COST EFFECTIVE DETECTION OF STRUCTURAL VARIANTS IN LONG-READ SEQUENCE – HOW DEEP IS ENOUGH?

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

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

SUMMARY

In recent years, the 1000 Bull Genomes Project has brought together short-read sequence for more than 6,000 cattle, playing a key role in detection of small variants and generating improved accuracy of genomic prediction for complex traits. While this continues to be an invaluable resource for SNP and small INDEL studies, it is not suited to detecting more complex structural variants (SV: variants with length > 50bp). However, SV often show large and sometimes deleterious effects on phenotypes and remain largely unexplored in livestock. Here, we use long-read sequences of two bovine parent-offspring trios to explore the optimal read depth to be cost effective whilst still maintaining a high chance of detecting SVs. This study shows that while sequencing from between 10X to 15X coverage resulted in some reduction in the SV discovery rate versus higher read depth, this may be an acceptable compromise for population scale studies to spread sequencing costs over a larger number of animals. However, if the purpose of using long-read sequencing is to discover a deleterious Mendelian mutation among a small group of known affected or carrier animals, the results here suggest that at least 20X cover would be preferable.

INTRODUCTION

Structural variants (SV) are genetic variations that involve the insertion, deletion, or rearrangement of large segments of DNA, typically affecting > 50 base pairs (Freeman *et al*. 2006). These types of variants can have significant impacts on gene function and expression, but their detection in livestock has been challenging due to limitations of short-read sequencing technology. To improve the accuracy and sensitivity of SV detection, several livestock genomics studies have deployed long-read sequencing technologies, mostly using PacBio and Oxford Nanopore Technologies (ONT) platforms. These studies have generally sequenced a relatively small sample of individuals at high read coverage, either to build reference pan-genomes or to pinpoint a deleterious SV (reviewed in Nguyen *et al*. 2023). Long-read sequencing is still relatively expensive for large population scale analyses, therefore it is critical to optimise read-depth for cost effective SV discovery. Therefore, we conduct a pilot experiment to study the effect of read-depth on discovery rate statistics of SV using two cattle parent-offspring trios.

MATERIALS AND METHODS

A flowchart of the methodology is illustrated in Figure 1. In brief, two Holstein trios (parents and offspring) were sequenced at ~60X coverage using ONT PromethION sequencer (flow cell 9.4.1 and ligation kit LSK110) following the manufacturer's recommendations. Post sequencing, the FAST5 files were re-basecalled using Guppy (v6.1.7) with the super high accuracy setting (SUP). The output FASTQ files were then trimmed using Filtlong (https://github.com/rrwick/Filtlong) with the default setting. Filtered reads were mapped to the ARS-UCD 1.2 + Btau5.0.1 Y reference genome ARS-UCD1.2 (Rosen *et al*. 2020) with additional Btau5.0.1 Y (Bellott et al. 2014) using Minimap2 aligner (Li 2018). Post sequencing and alignment, the recorded mapped coverage is estimated at 50X, so we considered this as the "baseline" read coverage. The aligned reads were used to detect SV with Sniffles2 (Sedlazeck *et al*. 2018) in individual samples and then merged using Sniffles2 joint genotyping function (default settings). Next, mapped reads at 50X coverage were scaled down using Sambamba (default settings: Tarasov *et al*. 2015) to an estimate of 3X, 5X, 10X, 15X, 20X coverage. These alignment files were re-exported to FASTQ format. The scaling down of read coverage was replicated three times for each animal at each coverage and replicate samples were then subjected to the same SV detection pipeline described above. For ease of analysis, we only considered SVs detected from autosomes (Chr 1 – 29). Finally, we deployed RTG Tools (https://github.com/RealTimeGenomics/rtg-tools) and its Mendelian plugin on merged SV calls to count the number of Mendelian consistent and inconsistent SV calls across each of the two trios (per replicate at each read cover). This plugin only counts SV that are genotyped in all individuals (i.e. excluding SV where one or more of the trio had a missing genotype).

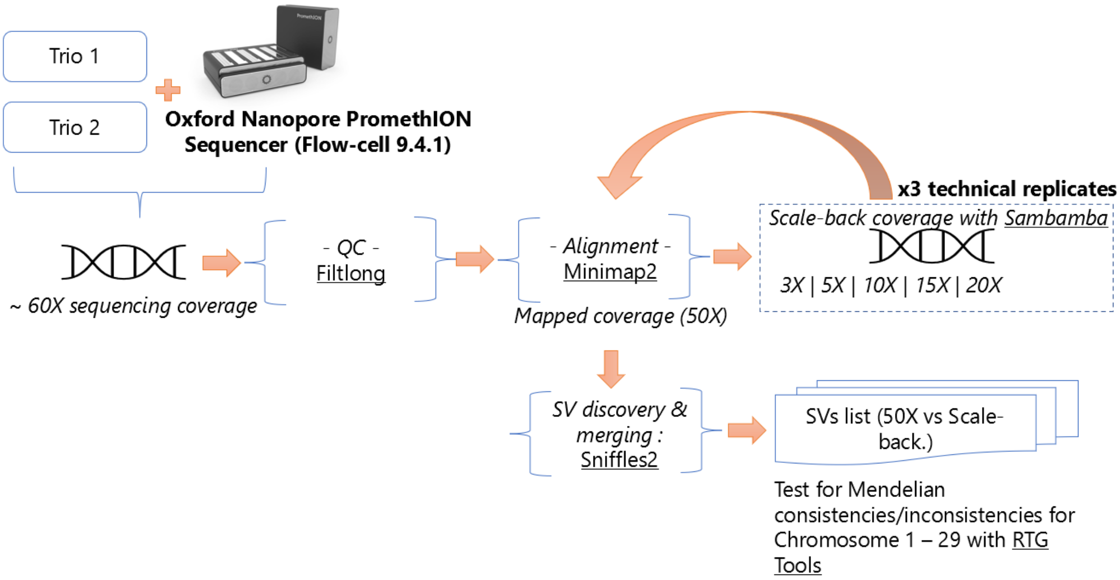
Trio 1
Trio 2
Oxford Nanopore PromethION Sequencer (Flow-cell 9.4.1)
x3 technical replicates
~ 60X sequencing coverage
- QC -
Filtlong
- Alignment -
Minimap2
Mapped coverage (50X)
Scale-back coverage with Sambamba
3X | 5X | 10X | 15X | 20X
SV discovery & merging :
Sniffles2
SVs list (50X vs Scale-back.)
Test for Mendelian consistencies/inconsistencies for Chromosome 1 – 29 with RTG Tools

Figure 1. Schematic workflow of the experiment to detect structural variants (SV) with different read coverage. Software used is shown with underlined text

RESULTS AND DISCUSSION

Table 1 summarises the SV discovery statistics at each scaled back read depth compared to the 50X cover. For example, on average at 3X we discover approximately 20% fewer SV, 93% fewer SVs with high quality genotypes (GQ > 10), and the sporadic missing genotype rate can be up to 21%. Missing rate is an important statistic to consider because the missing rate may impact the accuracy of imputing these sporadic missing for downstream use of SV genotypes. At 15X cover the summary statistics are much closer to the 50X cover compared to the 3X or 5X. In Table 2 we summarise the observed proportion of SV that violate Mendelian consistency at each scaled back read depth. Interestingly, the rate of Mendelian inconsistency only slightly increased with lower read cover: varying from 3-11% with larger variability between replicates at lower read depth. This estimate has some bias because we can only assess SV for which no individual in the trio has a missing genotype. Given that this number is relative to the total SV discovery at each coverage, this demonstrates that even at lower read-depth, if the SV are confirmed, the majority would be accurately genotyped in all animals. This might be expected because as coverage reduces it is likely that the merged set will be those SV that are relatively the easiest to detect.

It is important to note that a small number of Mendelian inconsistencies may arise from *de novo* mutation: for SNP this is in the order of ~30 given a bovine genome size of 3 Gb and a per base per generation mutation rate of 1x10^{-8}. It is likely that the number due only to SV would be similarly low. Therefore, here we made no attempt to differentitate *de novo* mutation from false positive SV as it would have no impact on our conclusions.

Table 1. Summary of Structural Variant (SV) discovery for different scaled back read depths, averaged across 2 trios of parent-offspring bovine (3 replicates) after join calling as well as in the original 50X read coverage. Average standard deviation between 2 trios is shown in brackets

	Number of SV called	**% non-missing genotypes among trio individuals**	**% genotypes where of all trio individual genotypes had GQ score[1] > 10**
3X	30,484.7 (40.4)	78.3 (0.1)	3.3 (0.1)
5X	37,015.5 (60.5)	93.9 (0.04)	29.4 (0.2)
10X	38,042.7 (57.8)	98.6 (0.04)	59.3 (0.1)
15X	38,292.7 (31.1)	99.2 (0.01)	73.9 (0.2)
20X	38,507.3 (8.5)	99.3 (0.02)	87.1 (0.1)
50X	38,513.5 (0)	99.3 (0)	96.1 (0)

[1] GQ is a composite mapping quality score that estimate the quality of the identified SV

Table 2. Summary of Structural Variants (SVs) observed in the offspring of two parent-offspring trios that show Mendelian consistency (Cons.) or inconsistency (Incons.) for a range of scaled back sequence coverage (average of 3 replicates) and in the original 50X coverage. Average standard deviation between the 2 trios is shown in brackets

	Number of Cons. SV	Number of Incons. SV	Rate of Incons.(%)
3X	21,585 (68.8)	2,189 (35.4)	9.2 (0.11)
5X	30,698 (29.7)	3,912 (58.3)	11.3 (0.13)
10X	34,470 (47.9)	3,001 (32.2)	8.0 (0.07)
15X	35,682 (43.4)	2,269 (26.3)	6.0 (0.06)
20X	36,616 (38.5)	1,598 (25.8)	4.2 (0.06)
50X	37,064 (0)	1,194 (0)	3.1 (0)

Our results demonstrate that the lower coverage of mapped reads increases the difficulty for the SV detection software to confidently call the genotype across multiple animals, particularly at 3X and 5X cover. This is likely partly due to the merging approach relying on there being at least one individual with good evidence of the SV and there being at least 5X cover of the SV region to call the genotype in each animal (as we are running with default settings), this perhaps explain the poor result of these two read depth coverages. Observing the missing rate, we can see that even at the highest read depth, in the merged calling of SV there are still around 1% of sporadically missing genotypes. We believe some of these missing genotypes are due to (i) complex SV that perhap require manual curation for accurate genotype calling, (ii) false positive SV that were either merged and/or joint called incorrectly. In addition, this study shows that while sequencing from between 10X to 15X coverage resulted in some reduction in the SV discovery rate versus higher read depth, this may be an acceptable compromise for population scale studies to spread sequencing costs over a larger number of animals. However if the purpose of using long-read sequencing is to discover a deleterious Mendelian mutation among a small group of known affected or carrier animals, the results here suggest that at least 20X cover would be preferable.

In this pilot study it is important to note that we deployed just one SV discovery program (Sniffles2), while there are several other programs currently available for this purpose. However, at the time of running this analysis, Sniffles2 was the software recommended for ONT long-read sequence with the best accuracy. In addition, Sniffles2 has an automated global/joint calling module that can automate calling of SV across population scale samples. Undoubtedly, results from this pilot study highlight the needs for further studies in resolving precise breakpoints and therefore, leading to more accurate genotyping of SV.

CONCLUSIONS

This study analysed the impact of sequencing read-depth on the detection of SV using two deeply sequenced bovine trios. Our results provide a means for future research to make decisions on optimising cost effective long-read sequencing cover for SV detection of either: (i) specific deleterious SV in a few individuals (iii) population scale genome-wide SV discovery or (iii) characterize an original set of SV such as for pan-genomes.

ACKNOWLEDGEMENTS

We are grateful for funding from DairyBio: a joint venture project between Agriculture Victoria (Melbourne, Australia), Dairy Australia (Melbourne, Australia) and the Gardiner Foundation (Melbourne, Australia).

REFERENCES

Bellott D.W., Hughes J.F., Skaletsky H., *et al.* (2014) *Nature* **508:** 494.0.

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

Freeman J.L., Perry G.H., Feuk L., *et al.* (2006) *Genome Res.* **16:** 949,

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

Rosen B.D., Bickhart D.M., Schnabel R.D., *et al.* (2020) *GigaScience* **9:** 1.

Sedlazeck F.J., Rescheneder P., Smolka M., *et al.* (2018) *Nat Methods*. **15:** 461.

Tarasov A, Vilella A.J., Cuppen E., Nijman I.J., Prins P. (2015) *Bioinformatics*..**31:** 2032.