LIBRARY PREPARATION METHOD AFFECTS OBSERVED MICROBIOME VARIATION WHEN USING OXFORD NANOPORE SEQUENCING

E.M. Ross[1], Z. Chen[1], L.T. Nguyen[1], S. Meale[2], C. T. Ong[1]

[1]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD, 4072 Australia
[2]School of Agriculture and Food Sciences University of Queensland, Gatton, QLD, 4343 Australia

SUMMARY

New sequencing technologies are opening up new opportunities to explore microbiome variation; however, the technical effects of the molecular methods used have not been characterized. In this study, we aimed to investigate the potential impact of different library preparation methods and base calling algorithms on the observed microbiome variation when using Oxford Nanopore Technologies sequencing. To achieve this, we sequenced technical replicates of a single rumen fluid sample from a cannulated Bos taurus. Our results showed that the use of higher accuracy base calling methods led to a significant increase in the number of classified reads, resulting in more usable data. We did not observe any alteration in the microbial profile due to the use of different base calling algorithms. We also found that the rapid library preparation sequencing kit, which uses an enzymatic method to cut the DNA and ligate the adapter, resulted in shorter sequence lengths and lower numbers of classified reads compared to the Ligation library preparation kit, which does not cut the DNA during library preparation. Importantly, we observed significant differences in the proportion of microbial species within the data generated using the Ligation versus the rapid library preparation kit. Our study suggests that the library preparation method used can impact the observed microbiome and is therefore important to consider in any downstream analysis.

INTRODUCTION

Metagenomics is a popular method to describe microbiome variation, with one important application being the investigation of the relationships between microbiome variation and host phenotype (e.g. Ross *et al*. 2013). Accurate representation of microbiome variation is essential to detect these associations. While short-read sequencing has been the primary method for microbiome analysis to date, the declining cost of long-read sequencing has made it a potential alternative (e.g. Ong *et al*. 2023). To confidently adopt long-read sequencing, specifically using Oxford Nanopore Technologies (ONT), it is crucial to investigate the technical effects of the molecular methods used, as well as the algorithms used to analyse the raw output signal. In this study, we aimed to test the hypothesis that the library preparation method used for generating the ONT sequencing library significantly affects the observed microbiome. Additionally, we tested the hypothesis that the base calling algorithm significantly affected the observed microbiome.

MATERIALS AND METHODS

Sample. This study used technical replicates from a single rumen fluid sample taken from a single 3-year-old cannulated cow (*Bos taurus*) under animal ethics number 2021/AE000991. The animal was fed with hay as a regular diet. Rumen fluid collection was performed by restraining the animal in a crush, removing the cannula, and collecting rumen contents. The rumen fluid was squeezed from the rumen contents and then sieved to remove large particulate matter. The rumen fluid was distributed into 1.5 mL tubes after homogenization and stored at -20°C until samples were processed.

DNA extraction. Thawed 1.5 mL rumen fluid samples were centrifuged at 14,000 rpm for 5 min at 4°C, followed by the removal of the supernatants. Multiple DNA extraction methods were performed to characterise microbiome differences between extraction kits compared to sequencing methods. DNA extraction was performed on the cellular pellet in triplicate for each method. The DNeasy Plant Mini Kit (QIAGEN, Germany) was performed following the manufacturer's protocol. The PowerFecal Pro DNA Kit (QIAGEN, Germany) was used according to the instruction from the manufacturer. The Puregene Blood Core Kit (QIAGEN, Germany) extraction was performed by following the Gram-positive bacteria protocol provided by the manufacturer. Chemical cell lysis was performed in the DNeasy Plant Mini Kit and Puregene Blood Core Kit, while the PowerFecal Pro DNA Kit was combining chemical and mechanical processes. The extracted DNA was stored at -20°C for subsequent use.

Sequencing. Two sequencing kits, the ligation kit (SQK-LSK109) and the rapid kit (SQK-RBK110.96), were used in this study. The Ligation Kit was used for the library preparation for all extraction methods. Exclusively, the rapid kit was used with the PowerFecal Pro DNA kit (Table 1). Barcoding during the library preparation of DNA samples from the Puregene Blood Core Kit was performed using EXP-NBD104. Library preparations were conducted according to the manufacturer's instructions with some modifications as previously described (Hayes et al., 2021). Sequencing was performed on the PromethION P24 (ONT, UK) with the MinKNOW v.22.03.4 software using FLO-PRO002 (R9.4.1) flow cells. Samples were sequenced for 24 hours. Three basecall models, named Fast basecalling (FA), High accuracy basecalling (HAC), and Super Accurate basecalling (SUP), as well as the barcode demultiplexing, were operated by Guppy v.6.0.7. The adapter and barcode trimming functions were not selected during the sequencing.

Bioinformatics. Porechop v.0.2.4 (Wick *et al*. 2017) was performed for the trimming of adapters and barcodes. Minimum Q scores for reads generated from FA, HAC, and SUP basecall models were 8, 9, and 10, respectively. Reads under the minimum Q scores of corresponding basecall methods and less than 100 bp were filtered by Nanofilt v.2.8.0 (De Coster *et al*. 2018). Read-based taxonomic classification was performed by Kraken2 v.2.1.2 (Wood *et al*. 2019) with a customized Kraken2 database. A customized Kraken2 database was used in this study to increase the taxonomic classification efficiency. The complete genomes of bacteria, fungi, archaea, and protozoa from the NCBI RefSeq were downloaded to construct the customized database, with the low-complexity sequences masked. The Vegan v.2.6-2 (Dixon 2003) and phyloseq v.1.40.0 (McMurdie and Holmes 2013) package implemented in R, were used for the calculation of alpha diversity (Shannon index). A linear model with the DNA preparation method and/or sequencing kit as covariates was employed to assess significance.

RESULTS AND DISCUSSION

The sequencing process generated a total of 49,917,517 raw reads. Following trimming and filtering, 2,096,033 reads (4.2%) were excluded, leaving 47,821,484 reads that passed quality control. These reads were subsequently classified using the Kraken2 tool (Figure 1). The N50 values for sequence data generated from the Ligation Kit were higher (6,558 to 7,941) than for the Rapid Kit (4,662 to 4,952) with Powerfecal kit extraction (Table 1). The N50 value was positively correlated with the proportion of classified reads ($r = 0.88$, $P < 0.001$). Increasing the basecalling accuracy led to an increase in the proportion of classified reads (Figure 1A), rising from a mean of 29.71 (FA) to 38.40 (SUP). Notably, within the Powerfecal excitation kit data the ligation library preparation kit resulted in a greater proportion of reads assigned to a taxon than the rapid library preparation kit (Figure 1B).

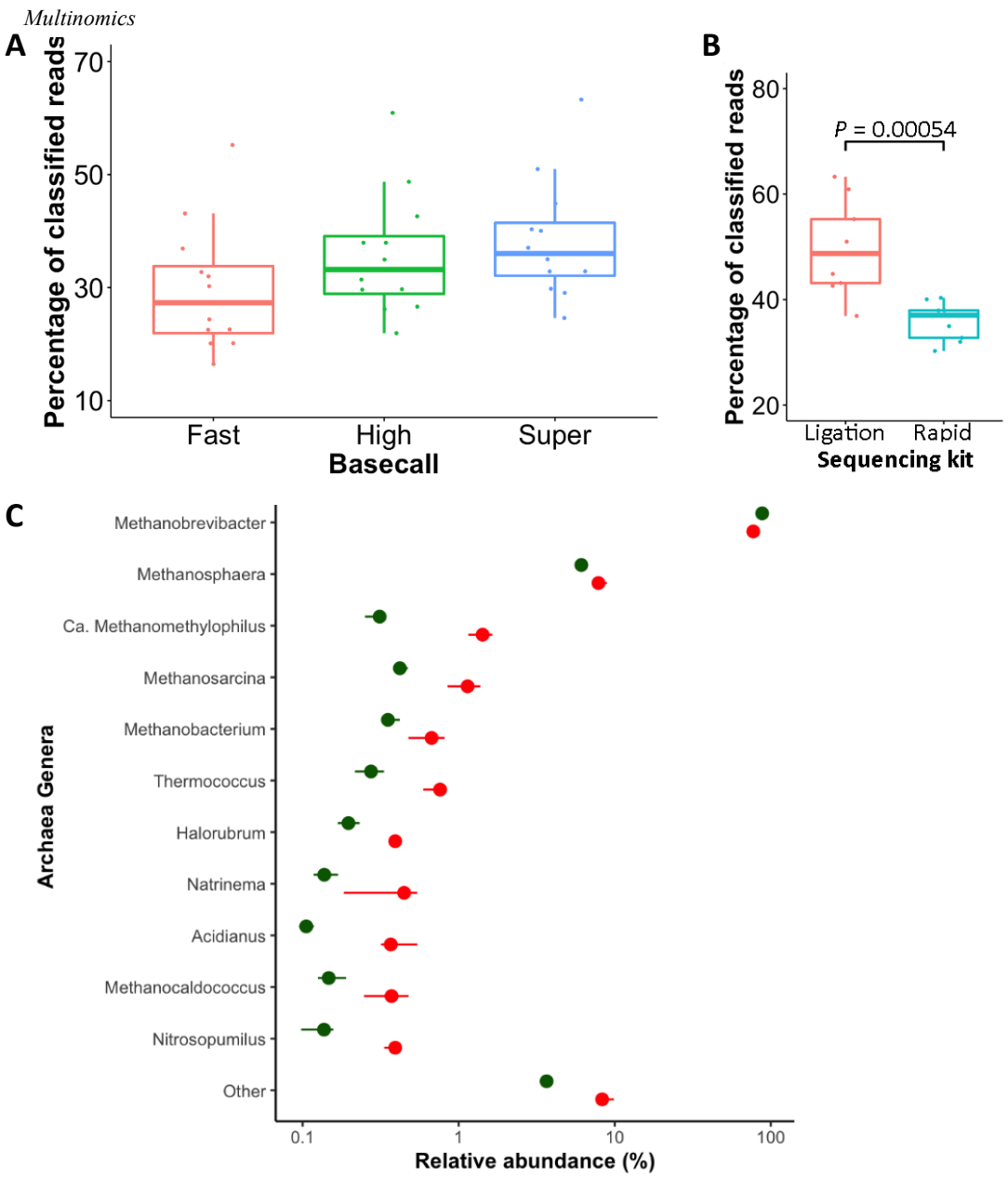

Figure 1. A) Percentage of reads assigned to a taxon for each base calling accuracy level. B) Percentage of reads assigned to a taxon for each library preparation kit (Powerfecal DNA extraction method only). A linear model was used to assess statistical significance. C) Within Archaea, the proportion of reads assigned to each genera from the ligation (Red) and rapid (green) sequencing kits

Table 1. Average lengths (N50) of the sequencing reads from each of the molecular methods

Extraction_Method	Sequencing_Kit	Mean	Sd	Median
DNeasy	Ligation_Kit	1532.22	587.87	1382.00
PowerFecal	Ligation_Kit	7460.89	659.32	7875.00
Puregene	Ligation_Kit	1431.44	149.80	1418.00
PowerFecal	Rapid_Kit	4792.89	115.46	4731.00

Microbial abundances at the Kingdom level were affected by DNA extraction ($P < 0.01$) and library methods ($P < 0.05$), but not basecall models ($P > 0.05$). Bacteria dominated the rumen microbial community (> 90.90%) for both extraction and sequencing kits. Significant effects were observed for the abundance of archaea genera (Figure 1C) based on both extraction and sequencing kits ($P < 0.05$), but not basecall models ($P > 0.05$). The Ligation Kit had a higher Shannon index ($H=2.56$) than the Rapid Kit ($H=2.08$). Conversely, the Rapid Kit had greater bacterial species diversity ($H=6.13$) than the Ligation Kit ($H=6.11$). The fungal diversity was slightly higher in the Rapid Kit than in the Ligation Kit ($H=4.40$ versus $H=4.46$, $P < 0.05$). Notably, DNeasy and Puregene extracted samples had less archaea abundance, but higher archaeal Shannon index compared to PowerFecal extracted samples. Basecall models did not affect the archaeal richness and evenness ($P > 0.05$).

CONCLUSION

Base calling accuracy in ONT sequencing of microbiome samples affects the proportion of reads that can be classified, but not species ratios, thereby impacting data acquisition costs. The choice of library preparation kit has a significant influence on the observed distribution of microbial species. Therefore, it is crucial to record the library preparation kit information in the metadata of public sequence repositories and account for it in statistical models.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the invaluable support of their colleagues at the Centre for Animal Science, within the Queensland Alliance for Agriculture and Food Innovation, and the staff at the Research Computing Centre of The University of Queensland. This research was funded by Meat and Livestock Australia (project P.PSH.2010) and Research Strategic Package 3 funding from the University of Queensland. The authors also acknowledge the use of ChatGTP, an AI algorithm from OpenAI, to assist with editing the final manuscript on February 16, 2023. A small number of corrections were made to the suggested edits to preserve the accuracy and nuance.

REFERENCES

De Coster W., D'Hert S., Schultz D.T., Cruts M. and Van Broeckhoven C. (2018) *Bioinformatics* **34:** 2666.

Dixon P. (2003) *J Veg Sci*. **14:** 927.

Hayes et al. (2021). *Frontiers in Genetics* **12:** 760450

McMurdie P.J. and Holmes S. (2013) *PLOS One* **8**: e61217.

Ong C.T., Boe-Hansen G., Ross E.M., Blackall P.J., Turni C., Hayes B.J. and Tabor A.E. (2022) *Microbiology Spectrum*. **10:** e00412.

Ross E.M., Moate P.J., Marett L.C., Cocks B.G. and Hayes B.J. (2013) *PLOS One* **8:** e73056.

Wick R.R., Judd L.M., Gorrie C.L. and Holt K.E. (2017) *Microb Genom.* **3:** e000132.

Wood D.E., Lu J. and Langmead B. (2019) *Genome Biol*. **20:** 257.