Proc. Assoc. Advmt. Anim. Breed. Genet. 25: 75 - 78

SIMULTANEOUS INVESTIGATION OF GENOMIC REGIONS OF INTEREST – THE USE OF ADAPTIVE SAMPLING

R.M. Clarke[1], A. Hess[2], A, Caulton[1], R. Brauning[1], K.M. McRae[1], A. Chen[3] and S.M. Clarke[1]

[1]AgResearch, Invermay Agricultural Centre, Mosgiel, Otago, NZ
[2]University of Nevada, Reno, Nevada, USA
[3]University of Otago, Dunedin, Otago, NZ

SUMMARY

Utilization of more complex genetic variation present within a population can help to address the challenge of identifying animals that perform optimally in their environment while reducing their environmental impact. The objective of this study was to determine if Oxford Nanopore sequencing technology provides a potential solution to capture these data types in a cost-effective and high-throughput method. Adaptive sampling was used to investigate regions of interest surrounding known genome-wide association studies peaks and copy number variation regions in sheep with higher enrichment achieved in targeted areas. Multiplexing of three animals was achieved, but further work is needed to determine cost-effectiveness of this tool for the animal industry.

INTRODUCTION

Providing energy-rich protein to the world while reducing the environmental impact is one of the largest challenges facing the animal industry today. To face this challenge, novel tools need to be adopted for methods to identify animals that perform optimally in their environment. This includes, but is not limited to, utilizing more complex variation, epigenetics, and microbial communities present within the host. A major hurdle in utilizing these data lies in the development of cost-effective and high-throughput methods for data capture. We propose the sequencing platform developed by Oxford Nanopore Technologies (ONT) as a potential solution.

Adaptive sampling, a software-controlled enrichment unique to the nanopore sequencing platform, enables targeted sequencing of specific regions of a genome or species of interest at higher coverage than non-selected regions of the genome. Adaptive sampling allows the sequencing of particular regions of DNA to be enriched through the comparison of the first 400bp of a strand of DNA to a provided sequence list. If the 400bp match the sequence list then the strand continues to be sequenced, if there is no match then the stand is ejected, and the pore is available for the next strand (Payne *et al* 2021). This enrichment approach not only circumvents the need for upfront sample manipulation but also enables simultaneous capture of multiple sources of information such as methylation, mutations, and structural variances in a single run, with the aim of reducing costs (Payne *et al*. 2021).

MATERIALS AND METHODS

DNA was extracted from 19 sheep (Montgomery and Sise 1990) and libraries were prepared using SQK-LSK109 with the native barcode expansion pack EXP-NBD104 (ONT) as per ONT protocols. Adaptive sampling was done on 52 regions of interest (ROI) surrounding previously identified GWAS peaks (unpublished data) and known structural variations such as the Haemoglobin region. High molecular weight (>60kb) DNA and fragmented DNA samples (10-20kb) were compared, as well single verseuse multiplexed samples to determine the optimal output for adaptive sampling in sheep.

Retained reads were analysed using Nanoplot (De Coster *et al*. 2018) to check quality and mapped to the Oarv3.1 (Jiang *et al*. 2014) and ARS-UI_Ramb_v2.0 (Rambv2.0; Davenport *et al*. 2022) sheep genomes using minimap2 (Li, 2018). Mosdepth (Pedersen and Quinlan 2018) was used

to determine the mean and median read depth across the ROI and the whole genome. Coverage results were displayed using R studio and Samplot (Belyeu *et al*. 2021).

Faecal samples were taken from sheep challenged with a single isolate of the gastrointestinal nematode *Haemonchus contortus*. DNA was extracted using the QIAamp PowerFecal Pro DNA kit (QIAGEN), and sequenced using random Genotyping by Sequencing (GBS; Dodds *et al*. 2015) to determine the percentage of DNA mapping to the *Haemonchus* (Doyle *et al.* 2020) and sheep (Oarv3.1) genomes. To determine if parasite and host DNA can be detected from faecal samples using ONT, adaptive sequencing was completed with enrichment for *Haemonchus* genome sequence. The passed reads were mapped to the *Haemonchus* genome using minimap2 and the percent of accepted reads and reads mapped to the genome were calculated. The passed reads and the failed reads, which are the first 400bp that are sequenced then rejected as not matching, were mapped against the Oarv3.1 genome using minimap2. The reads that passed were also BLAST searched using BLASTn to determine the possible source of the DNA. The percent of reads mapped to *Haemonchus* and the sheep genome were compared between techniques.

RESULTS AND DISCUSSION

Multiplexing of the 52 ROI showed that higher enrichment is seen in the selected regions versus non-selected regions and that the median coverage of these regions ranges from 1-3x coverage when three samples are multiplexed together in the same run (Figure 1).

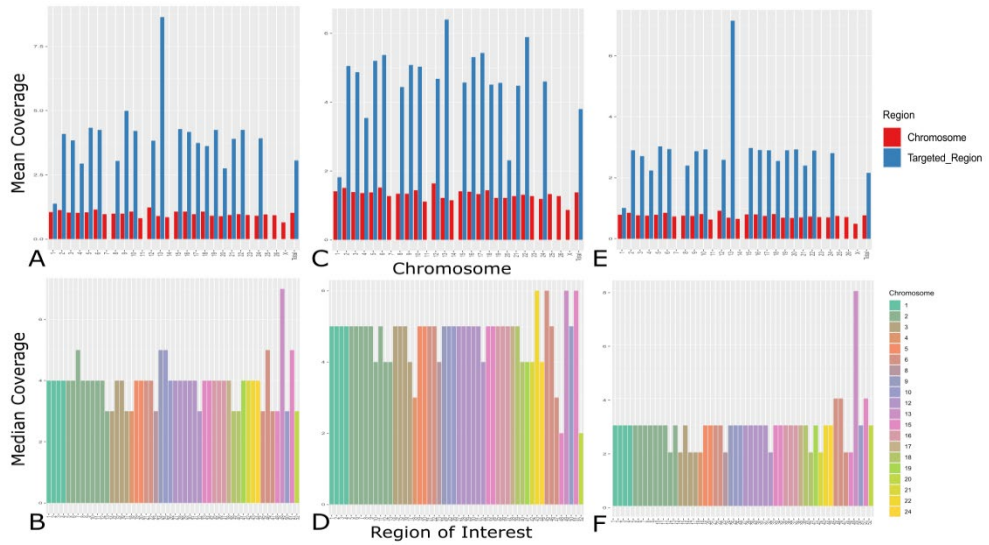
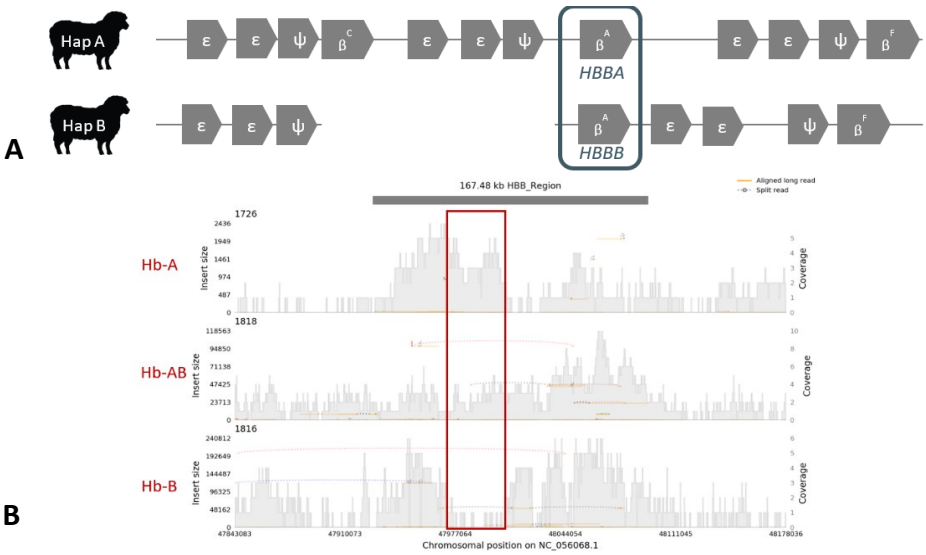


Figure 1. Mean coverage of the target regions compared to the chromosomes and the median coverage per target region for three multiplexed samples. The top panel (A, C, E) shows the mean coverage of each chromosome (red) and the mean coverage of the target regions per chromosome (blue). The x-axis labels the chromosomes. The bottom panel (B, D, F) shows the mean coverage of each of the 52 ROI, x-axis labels the regions of interest (1-52) with the colours indicating the chromosome on which the region is located. Barcode 1 (A-B), 2 (C-D) and 3 (E-F) are three individual animals that were multiplexed.

The β-globin locus of the ovine genome was chosen as an exemplar in this study, which is a region on chromosome 15 formed by the duplication of an ancestral four-gene set consisting of two embryonic-like genes, a pseudogene, and a β globin gene. Each set contains a different form of the β-globin gene, which is synthesised to make different forms of haemoglobin during development; in the foetus (F), pre-adult (C) and adult (A). There are two haplotypes, a long one comprising of all

three gene sets (haplotype A) and a short one (haplotype B), where the juvenile set is missing (Figure 2) (McRae *et al*. 2022). These two haplotypes are tagged by two SNPs and the individuals are genotyped allowing a comparison between the SNP genotype and the adaptive sampling targeting the entire haemoglobin locus. The adaptive sampling shows a corresponding genotype as the SNPs with the long version of the locus Hb-A having reads across the area and the short Hb-B haplotype not having reads across the juvenile gene set (Figure 2).

Figure 2. A) Schmetic of the Haemoglobin locus with both the short (HapB), the long (HapA) haplotype. B) Mapped examples of the long (Hb-A), short (Hb-B) and the heterozygous (Hb-AB) aligned to the Rambv2.0 (long haplotype) genome. Left axis shows the aligments scaled by insert size (distance between pair ends) and the right axis shows the per base coverage



The comparison between the GBS and adaptive sampling shows that when targeting the entire *Haemonchus* genome, a similar percentage of reads mapped to the genome using either GBS or adaptive sequencing (Table 1). When the adaptive sequencing reads were mapped against the *Haemonchus* genome only 0.38% of the reads were mapped. These reads were BLASTn searched to determine the source and the top 5 hits are shown in table 2. This suggests that the first 400bp of the read which adaptive sampling uses to make its decision to accept, or reject is matching a common sequence in the genome that is present in other species. To make this more specific to parasite DNA in faecal samples, the ITS2 region could be provided as a target region. The percentage of host DNA that is detected in adaptive sampling from both the accepted and failed reads combined shows a higher percentage of host DNA than is detected in the GBS (Table 1).

CONCLUSIONS

We have utilized adaptive sampling to investigate ROI surrounding known GWAS peaks and CNV regions with ~2-15x higher enrichment in selected areas versus non-selected areas. Enrichment of both host and parasite DNA from faecal samples shows that this technique can be utilized for different sample types and has flexibility in the information acquired. The results show that the use of the whole genome of a single parasite as the target sequence resulted in reads being accepted from a range of sources and not only the intended targets. To overcome this, adaptive sampling targeting the ITS2 region may provide a better sequencing enrichment of parasite DNA from faecal samples.

Table 1. Reads mapped to the *Haemonchus* and sheep genomes from both GBS and adaptive sampling on the same DNA faecal samples

	GBS (% mapped)	**Adaptive sequencing** (% reads accepted)	**Accepted reads mapped to corresponding genome** (%)	**Accepted and failed reads mapped to corresponding genome** (%)
Haemonchus	0.14	0.15	0.38	
Sheep	0.01			0.14

Table 2. Top 5 BLASTn matches for reads that were accepted as matching to the *Haemonchus* genome

BLASTn match	Accepted reads matched (%)
Haemonchus contortus	22
Plasmodium berghei ANKA	18
Chrysodeixis includens	13
Heterocephalus glaber	11
Bos taurus	3

We have also shown multiplexing can be achieved in conjunction with adaptive sequencing, but the current level of multiplexing that can be achieved to still provide the required coverage suggests that while this tool is useful for discovery and validation, it is not at the point of moving through to industry uptake. If the number of samples that can be run in one multiplexing run can be increased by, for example, using the R.10.4 flow cells on the PromethION, where current predictions are 15 samples for multiplexing, this would provide a more cost-effective option for the industry.

ACKNOWLEDGEMENTS

Thanks must go to Tania Waghorn for providing the ovine faecal samples. Funding provided by Genomics Aotearoa and AgResearch's Ministry of Business, Innovation and Employment (MBIE) Strategic Science Investment Funding (SSIF).

REFERENCES

Belyeu J.R., Chowdhury M. ,Brown J., … and Layer R.M. (2021) *Genome Biol.* **22:**1.
Davenport K.M., Bickhart D.M., Worley K., … & Rosen B.D. (2022) *Gigascience*, **11**.
De Coster W., D'hert S., Schultz D.T., Cruts M., and Van Broeckhoven C. (2018) *Bioinformatics*, **34:** 2666.
Dodds K.G., McEwan J.C., Brauning R., … & Clarke S.M. (2015) *BMC Genomics* **16:** 1.
Doyle S.R., Tracey A., Laing R., … & Cotton J.A. (2020) *Commun.Biol.* **3**: 656.
Jiang Y., Xie M., Chen,W., Talbot R., … and Dalrymple B. (2014) *Science*, **344:** 1168.
Li H. (2018) *Bioinformatics* **34:** 3094.
McRae K.M., Rowe S.J., Johnson P.L., … and Clarke S.M. (2021) *Genes* **12:** 1560.
Montgomery G.W., & Sise J. (1990) *New Zeal. J.Agr. Res.* **33:** 437.
Payne A., Holmes N., Clarke T., … and Loose M. (2021). *Nat. Biotechnol.* **39:** 442.
Pedersen B.S., and Quinlan A.R. (2018). *Bioinformatics*, **34:**867-868.