POPULATION SCALE LONG-READ SEQUENCE DATABASES: ARE THEY USEFUL FOR ACCURATE SNP AND INDEL DISCOVERY?

I.M. MacLeod[1,2], T.V. Nguyen[1], J. Wang[1], C.J. Vander Jagt[1] and A.J. Chamberlain[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

Several animal industries, including cattle, have built population scale whole-genome reference databases of genetic variants, SNPs and small INDELs, that have been discovered using short-read sequencing. These databases have proved invaluable: enabling development of genetic tools to breed healthier and more productive animals. However, while accurate and cost effective, short-read sequencing is not well suited to the discovery of larger genetic variants called structural variants (defined as > 50 base pairs in length). Thus, there is interest in creating population scale long-read databases for structural variant discovery and downstream applications. Ideally, for cost efficiencies, these would also contribute to the sequence database of SNPs and INDELs and enable imputation of all variants. Therefore, we explored the effect of long-read coverage on accuracy of SNP and INDEL discovery compared to a truth set from short-read sequence. The results show that at all read depths, recall and precision of SNP was considerably higher than for INDEL. At ≥ 10X read depth, SNP recall was 0.95 and reached 0.99 at 50X cover. The precision for SNPs and particularly INDELs suggested that the long-read variant calls included a relatively high, but likely overestimated proportion of false positives. We conclude that SNP and INDEL discovery in long-read data is useful, particularly if extensive 'truth' variant sets exist that could help remove false positives.

INTRODUCTION

Several animal industries, including cattle, have built population scale whole-genome reference databases of small genetic variants (SNPs, and INDELs < 50 base pairs) that have been discovered using short-read sequencing (Daetwyler *et al.* 2017). These databases have proved invaluable for the detection of recessive deleterious mutations, for sequence imputation and enabling the development of genetic tools to breed healthier and more productive animals. However, while short-read sequencing is highly cost effective and accurate for SNP and INDEL discovery, it is not well suited to the discovery of larger genetic variants (> 50bp in length) called structural variants (SVs). Instead, long-read sequencing is much better suited to genome-wide SV discovery. Limited research in livestock, and experience from human genetics research suggests that SVs may often have large effects on both mendelian and quantitative traits (reviewed by Nguyen *et al.* 2023a).

Until recently, two major deterrents for long-read sequencing have been the higher cost and lower per base accuracy, where the latter resulted in low quality SNP and INDEL calls compared to short-read sequencing. However, two key competitors in the field of long-read sequencing, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have made significant improvements in both per base accuracy and cost. Thus, there is now considerable interest in exploring the SV landscape at a population scale in cattle (Chamberlain *et al.* 2023) and potentially other livestock. For livestock studies, it is critical to consider how to reduce costs per individual without unduly compromising on the accuracy of variant discovery. The sequencing read depth is a key factor regulating cost, and Nguyen *et al*. (2023b) have used ONT long-read sequencing to explore the impact of read depth on the accuracy of SV discovery. Additionally, to maximise the cost effectiveness of long-read sequencing and to enable SV imputation, it is desirable to use these same sequences to develop new, or expand existing, whole-genome SNP and INDEL databases. Therefore, the aim of this paper was to explore the accuracy of SNP and INDEL discovery in long-

read sequencing at a range of read depths. Additionally, the paper considers the impact of incomplete discovery of these variants for population scale studies or smaller scale studies of recessive deleterious mutations.

MATERIALS AND METHODS

Three Holstein animals were each sequenced at approximately 50X coverage using an ONT PromethION sequencer, with flow cell 9.4.1 and ligation kit LSK110. To achieve maximum accuracy, the bases were re-called using Guppy v6.1.7 with the 'super high accuracy' setting (SUP). The output FASTQ files were trimmed using Filtlong (default settings: https://github.com/rrwick/Filtlong). Filtered reads were mapped to the ARS-UCD 1.2 reference genome (Rosen *et al.* 2020) using Minimap2 (Li 2018). Clair3 software was used to call SNPs and INDELs in individual sequences (default settings: Zheng *et al* 2022) and for comparison, Longshot software was also used to call SNPs (default settings: Edge and Bansal 2019).

Next, mapped reads at 50X coverage for each individual were subsampled using Sambamba (default settings: Tarasov *et al* 2015) to 3X, 5X, 10X, 15X and 20X coverage and the data at each read depth was processed as for the 50X coverage to re-call SNPs and INDELs. For each of the three animals, three chromosomes were chosen as technical replicates (chromosome 1, 19 and 25) to investigate the accuracy of SNP and INDEL discovery at each of these read depths. The same three animals had also been sequenced using short-read Illumina technology at approximately 12X, 15X & 18X read depth and were previously processed in Run8 of the 1000 Bull Genomes Project according to project guidelines (Daetwyler *et al.* 2017) with GATK joint variant calling according to GATK best practices (DePristo *et al.* 2011). The SNPs and INDELs discovered in the short-read data of the three animals were used as the gold standard 'truth set' of variants for comparison with the SNPs and INDELs discovered in the long-read sequencing for the same animals. To ensure a high quality truth set, we retained only biallelic variants with minor allele count of > 3, GATK Variant Quality Score Recalibration Tranche < 99.0, and indel < 50bp.

Hap.py software (https://github.com/Illumina/hap.py) was used to compare the variant truth set with the SNPs and INDELs discovered in the long-read sequencing that passed default software filters at each read depth ('query sets'). The following three sets of variants were identified from this comparison: 1) true-positive variants/genotypes (TP) that match in truth and query variant sets, 2) false-negative variants (FN) missed in the query set but present in the truth set, and 3) false-positive variants (FP) that have mismatching genotypes or alternate alleles in query versus truth set. The summary statistics calculated were; Recall = TP/(TP+FN) and Precision = TP/(TP+FP).

RESULTS AND DISCUSSION

The results were calculated for the combined truth variant sets across the three animals and three chromosomes, resulting in comparisons for a total of 1,894,775 SNPs and 158,338 INDELs at each read depth. As expected, accurate discovery of both SNPs and INDELs in long-read sequence was affected by read depth: declining more rapidly once read depth fell below 10X coverage, compared to higher read depths of 15X, 20X and 50X. The "recall" statistic (Figure 1A) indicates the proportion of variants that were discovered in the long-read data that were also in the truth set ("true positives": TP). There was excellent recall of SNPs from the long-read sequencing at 10X to 50X read coverage using Clair3 software, plateauing at between 0.95 to 0.99 (i.e. only 1 to 5% of SNPs in the truth set were not detected in the long-read sequence). Even at 5X coverage, Clair3 only missed 14% of SNPs. Longshot software showed much lower SNP recall, particularly at lower read coverage and even at 50X read depth 17% of SNPs were missed. This was expected because Longshot implements a less sophisticated variant calling approach (pileup only) compared to Clair3 which combines both pileup and full alignment in a deep learning-based variant calling algorithm (Zheng *et al.* 2022). Furthermore, Longshot is recommended for use with at least 30X read depth

and it cannot be used to call INDELs. The precision of SNP discovery was very similar for both Clair3 and Longshot (Figure 1B) and suggested that the proportion of false positives among all SNPs discovered in long-read sequences was between 12 to 28%. The precision was lower than for high quality human data reported to be 0.99 at 20X coverage (Zheng *et al.* 2022). However, there are several reasons why we would expect our precision to be lower: (1) our strict filtering of variants to create the 'truth set' from the short-read data would likely result in a proportion of real SNPs and INDELs being excluded so if found in the long-read data they appear to be false positives, (2) the human field has put tremendous effort into creating high quality truth sets through the "Genome in a Bottle Consortium" with higher short-read depth (35X) (e.g. Olson *et al.* 2022) while our lower coverage short-read data likely missed some real variants, and (3) Clair3 software algorithms were trained on human data with difficult to map regions excluded. Thus, our less accurate truth set compared to the human field will inflate the estimated false positive rate and this biases downwards our estimate of precision. There is clearly a need for high accuracy truth sets in cattle for improved benchmarking.

The recall and precision for INDELs using Clair3 was much lower than for SNPs, for example, recall ranged from 0.27 at 3X to 0.89 at 50X read depth (Figure 1a). Additionally, the recall rate kept improving with increased coverage compared to the plateau observed for SNP at around 15X coverage. As mentioned above, there is likely to be some downward bias in the estimate of precision. However, even in more accurate human data the precision for INDELs at 20X coverage was lower than for SNPs at around 0.87. INDEL calls in long-read data are known to be more error prone than for short-read sequence, particularly in homopolymer regions (consecutive repeat bases) where sequencing difficulty creates false positives (Amarasinghe *et al.* 2020; Delahaye and Nicolas 2021).

The high recall rates for SNPs suggests that long-read data of at least 10X coverage is likely to be of considerable value in augmenting or developing whole-genome SNP databases at population scale. This would be convenient because the study by Nguyen *et al*. (2023b) also suggested that read depth of ≥ 10X is preferable for population scale structural variant discovery. Furthermore, if the false negative rate for SNP in long-read data is around 10% or less and is largely sporadic (i.e., there is a different set of SNPs missing in each animal) this should enable highly accurate imputation of the missing SNPs where there are reasonable sized sequence databases. We examined the distribution of missing variants in our animals at 10X read depth (Chromosome 1) and found that only 4% of missing SNPs overlapped between each pair of animals on average. However, the overlap of the missing INDEL sets was much higher than for SNPs, averaging 16% between pairs of animals at 10X coverage. Therefore, if these INDELs are missed in most or all individuals and given the higher overall missing rate of INDELs compared to SNPs, then accurate imputation would require an existing reference population with accurately genotyped INDELs. If SNPs and SVs are accurately genotyped in long-read data then it will be possible to impute SVs into large populations of cattle with SNP panel genotypes using a reference population with long-read sequences.

Although the results suggest relatively high false positive rates, if there are existing short-read databases of variants (such as the 1000 Bull Genomes project: Hayes and Daetwyler 2019) then these could be used as a filter/training set to help remove false positive SNPs and INDELs from long-read data. In the case where research may be undertaken to discover a mendelian mutation of large effect in a small cohort of animals, Nguyen *et al*. (2023b) recommend long-read sequencing at ≥ 20X coverage for high accuracy discovery of a causal SVs in the data. Thus, if the mendelian mutation might equally be a SNP or INDEL, and no short-read sequence was available on the same animals, then sequencing (≥ 20X) of parent-offspring trios would be necessary to filter putative false positive variants (particularly INDELs) that do not show mendelian inheritance (although this would remove *de novo* mutations). Although INDELs constitute around 10% of all variants in Run8, they are important. For example, in Run8 of the 1000 Bull Genomes project Variant Effect Predictor software (VEP: McLaren *et al.* 2016) annotated 0.28% of INDEL, versus only 0.01% of SNP, to

have a high impact on a protein (i.e. loss of function, truncation and/or triggering nonsense mediated decay). A caveat of our study is that the ONT flow cell 9.4.1 used here for long-read sequencing is now superseded by a newer flow cell that should increase accuracy. Nonetheless, our results provide a useful benchmark, with the expectation that a range of advances will result in improved accuracy.

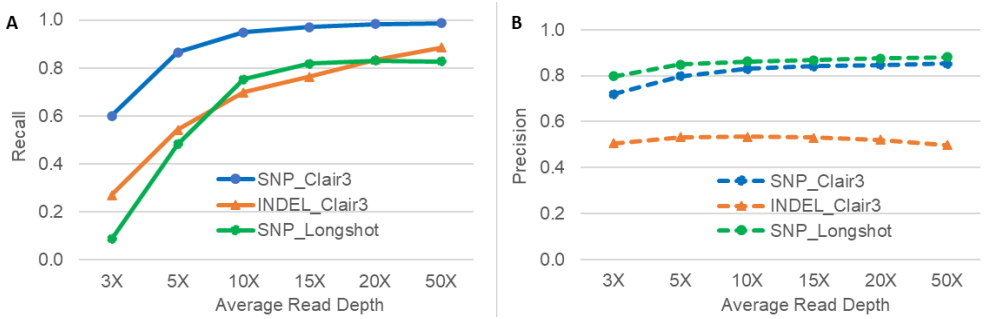

Figure 1. Recall (A) and precision (B) for SNP and INDEL discovery in long-read sequence of different read depths, using Clair3 (SNP and INDEL) and Longshot software (SNP only)

CONCLUSIONS

This study shows that with the use of existing truth sets of SNPs and INDELs, we can curate useful SNP and INDEL databases from long-read sequences. While there are some limitations particularly for small INDEL discovery in long-read sequence, it is likely that this will continue to improve with modifications in hardware, chemistry and variant calling algorithms. Also, there is a need to further develop truth sets in cattle of sequence variants for future benchmarking studies.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from DairyBio, a joint venture between Agriculture Victoria, Dairy Australia and the Gardiner Foundation. We acknowledge partners in the 1000 Bull Genomes Project for access to the data and thank Prof. Paul Stothard for providing VEP annotation.

REFERENCES

Amarasinghe S.L., Su S., Dong X. *et al.* (2020) *Genome Biol.* **21:** 30.
Chamberlain A.J., Nguyen N.V., Wang J. and MacLeod I.M. (2023) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **25:** *These proceedings*.
Daetwyler H.D., Brauning R., Chamberlain A.J., *et al*. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22:** 201.
Delahaye C. and Nicolas J. (2021) *PLOS ONE* **16**(10): e0257521.
DePristo M.A., Banks E., Poplin R., *et al*. (2011) *Nat. Genet.* **43:** 491.
Edge P. and Bansal V. (2019) *Nat. Comms.* **10:** 1.
Hayes B.J. and Daetwyler H.D. (2019) *Annual Review of Animal Biosciences* **7:** 89.
Li H. (2018). *Bioinformatics*. **34:** 3094.
McLaren W., Gil L., Hunt S.E., *et al*. (2016) *Genome Biology* **17:** 122.
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. and MacLeod I.M. (2023b) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **25:** *These proceedings*.
Olson N.D., Wagner J., McDaniel J. *et al.* (2022) *Cell Genom.* **2:** 100129.
Rosen, B.D., Bickhart, D.M., Schnabel, R.D., *et al*. (2020) GigaScience, **9:** 1.
Tarasov A, Vilella AJ, Cuppen E, Nijman IJ and Prins P. (2015). *Bioinformatics*. **31:** 2032.