THE FEASIBILITY OF AN IMPUTATION REFERENCE POPULATION FOR STRUCTURAL VARIATION IN CATTLE

A.J. Chamberlain[1,2], T.V. Nguyen[1], J. Wang[1], X. Wang[3], C.J. Vander Jagt[1] and I.M. 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
[3] Northwest A&F University, Yangling, China

SUMMARY

To understand the full impact of structural variation on traits important to cattle industries it would be desirable to generate a reference population for the imputation of large numbers of structural variants (SV) into existing populations with detailed phenotypic records for the traits of interest. This pilot study investigates two characteristics of SV that could impact their imputation; the precision of SV calling and their linkage disequilibrium (LD) with known single nucleotide polymorphisms (SNP) in the genome. Results indicate there are a number of SV that are called with low to zero standard deviation in length and starting position. Also, the pattern of LD between SV and SNP was similar to that between SNP and other SNP. Combined these results indicate that it is feasible to impute many SV into large existing populations with SNP genotypes which will enable exploration of their impact on traits important to cattle industries.

INTRODUCTION

Structural variants (SV) can be large insertions or deletions (INDEL, >50 base pairs), inversions, translocations, copy number variations or segmental duplications. Human studies estimate that SV together occupy a proportion of the genome that is equal to or greater than that of single nucleotide polymorphisms (SNP) and small INDEL (Feuk *et al.* 2006; Ho *et al.* 2020) and cause greater diversity at the nucleotide level than any other form of genetic variation (Chaisson *et al.* 2019). Multiple studies in cattle have demonstrated that SV impact classic mendelian traits, quantitative traits and gene expression (Kadri *et al.* 2014; Rothammer *et al.* 2014; Lee *et al.* 2021).

The full impact of SV on traits important to cattle industries has not yet been explored. It is therefore desirable to generate a large reference population that would enable imputation of SV into large populations with phenotypic records for the traits of interest, as has been done extensively with SNP and small INDEL from the 1000 Bull Genomes project (Hayes and Daetwyler 2019). To this end, the Bovine Long Read Consortium (BovLRC, Nguyen *et al.* 2023a) was formed to characterise structural variation at population scale as well as generate an imputation reference panel for SV. For this we developed a process (Figure 1) that was computationally efficient and scalable. It requires partners to pre-process sequences to smaller summary files using a custom built nextflow pipeline (Nguyen *et al.* 2023b), which standardises data processing and avoids sharing of large sequence or alignment files. Following joint calling of SV as well as SNP and then imputation of sporadic missing genotypes the reference database would be distributed to all partners for imputation into their own populations.

We previously assessed the mendelian consistency of SV called in cattle with the same pipeline used in this study (Nguyen *et al.* 2023b). The accuracy of SV imputation would depend on the accuracy of SV characterisation and genotype calling in the reference population. Genotype calling accuracy of SV is influenced by multiple factors; the aligner, genotyping method, depth of coverage, sequence data type, SV size and precision of breakpoints among others (Duan *et al.* 2022). Determining SV breakpoints is particularly challenging, and SV detection methods often generate SVs with imprecise breakpoints (Kosugi *et al.* 2019). Duan *et al.* (2022) showed that modifying SV breakpoints resulted in decreased F1 scores for five SV detection software, where F1 is the harmonic mean of precision and recall. They also found F1 scores varied with SV length. The degree of linkage disequilibrium (LD) between the SV and SNP in the reference population also impacts the ability to impute SV. Therefore, this pilot study investigates 1) the precision of SV breakpoints and 2) LD between SV and known SNP in two breeds to determine the feasibility of an imputation reference population for SV.

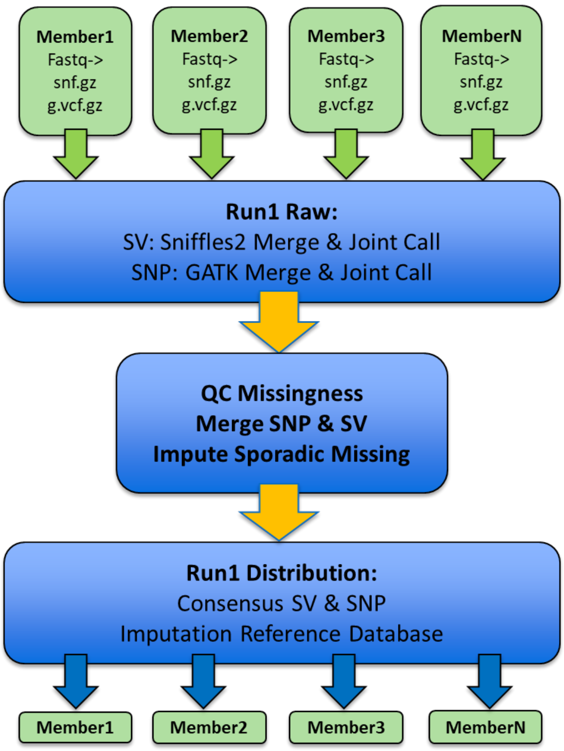

Figure 1. A process for the characterisation of structural variation in the bovine genome at population scale and the subsequent generation of an imputation reference panel for SV. Note fastq -> snf/g.vcf.gz is undertaken with nextflow pipeline nf-EXPLOR which uses Sniffles2 to call SV and Clare3 to call SNP within individuals. Green objects represent steps undertaken by participating organisations and blue those undertaken by Agriculture Victoria.

MATERIALS AND METHODS

DNA sequencing, read processing and genotype calling. 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 (Oxford Nanopore Technology, ONT) according to manufacturer's instructions and sequenced on R9.4.1 or R10.4.1 flowcells on a PromethION (ONT). Super high accuracy base calling was undertaken with Guppy v6.1.7 or Dorado v0.7.0. and reads with q-score greater than 7 retained for analysis. All samples were processed as outlined in Figure 1. Briefly nf-EXPLOR (https://github.com/tuannguyen8390/nf-EXPLOR) was used to trim poor quality bases from reads using FiltLong (https://github.com/rrwick/Filtlong), map filtered reads to the bovine genome ARS-UCD2.0 (Rosen *et al.* 2020) using Minimap2 (Li 2018) and detect SV with Sniffles2 v2.6.1 (Sedlazeck *et al.* 2018) for each individual. Sniffles2 was used to subsequently merge SV from all animals and re-genotype. SV larger than 3Mb were removed. Genotypes with quality score less than 5 were set to missing, then SV with greater than 20% missing genotypes were removed.

SV breakpoint analysis. Long read sequences from 108 animals, 50 Holstein and 58 Jersey, were generated and processed as outlined above. From the variant call format (VCF) file the standard deviation (SD) of the SV length and start position was extracted for 30,668 deletions and 37,486 insertions, which were then plotted in frequency histograms with bins of 20bp.

Linkage disequilibrium analysis. Long read sequences for 41 animals, 19 Holstein and 22 Jersey, were used to discover SV (Chamberlain *et al*. 2023) as outlined above. The insertions and deletions were then called in 93 Holstein and 105 Jersey with short read sequences, that had greater than 10x coverage, using Paragraph v2.0 (Chen *et al*. 2019). Short read sequences were publicly available and taken from Run 8 of the 1000 Bull Genomes Project. SV genotypes were then merged with SNP genotypes called from the same short read sequences. LD was calculated as r^2 between all pairs of SV (2,513) and SNP (972,276) on chromosome 1 with minor allele frequency > 0.01 within windows of 500kb using PLINK v1.9 (Chang *et al*. 2015). Similarly, LD was calculated between pairs of SNP however, only a random 10% of SNP were used for computational efficiency. LD decay was then plotted as the mean r^2 within bins of 1000bp.

RESULTS AND DISCUSSION

The accuracy of SV imputation would depend on the accuracy of SV characterisation and genotype calling in the reference population. Key characteristics of SV are their length and position in the genome, both of which are challenging to estimate because of the need to call SV within individuals and then merge and genotype across the population. For this reason, many software, including Sniffles2, estimate the standard deviation of length and position in merged datasets, where low standard deviation reflects consistent calling across the population. Figure 2 A-D show frequency histograms of the SD of length and starting position of 30,668 deletions and 37,486 insertions discovered in 108 animals Though there were SV with large SD the majority had very low to zero SD for both length and position. Based on the results of Duan *et al*. (2022) such SV are likely to be genotyped with high precision and recall, something further studies will confirm. This indicates that there are a significant number of SV that Sniffles v2.6.1, the variant detection software used, was able to call consistently, and likely accurately, within this population. It is these SV that will be good candidates to impute into larger populations.

LD between the SV and SNP in the reference population is required for an accurate imputation. It is common practice for large target populations, with detailed phenotypic records for the traits of interest, to be genotyped with low density (~50K) SNP panels which are subsequently imputed to full sequence using reference populations like the 1000 bull genomes dataset. Therefore, we assessed the LD between SV and SNP genotypes called from whole genome short read sequence data from Run 8 of the 1000 bull genomes project. Figure 2E shows the decay in LD with distance between SV and SNP was similar to that between SNP with SNP, indicating that SV could be imputed with similar accuracy to SNP, provided that genotyping of SV is highly accurate.

CONCLUSION

Results from this pilot study in two breeds indicate that a significant proportion of SV could be imputed into large target populations with SNP genotypes and phenotypic records for traits of interest. Further work is required to test the accuracy of imputation. However, imputation of SV will enable a more detailed understanding of the impact of SV on the cattle genome and phenotypes of interest to cattle industries.

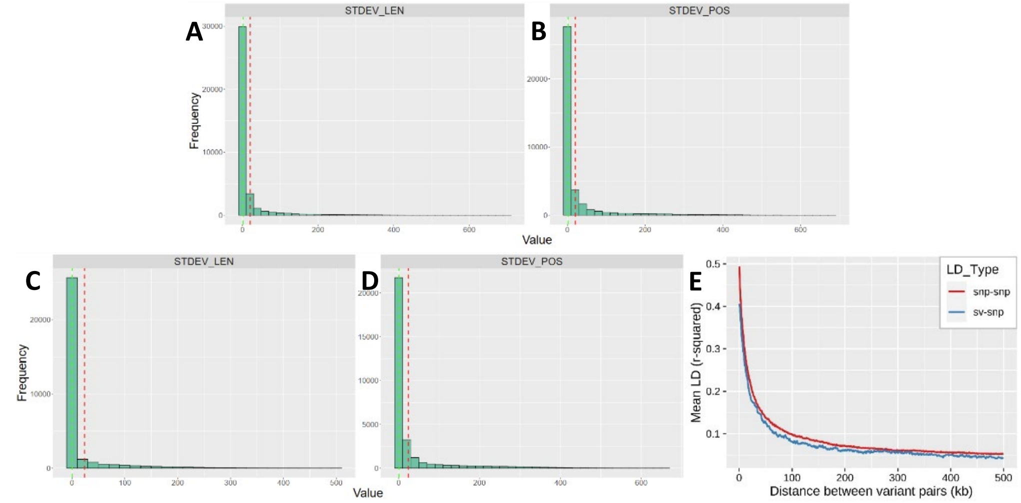

Figure 2. Frequency histograms of the standard deviation of length and position of insertions (A and B respectively) and deletions (C and D respectively) detected from long read sequence of 108 animals from 2 breeds. Red and green lines indicate the mean and median respectively. E) Mean linkage disequilibrium (LD as r^2) between variant pairs, either SV (insertions and deletions only) with SNP (blue) or SNP with other SNP (red) within 500kb on chromosome 1.

ACKNOWLEDGEMENTS

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

REFERENCES

Chaisson M.J.P., Sanders A.D., Zhao X., *et al.* (2019) *Nat. Commun.* **10**: 1784.

Chamberlain A.J., Nguyen T., Wang J., *et al.* (2023) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **25**: 274.

Chen S., Krusche P., Dolzhenko E., *et al.* (2019) *Genome Biol.* **20**: 291.

Chang C.C., Chow C.C., Tellier L.C.A.M., *et al.* (2015) *GigaScience* **4**: s13742–015–0047–8.

Duan X., Pan M. and Fan S. (2022) *BMC Genomics* **23**: 324.

Feuk L., Carson A.R. and Scherer S.W. (2006) *Nat. Rev. Genet.* **7**: 85.

Hayes B.J. and Daetwyler H.D. (2019) *Annu. Rev. Anim. Biosci.* **7**: 89.

Ho S.S., Urban A.E. and Mills, R.E. (2020) *Nat. Rev. Genet.* **21**:171.

Kadri N.K., Sahana G., Charlier C., *et al.* (2014) *PLoS Genet.* **10**: e1004049.

Kosugi S., Momozawa Y., Liu X., *et al.* (2019) *Genome Biol.* **20**: 117

Lee Y.-L., Takeda H., Costa Monteiro Moreira G., *et al.* (2021) *PLoS Genet.* **17**: e1009331.

Li H. (2018) *Bioinformatics* **34**: 3094.

Nguyen T.V., Vander Jagt C.J., Wang J., *et al.* (2023a) *Genet. Sel. Evol.* **55**: 9.

Nguyen T.V., Wang J., Chamberlain A.J., *et al.* (2023b) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **25**: 202.

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.