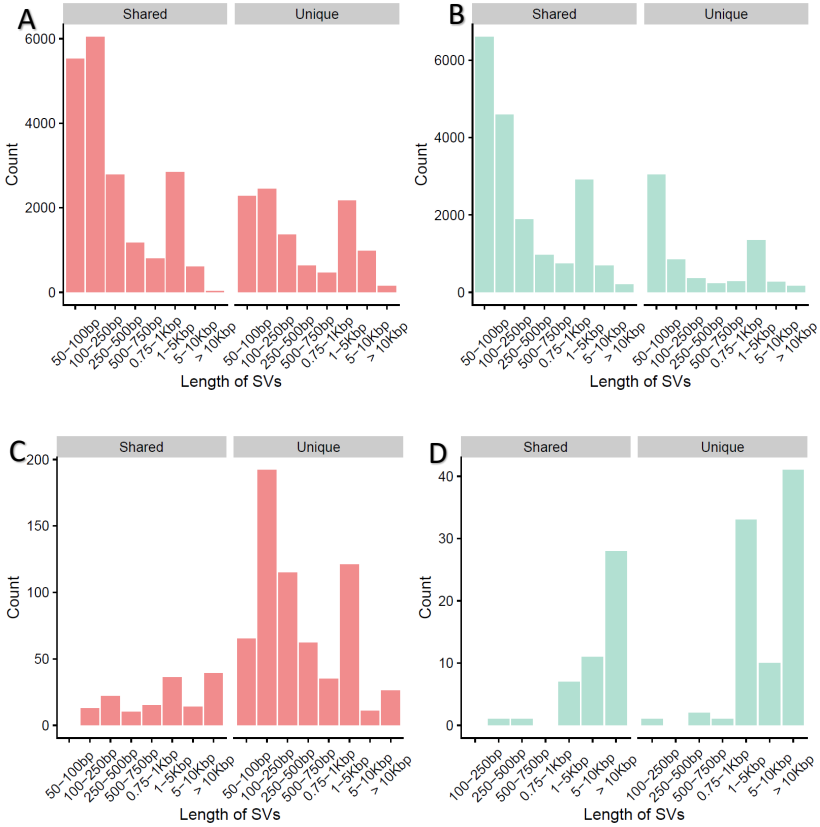

Figure 2. Numbers of deletions (A), insertions (B), inversions (C), and duplications (D) of different length for across breed (shared) and breed specific (unique) SVs. Note x-axis is not to scale

In agreement with other studies (Boussaha *et al*. 2015; Upadhyay *et al*. 2021; Zhou *et al*. 2022) insertions and deletions tended to be smaller and duplications larger (Figure 2). In this study we found insertions and deletions more often occurred across both breeds (Figure 2A and 2B), while inversions were much more often breed specific (Figure 2C). Few duplications were detected but they tended to be slightly more likely to be breed specific. However, reads that span the structural variant are required to call them accurately, therefore this dataset with read N50 of 26-30Kb has

limited power to detect very large SVs, likely partially accounting for the low numbers of duplicates found. Other studies also find lower numbers of large duplications compared with insertions and deletions (Mei *et al*. 2020; Zhou *et al*. 2022). It's also likely that many duplications were removed when SVs were merged across animals due to difficulty deciphering breakpoints for SVs. Given the small population size used here, read length N50 and the difficulties associated with accurate annotation of large and complex SVs this study had limited power to detect large and rare SVs.

CONCLUSION

This small pilot study in 2 breeds highlights that it would be beneficial to have a dataset with large numbers of animals and breeds to understand the structural variation that exists in the bovine genome. The BovineLRC has been formed to achieve this. It also highlights that more work is required to accurately annotate and genotype large and complex SVs. Further work is required to understand the impact of the SVs detected in this study on traits important to the dairy industry.

ACKNOWLEDGEMENTS

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

REFERENCES

Bellott D.W., Hughes J.F., Skaletsky H., *et al*. (2014) *Nature* **508:** 494.
Boussaha M., Esquerré D., Barbieri J., *et al*. (2015) *PLOS ONE* **10:** e0135931.
Chaisson M.J.P., Sanders A.D., Zhao X., *et al*. (2019) *Nature Communications* **10:** 1784.
Chen L., Pryce J.E., Hayes B.J., *et al*. (2021) *Animals* **11:** 541.
Couldrey C., Keehan M., Johnson T., *et al*. (2017) *J. Dairy Sci.* **100:** 5472.
Crysnanto D., Leonard A.S., Fang Z.-H., *et al*. (2021) *Proceedings of the National Academy of Sciences* **118**:e2101056118.
Daetwyler H.D., Capitan A., Pausch H., *et al*. (2014) *Nature Genet.* **46:** 858.
Feuk L., Carson A.R. and Scherer S.W. (2006) *Nature Reviews Genetics* **7:** 85.
Hayes B.J. and Daetwyler, H.D. (2019) *Annual Review of Animal Biosciences* **7:** 89.
Ho S.S., Urban A.E. and Mills R.E. (2020) *Nature Reviews Genetics* **21:** 171.
Hu Y., Xia H., Li M., *et al*. (2020) *BMC Genomics* **21:** 682.
Kadri N.K., Sahana G., Charlier C., *et al*. (2014) *PLoS Genet.* **10:** e1004049.
Lee Y.-L., Takeda H., Costa Monteiro Moreira G., *et al*. (2021) *PLoS Genet.* **17:** e1009331.
Li H. (2018) *Bioinformatics* **34:** 3094.
Low W.Y., Tearle R., Liu R., *et al*. (2020) *Nature Communications* **11:** 2071.
Mei C., Junjvlieke Z., Raza S.H.A., *et al*. (2020) *Genomics* **112:** 831.
Mesbah-Uddin M., Guldbrandtsen B., Iso-Touru T., *et al*. (2017) *DNA Research* **25:** 49.
Mielczarek M., Frąszczak M., Nicolazzi E., *et al*. (2018) *BMC Genomics* **19:** 410.
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.
Sedlazeck F.J., Rescheneder P., Smolka M., *et al*. (2018) *Nat Methods* **15:** 461.
Upadhyay M., Derks,M.F.L., Andersson G., *et al*. (2021) *Genomics* **113:** 3092.
Wick R. Flitlong. https://github.com/rrwick/Filtlong
Zhou Y., Yang,L., Han X., *et al*. (2022) *Genome Res* **32:** 1585.