BMC Genomics

RESEARCH Open Access

Biases from Oxford Nanopore library preparation kits and their effects on microbiome and genome analysis

Ziming Chen[1], Chian Teng Ong[1], Loan To Nguyen[1], Harrison J. Lamb[1], O. González-Recio[2], M. Gutiérrez-Rivas[2], Sarah J. Meale[3] and Elizabeth M. Ross[1*]

Abstract

Background Oxford Nanopore sequencing is a long-read sequencing technology that does not rely on a polymerase to generate sequence data. Sequencing library preparation methods used in Oxford Nanopore sequencing rely on the addition of a motor protein bound to an adapter sequence, which is added either using ligation-based methods (ligation sequencing kit), or transposase-based methods (rapid sequencing kit). However, these methods have enzymatic steps that may be susceptible to motif bias, including the underrepresentation of adenine-thymine (AT) sequences due to ligation and biases from transposases. This study aimed to compare the recognition motif and relative interaction frequencies of these library preparation methods and assess their effects on relative sequencing coverage, microbiome, and methylation profiles. The impacts of DNA extraction kits and basecalling models on microbiome analysis were also investigated.

Results By using sequencing data generated by the ligation and rapid library kits, we identified the recognition motif (5′-TATGA-3′) consistent with MuA transposase in the rapid kit and low frequencies of AT in the sequence terminus of the ligation kit. The rapid kit showed reduced yield in regions with 40–70% guanine-cytosine (GC) contents, while the ligation kit showed relatively even coverage distribution in areas with various GC contents. Due to longer reads, ligation kits showed increased taxonomic classification efficiency compared to the rapid protocols. Rumen microbial profile at different taxonomic levels and mock community profile showed significant variation due to the library preparation method used. The ligation kit outperformed the rapid kit in subsequent bacterial DNA methylation statistics, although there were no significant differences.

Conclusions Our findings indicated that careful and consistent library preparation method selection is essential for quantitative methods such as bovine-related microbiome analysis due to the systematic bias induced by the enzymatic reactions in Oxford Nanopore library preparation.

Keywords Oxford Nanopore sequencing, DNA extraction, Library preparation, GC bias, Microbiome

*Correspondence:
Elizabeth M. Ross
e.ross@uq.edu.au
[1] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD 4072, Australia
[2] Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, INIA-CSIC, Madrid 28040, Spain
[3] School of Agriculture and Food Sustainability, University of Queensland, Gatton, QLD 4343, Australia

Background

Since the introduction of next-generation sequencing (NGS), sequencing technologies have been applied to conduct a wide range of biological studies, such as genome assembly and medical diagnosis, due to their increased efficiency compared to Sanger sequencing. The parallel sequencing of billions of DNA molecules in NGS can generate large amounts of output data. Currently, short-read and long-read sequencing are two primary approaches to obtaining DNA sequences. Present widely used short-read sequencing technologies include Illumina and DNBSEQ, which allow the yield of data with higher coverage and accuracy (over 99.9%), while with the DNA at a shorter read length (250–800 bp) [1]. In comparison, long-read sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) allow longer reads (over 10 kb) [1]. Although relatively lower per-base accuracy was seen previously in long-read sequencing platforms, the current improved chemistries in PacBio and ONT allow their increasing per-base accuracy to reach up to 99.9% [2, 3].

Library preparation steps are required for NGS, which aims to attach DNA strands with adapters compatible with sequencers. Different sequencing platforms have designed a variety of commercial library preparation kits to achieve this. On the ONT long-read sequencing platform, two major types of library preparation techniques are currently provided for DNA samples, which are based on transposases (rapid kits) and ligases (ligation kits). Enzymes are used in library preparation for DNA fragmentation or adaptor ligation. Transposase-based kits offer simplified protocols that notably reduce the time for library preparation, compared to those requiring end-repair and ligases. Generally, in ONT transposase-based protocols, the transposome complex involved in the library kits hydrolyzes DNA strands into shorter fragments and simultaneously attaches barcodes or adapters to cleaved ends. In a patent [4], ONT indicates the involvement of MuA transposase in their transposase-based protocols. In comparison, ligation protocols for ONT utilize T4-relevant enzymes and Taq DNA polymerase or DNA Polymerase I, Large (Klenow) Fragment for library construction. In brief, after the generation of blunt-end DNA and phosphorylated 5' end with enzyme mixtures, such as T4 polymerases and T4 polynucleotide kinase, dA tails are created in DNA strands by Taq DNA polymerase at elevated temperatures or the mesophilic DNA Polymerase I, Large (Klenow) Fragment, which facilitates subsequent adapter ligation by T4 ligase [5–7].

Table 1 Bovine and microbiome data used in this study

Location	Sample type	Extraction method	Library protocol	Biological replicates	Technical replicates
Australia	Rumen fluid	PowerFecal Pro	SQK-LSK109 (ligase-based)	1	3
	Rumen fluid	PowerFecal Pro	SQK-RBK110.96 (transposase-based)	1	3
	Rumen fluid	DNeasy Plant Mini	SQK-LSK109 (ligase-based)	1	3
	Rumen fluid	Puregene	SQK-LSK109 with EXP-NBD104 (ligase-based)	1	3
Spain	Rumen fluid	PowerSoil	SQK-LSK109 (ligase-based)	4	1
	Rumen fluid	PowerSoil	SQK-RBK004 (transposase-based)	4	1
Australia	Cattle ear punch	Puregene	SQK-NBD114.24 (ligase-based)	15	1
	Cattle ear punch	Puregene	SQK-RBK110.96 (transposase-based)	15	1

Sequencing bias can be introduced during the library preparation from different sources. The polymerase chain reaction (PCR) amplification is regarded as a major cause of uneven sequencing coverage in extreme guanine-cytosine (GC) content regions [8, 9]. In addition, it was reported that the insertion preference of Tn5 transposase could introduce lower sequencing depth in the GC-poor spectra [10]. Likewise, a previous study found that MuA transposase showed insertion bias, contributing to imbalanced sequencing data coverage [11]. Furthermore, the increased incubation temperature during sample and library preparation can also contribute to low sequencing coverage of high adenine-thymine (AT) regions [12, 13]. This biased coverage resulting from library preparation can lead to the loss of single nucleotide variants, the misrepresentation of the microbial composition in microbiome studies, and the misinterpretation of chromatin accessibility data [9, 14, 15].

This study aimed to characterize different forms of bias in Oxford Nanopore library preparation by comparing the recognition motif, interaction frequencies, sequencing coverage, and microbiome profiles of two of the most commonly used Oxford Nanopore library preparation kits (rapid kits and ligation kits). We hypothesized: (1) the MuA transposase has a bias in insertion sites and regions of coverage; and (2) the observed microbiome profiles vary based on the sequencing kits used.

Results

Two types of data were used in this study: cattle ear tissue DNA and the rumen microbiome DNA (Table 1; Supplemental Table 1; Supplemental Table 2). The ear tissue data was used to identify the sequencing bias from two

ONT library protocols (ligase-based and transposase-based). The rumen fluid data were used to characterize the microbiome profile variation between two ONT library protocols (ligase-based and transposase-based).

Enzymatic interaction motifs

The MuA enzyme is used in the ONT rapid kit to cut the DNA and anneal the adapters. We tested the hypothesis if the cleavage motif of MuA would be a source of systematic bias in the rapid kit data, but not the ligation kit data. To test this hypothesis we took 31-bp window bovine fasta files extracted from the alignment results and utilized them to generate motif graphs. To account for mapping direction, instead of using the start positions, the end positions of reverse-complement mapped reads in the bed file were used as the start points of the reads. The background frequencies of four nucleotides were also used in the motif identification to remove the background bias from the *Bos taurus* reference genome. The read start sites from the rapid kit showed a preference for the 5'-TATGA-3' motif (Fig. 1A) while the read start sites from the ligation kit showed a preference for the recognition motif 5'-AT-3' (Fig. 1B).

To investigate the impact of the cut or ligation site bias on sequencing coverage, the percentage of each nucleotide at each position around the read start site was calculated. Both library preparation kits showed fluctuated nucleotide frequencies around the read start sites (-5 to

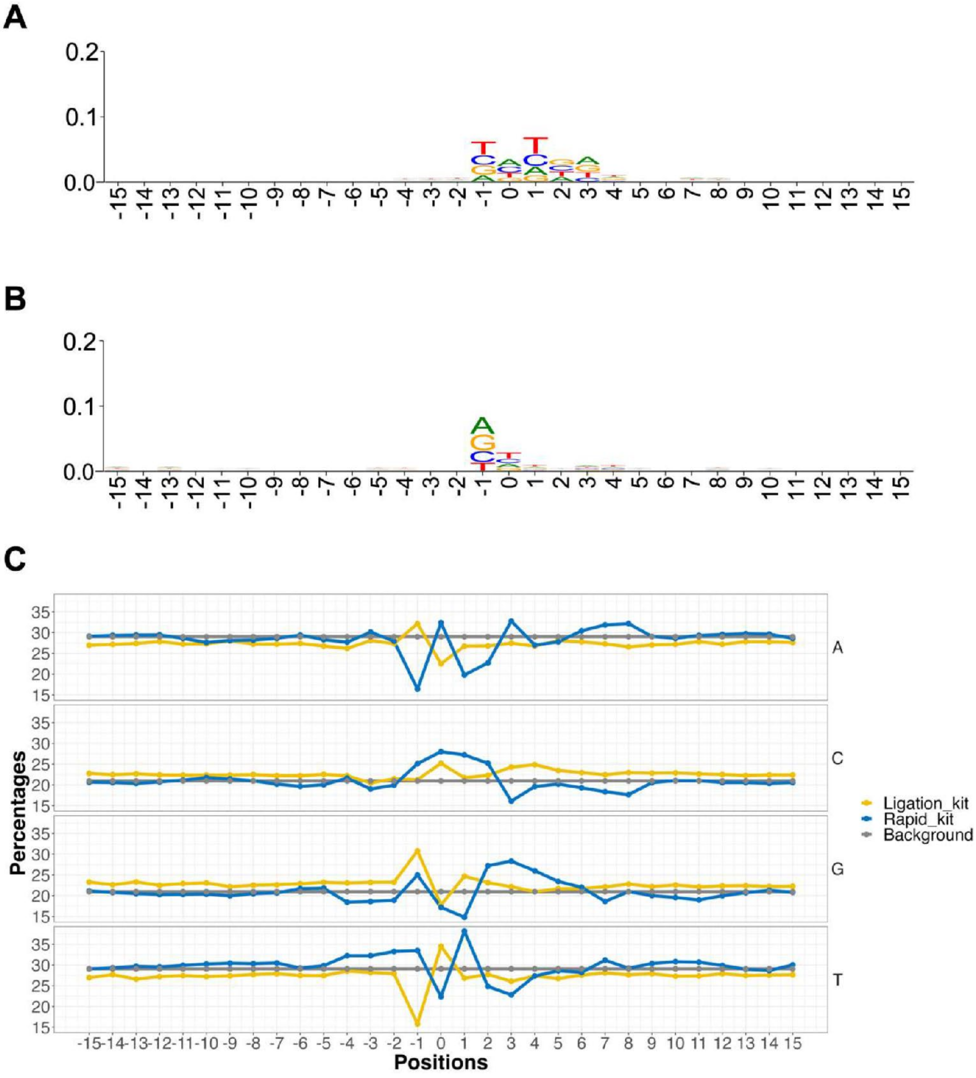

Fig. 1 Recognition motifs of (**A**) the rapid sequencing kit SQK-RBK110.96 and (**B**) the ligation sequencing kit SQK-NBD114.24, and (**C**) nucleotide frequencies in a 31 bp window between the rapid sequencing kit SQK-RBK110.96 and the ligation sequencing kit SQK-NBD114.24. The background frequency of each nucleotide in the ARS-UCDv1.2 reference genome was included to generate the recognition motifs of enzymes. 0 of the x-axis was the start position of mapped regions. The y-axis indicates the information content in bits (**A** and **B**) or the nucleotide percentages (**C**). Background frequencies for nucleotides in the ARS-UCDv1.2 reference genome were: A = 29.00%, T = 29.06%, C = 20.96%, G = 20.97%. A: Adenine; T: Thymine; C: Cytosine; G: Guanine

+ 5), indicating the sequence preferences of enzymes in two ONT sequencing protocols. In the ligation kit, constantly lower A (27.31 ± 1.71%) and T (27.34 ± 2.69%) frequencies were observed, compared to the rapid kit (Fig. 1C). In comparison, the nucleotide proportion of the rapid kit almost followed the pattern of background frequency of the reference genome, with C at 20.96 ± 2.64% and G at 20.94 ± 2.85%, respectively. Fasta files generated from 1001-bp windows (500 bp up and down stream of the ligation or interaction site) also showed a constant underrepresentation of AT content in the ligation kits (see Supplemental Fig. 1).

Enzyme-DNA interaction biases of ONT library Preparation kits

Based on the consensus recognition site of the rapid sequencing kit SQK-RBK110.96, we hypothesized that the rapid sequencing kit had a strong interaction bias in AT-rich regions. To ensure an unbiased analysis, the interaction frequency was normalized using the background frequency to counter the uneven distribution of regions with different GC values which may lead to misinterpretation of the sequencing bias (see Supplemental Fig. 2). Again, the enzyme-DNA interaction site (or read start site), was considered an indicator of whether the nucleotide that the enzymes from the ligation sequencing kits interacted with, or the binding site of the transposase for the rapid sequencing kits. The frequency of

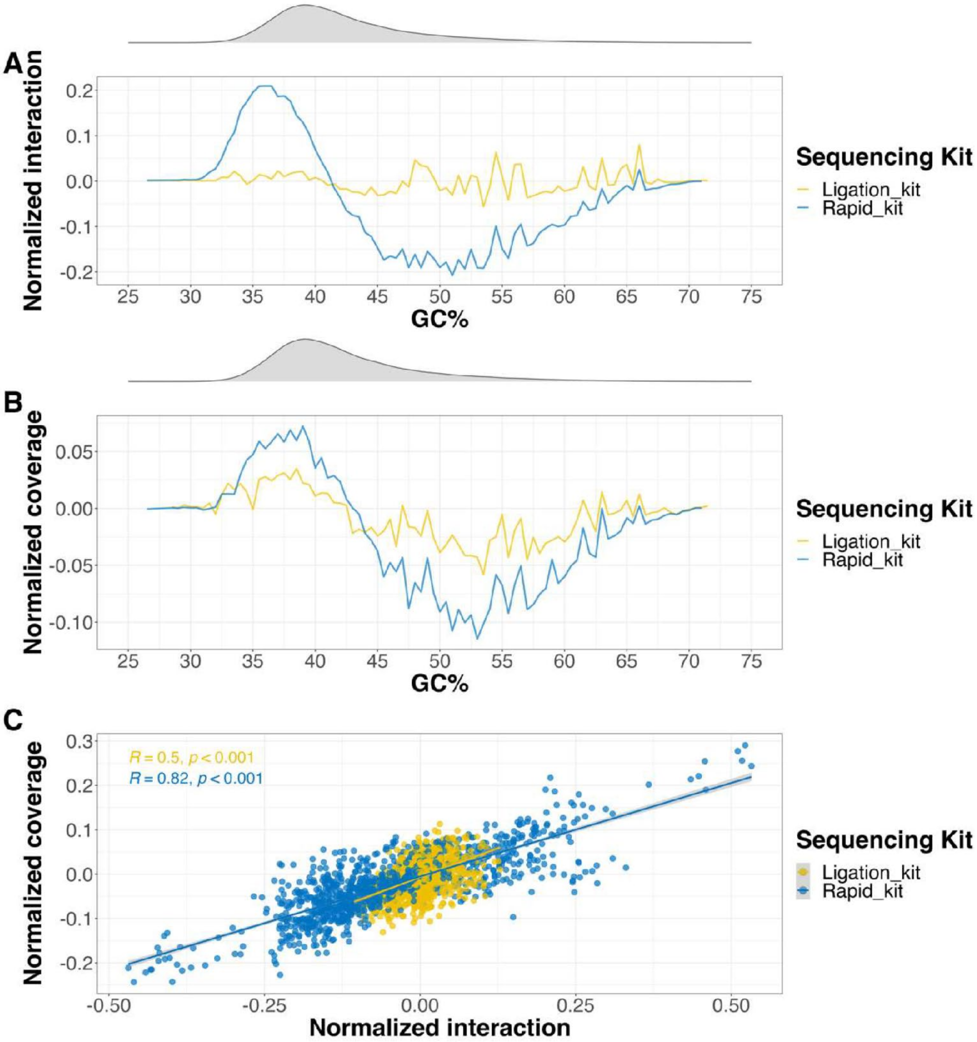

Fig. 2 (**A**) Normalized interaction frequencies and (**B**) normalized coverage, and (**C**) correlation between normalized interaction frequency and sequencing coverage at the 0.5% interval of GC contents between the rapid sequencing kit SQK-RBK110.96 and the ligation sequencing kit SQK-NBD114.24. The side plot in grey color on the top is the overall GC density of the reference genome. The x-axis indicates the GC contents (**A** and **B**) or the normalized interaction frequency (**C**), and the y-axis indicates the normalized interaction frequency (**A**) or the normalized coverage (**B** and **C**). Lines in the interval graph represented the mean normalized interaction frequency (**A**) or the mean normalized coverage (**B**). The interaction site was an indicator of whether the nucleotide that the enzymes from the ligation sequencing kits interacted with, or the binding site of the transposase for the rapid sequencing kits

the enzyme-DNA interaction across the GC distribution was significantly different between the ligation and rapid protocols (Fig. 2A). When the rapid kit was used, the enrichment of enzyme-DNA interaction at 30–40% GC regions was seen, with the highest mean normalized interaction frequency of 0.21 ± 0.10. The cleavage event of the rapid kit notably decreased in windows over 40% GC ratios, with the lowest normalized frequency at only − 0.21 ± 0.07. In contrast, the interaction distribution of the ligation kit was more even and close to the background probability across regions with different GC contents, although some fluctuations were still present.

Coverage biases of ONT library Preparation kits

Because of the longer DNA fragments produced by ONT sequencing, we expected that the bias in the enzyme-DNA interaction site locations would not translate to an overall bias in the read sequence, especially in the ligation kit, which produces longer reads than the rapid kit [16]. Similar to the interaction profiles, a rise in sequencing coverage to 0.072 ± 0.036x was seen at regions where the GC content reached 39% when the data was sequenced with the rapid kit. However, this increase was followed by a continuous coverage decline to -0.108 ± 0.058x in regions with 51% GC ratios (Fig. 2B). Combining the normalized coverage and interaction frequency data with 0.5% GC content intervals, the interaction frequencies of the rapid kit had a strong positive correlation with sequencing depth ($R = 0.82$), suggesting that the enzyme-DNA interaction bias directly led to less even sequencing coverage across the genome. On the other hand, the coverage distribution of the ligation kit was more even although it showed a similar pattern to the rapid kit. The data generated with the ligation kit showed less divergence, with R at 0.5 (Fig. 2C).

Microbiome read N50 and taxonomic classification analysis

Datasets generated using different basecalling models posed different accuracies (see Supplemental Fig. 5 and Supplemental Fig. 6), which could also affect the taxonomic classification performance. Therefore, the Australian PowerFecal and Spanish microbiome datasets basecalled under three algorithms were used to identify the effects of basecalling models. Our results demonstrated the high-accuracy basecalling (HAC) model increased the percentage of classified reads by 3.62–5.34% compared to the fast basecalling (FAST) model; and the SUP model increased the percentage of classified reads by 5.50–7.48% compared to the FAST model (Fig. 3A; Table 2; Supplemental Table 37; $P < 0.05$).

The impacts of extraction methods and sequencing methods on the read length were investigated using the Australian and Spanish datasets basecalled with super accurate basecalling (SUP) mode. Our results indicated that significantly longer DNA reads were generated by the PowerFecal kit (7469.67 ± 754.50 bp; $P < 0.001$; Fig. 4B), compared to the other two extraction methods. Additionally, our results showed the ligation kits (Australia: 7469.67 ± 754.50 bp; Spain: 4275.50 ± 1018.10 bp) outperformed the rapid kits (Australia: 4856.67 ± 111.63 bp; Spain: 2452.75 ± 1134.44 bp) by resulting in reads with higher N50 values (Fig. 4A; Australian dataset $P < 0.05$; Spanish dataset $P < 0.01$).

The impact of the DNA length, indicated by the N50, on the classification performance was investigated using the Australian and Spanish dataset basecalled with SUP mode. Both Australian and Spanish datasets showed a greater proportion of classified reads from the ligation protocols compared to the rapid kits (Fig. 3B). Notably, the ligation method in the Spanish dataset had a significantly higher percentage of classified reads (52.67 ± 9.84%) compared to the rapid protocol (32.13 ± 6.60%, $P < 0.01$). In addition, the highest classified read percentages were also observed in the samples extracted with PowerFecal kit (55.40 ± 9.52%, $P < 0.05$) (Fig. 3C). The N50 values showed a strong positive correlation to the classified proportion of the dataset ($R = 0.89$; Fig. 4C). The higher classification performance of both PowerFecal kit and ligation kits was attributed to their capacities in harvesting reads with higher N50 values, which provided more information to the classification tool for taxonomic assignments.

Microbiome analysis – alpha and beta diversity

Microbiome profiles from sequence data can be used to compare changes or differences in microbiome populations [17]. However, systematic bias in the observed population due to molecular methods could cause erroneous results. Therefore, DNA was extracted from rumen fluid by three extraction kits and sequenced by two library protocols to examine these effects. The sequencing kit effects on the microbial diversity of rumen samples were evaluated using the Australian PowerFecal and Spanish microbiome datasets. Data were basecalled under the SUP mode before analysis. Shannon indexes were calculated to identify the species richness variations among different protocols. Shannon index differences were observed between the two library protocols in both locations (Fig. 5A). Geographical variations were observed, where Australian samples had a higher species richness in the ligation protocols (6.18 ± 0.06, $P > 0.05$), but a decreased diversity was seen in the Spanish ligation samples (5.67 ± 0.23, $P < 0.05$).

The impacts of extraction kits on microbial diversity were also evaluated using the Australian rumen microbiome samples extracted using three DNA kits. Data were basecalled under the SUP mode before analysis. Likewise, the DNA extraction protocol selection also generated

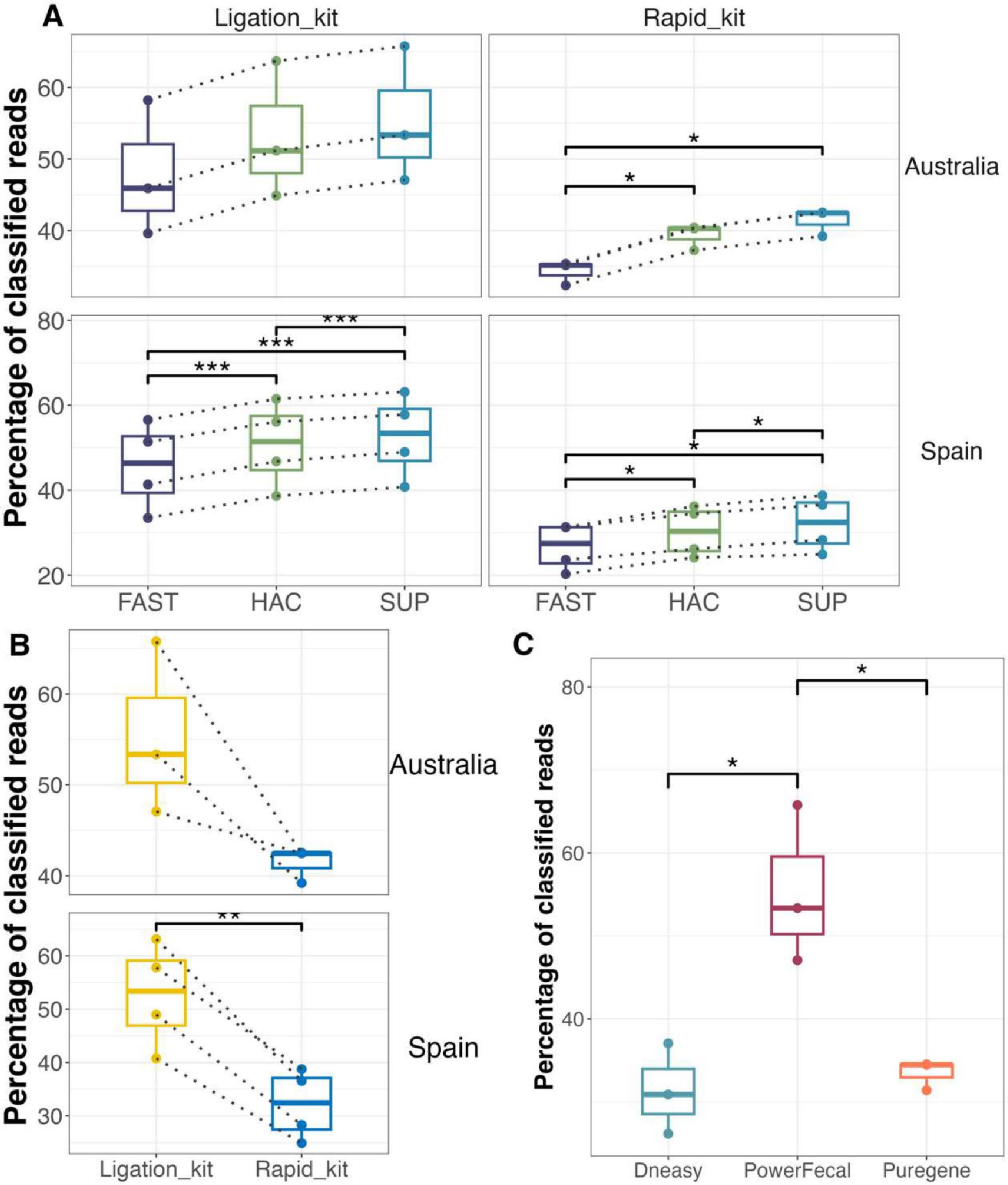

Fig. 3 Proportions of classified reads of (**A**) basecalling models, (**B**) sequencing and (**C**) extraction kits. The Australian dataset here only included the samples extracted by the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. The DNA extraction kit dataset only used samples sequenced by the ligation kit SQK-LSK109 from the Australian dataset. Data for sequencing and extraction kit analysis were basecalled under SUP mode. Dotted lines in boxplots link the samples from the same DNA sample. A t-test was used to compare the means of different basecalling, extraction, and sequencing protocols (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***. FAST: Fast basecalling; HAC: High-accuracy basecalling; SUP: Super accurate basecalling

Table 2 Classified read percentage of different basecalling models

Location	Sequencing method	Basecall mode[1]	Mean classified percentage	SD classified percentage	Increased percentage[1]
Australia	SQK-LSK109	FAST	47.91	9.47	NA
	SQK-LSK109	HAC	53.25	9.57	5.34
	SQK-LSK109	SUP	55.40	9.52	7.48
	SQK-RBK110.96	FAST	34.29	1.69	NA
	SQK-RBK110.96	HAC	39.32	1.80	5.03
	SQK-RBK110.96	SUP	41.40	1.90	7.11
Spain	SQK-LSK109	FAST	45.70	10.31	NA
	SQK-LSK109	HAC	50.76	10.11	5.06
	SQK-LSK109	SUP	52.67	9.85	6.97
	SQK-RBK004	FAST	26.63	5.57	NA
	SQK-RBK004	HAC	30.25	5.98	3.62
	SQK-RBK004	SUP	32.13	6.60	5.50

[1] FAST: Fast basecalling; HAC: High-accuracy basecalling; SUP: Super accurate basecalling

[2] The increased percentage indicated the corresponding increased classified read proportion compared to the FAST model

Shannon index divergences. The PowerFecal recorded the highest species diversity (6.20 ± 0.06, $P < 0.05$) compared to the other two DNA extraction kits (Fig. 5B). However, no significant effect on species diversity was seen from different basecalling modes (Supplemental Fig. 8; $P > 0.05$).

Principal coordinates analysis was performed to evaluate the variation among different protocols. Notable geographical differences were observed along PCo1, which accounted for 52.7% of the variation (Fig. 5C). Samples from two library preparation protocols also separated from each other. DNA extraction methods separated along PCo1, which accounted for 56.4% of the variation (Supplemental Fig. 15). However, basecalling approaches had minor impacts on microbial profiles and did not consistently correlate with PCo1 or PCo2 (Fig. 5C and Supplemental Fig. 15).

Microbiome analysis – relative abundance

The effects of library preparation kits on the kingdom-level analysis were identified by using the Australian PowerFecal and Spanish microbiome datasets. Data were basecalled under the SUP mode before analysis. At the kingdom level, Bacteria were the most abundant in both library preparation methods and higher proportions were seen in the libraries prepared with ligation protocols (Australia: 95.56 ± 0.57%, Spain: 96.99 ± 0.28%) compared to the rapid protocols (Australia: 91.42 ± 0.41%; Spain: 95.37 ± 0.67%) (Fig. 6A; Australian dataset $P < 0.001$; Spanish dataset $P < 0.05$). The rapid kits showed significantly higher abundances of Archaea (Australia: 7.65 ± 0.40%; Spain: 0.61 ± 0.03%), compared to those of the ligation kits (Australia: 3.12 ± 0.31%; Spain: 0.45 ± 0.04%) (Fig. 6A; Australian dataset $P < 0.001$; Spanish dataset $P < 0.05$). The relative abundances of Eukaryotes from the two sequencing kits varied geographically. In the Australian data, the Eukaryotes accounted for a lower proportion of reads when the rapid library preparation kit was used (0.93 ± 0.02%) compared to the ligation kit (1.33 ± 0.28%). However, the Eukaryotic proportion of the Spain rapid dataset was significantly higher (4.02 ± 0.65%) than the ligation protocol (2.56 ± 0.26%, $P < 0.05$).

It has been well-documented that the DNA extraction method affects the observed microbiome [18, 19]. The kingdom-level profile variations among DNA extraction kits were evaluated using the Australian rumen microbiome samples extracted with three extraction kits. Data were basecalled under the SUP mode before analysis. Similar to the profile of sequencing kits, Bacteria are also the most abundant kingdom in the DNA extraction dataset (Fig. 6B). Except for the Eukaryotes, significant abundance variations were observed among DNA extraction kits. Archaeal sequences were significantly higher in the PowerFecal extraction kit (3.14 ± 0.25%, $P < 0.05$) compared to the other kits. The bacterial percentage in PowerFecal protocol decreased (95.50 ± 0.61%, $P < 0.05$), compared to other extraction methods.

The basecalling effects on the kingdom-level profile were analyzed using the microbiome datasets from the Australian PowerFecal kit and Spain basecalled under three models. Basecalling algorithms showed significant effects on kingdom-level classification in the Spanish dataset (Supplemental Fig. 16; $P < 0.05$). For instance, significantly higher Eukaryotic proportions (from 1.36 ± 0.08% to 5.60 ± 0.47%, $P < 0.05$) were seen in the FAST algorithm in both Australian and Spanish datasets, compared to other basecalling models.

The effect of library preparation kits on the relative abundances of the top 11 genera in the microbiome datasets was examined. The data from Australian PowerFecal and Spain basecalled under the SUP mode were incorporated. *Prevotella* was the most abundant genus in both library preparation methods and the highest number was recorded in the Spanish rapid method (38.35 ± 5.43%)

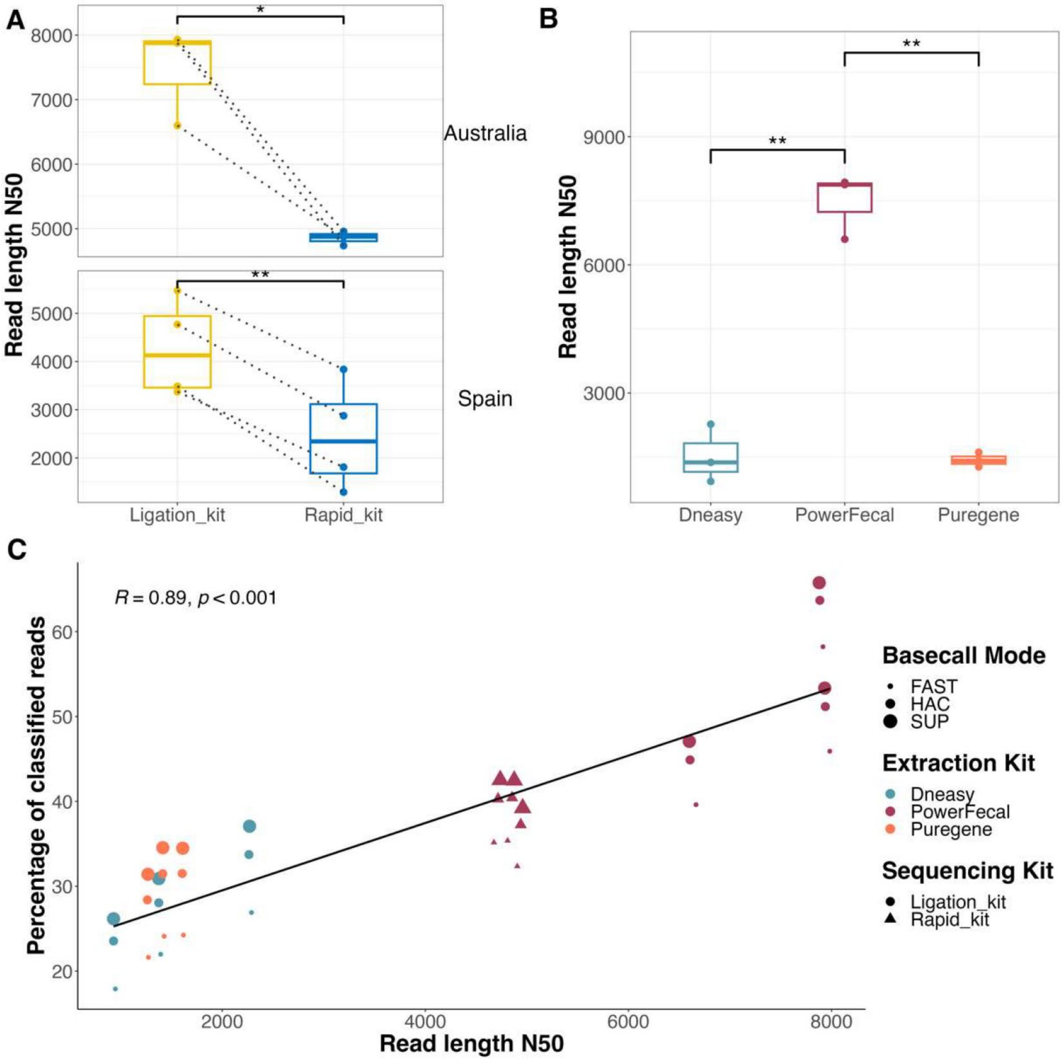

Fig. 4 Read length N50 of (**A**) sequencing and (**B**) extraction kits. (**C**) Correlation between read length N50 and classified read proportion. The Australian dataset only included the samples extracted by the PowerFecal kit for sequencing kit analysis. The Australian dataset used the ligation kit SQK- LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. The DNA extraction kit dataset only used samples sequenced by the ligation kit SQK-LSK109 from the Australian dataset. Data for sequencing and extraction kit analysis were basecalled under SUP mode. Dotted lines in boxplots link the same DNA sample. A t-test was used to compare the means of different extraction and sequencing protocols (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***. R is the Pearson correlation coefficient between read length N50 and classified read proportions. FAST: Fast basecalling; HAC: High-accuracy basecalling; SUP: Super accurate basecalling

(Fig. 7A). *Xanthomonas*, a bacterial genus with around 65% GC content, was identified at a significantly lower level in the rapid kit protocol (5.05 ± 2.76%) in the Spanish dataset, compared to the ligation protocol (14.07 ± 3.63%, $P < 0.01$). This pattern was also seen in the Australian dataset, although the difference was not significant ($P > 0.05$). On the other hand, the genus, *Photobacterium*, with around 44% GC ratio in its genome, showed lower abundance in the ligation kits (Australia: 0.06 ± 0.01%; Spain: 0.09 ± 0.02%), compared to the rapid protocols (Australia: 1.72 ± 1.06%; Spain: 0.20 ± 0.04%) (Fig. 7A; Spanish dataset $P < 0.05$). Interestingly, the proportion of *Anabaena* genus (around 38% GC) in the rapid kit was lower (Australia: 0.10 ± 0.03%; Spain: 0.53 ± 0.69%), compared to the ligation kit (Australia: 0.49 ± 0.29%; Spain: 2.59 ± 0.76%) (Fig. 7A; Spanish dataset $P < 0.05$).

The relative abundance profiles of genera under the Archaea kingdom between two library preparation protocols were also investigated. The data from Australian PowerFecal and Spain basecalled under the SUP mode were used. The dominant genus in both Australian and Spanish datasets was *Methanobrevibacter* (Fig. 7B). However, in both datasets, the rapid kit showed significantly enriched *Methanobrevibacter* abundance (Australia:

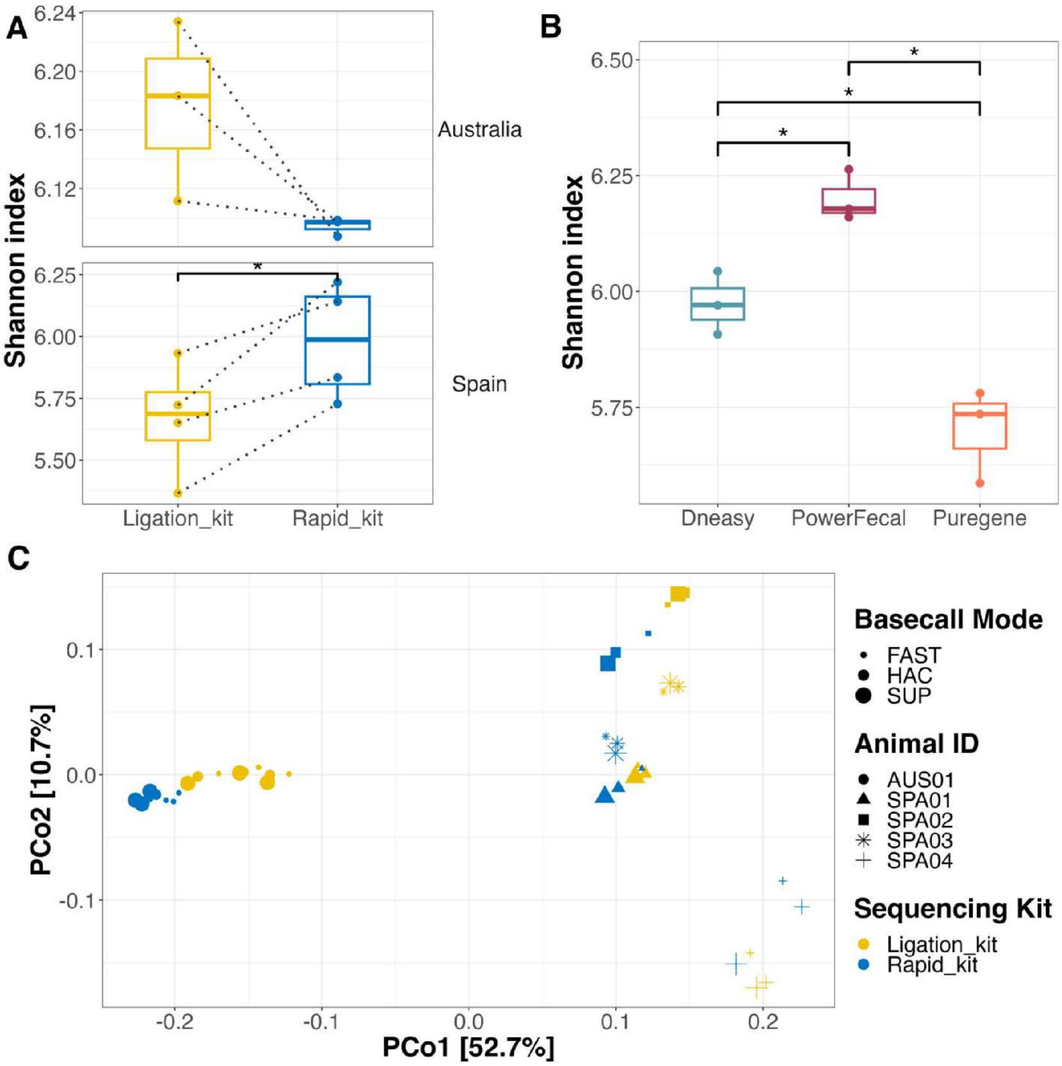

Fig. 5 Shannon indexes of rumen metagenome data for (**A**) library preparation kits and (**B**) DNA extraction kits, and (**C**) principal coordinates analysis. The Australian dataset for sequencing kit analysis only included the data from the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96. The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. The DNA extraction kit dataset only used samples sequenced by the ligation kit SQK-LSK109 from the Australian dataset. Data for sequencing and extraction kit analysis were basecalled under SUP mode. Dotted lines in boxplots link the same DNA sample. Animal ID AUS01 was from the Australian dataset, while the other animals were from the Spanish dataset. A t-test was used to compare the means of different extraction and sequencing protocols (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***

87.40 ± 0.58%; Spain: 19.76 ± 6.11%), compared to the ligation kit (Australia: 77.11 ± 1.60%; Spain: 11.89 ± 3.93%) (Fig. 7B; Australian dataset $P < 0.01$; Spanish dataset $P < 0.01$). In addition, the *Candidatus Methanomethylophilus* was significantly depleted when the rapid kit was used (Australia: 0.38 ± 0.09%; Spain: 1.24 ± 1.00%), compared to the ligation kit (Australia: 1.37 ± 0.39%; Spain: 4.09 ± 2.01%) (Fig. 7B; Australian dataset $P < 0.05$; Spanish dataset $P < 0.05$).

The genus-level microbial profile variations of extraction kits were identified using the Australian rumen microbiome samples extracted from three DNA kits. Data were basecalled under the SUP mode before analysis. The proportions of most genera showed significant differences among various DNA extraction kits (Fig. 8). Similar to the profiles of library preparation protocols, *Prevotella* was the dominant genus in all extraction protocols, with almost 20–40 times proportions higher than other genera. However, the *Prevotella* abundance was significantly lower in the PowerFecal method (28.23 ± 2.28%, $P < 0.05$) compared to other protocols. Interestingly, gram-positive genera, such as *Clostridium* (1.55 ± 0.14%) and *Streptomyces* (3.07 ± 0.19%), showed increased abundances in the PowerFecal protocol compared to other extraction methods although no significant difference was found ($P > 0.05$). However, although a Gram-positive specific

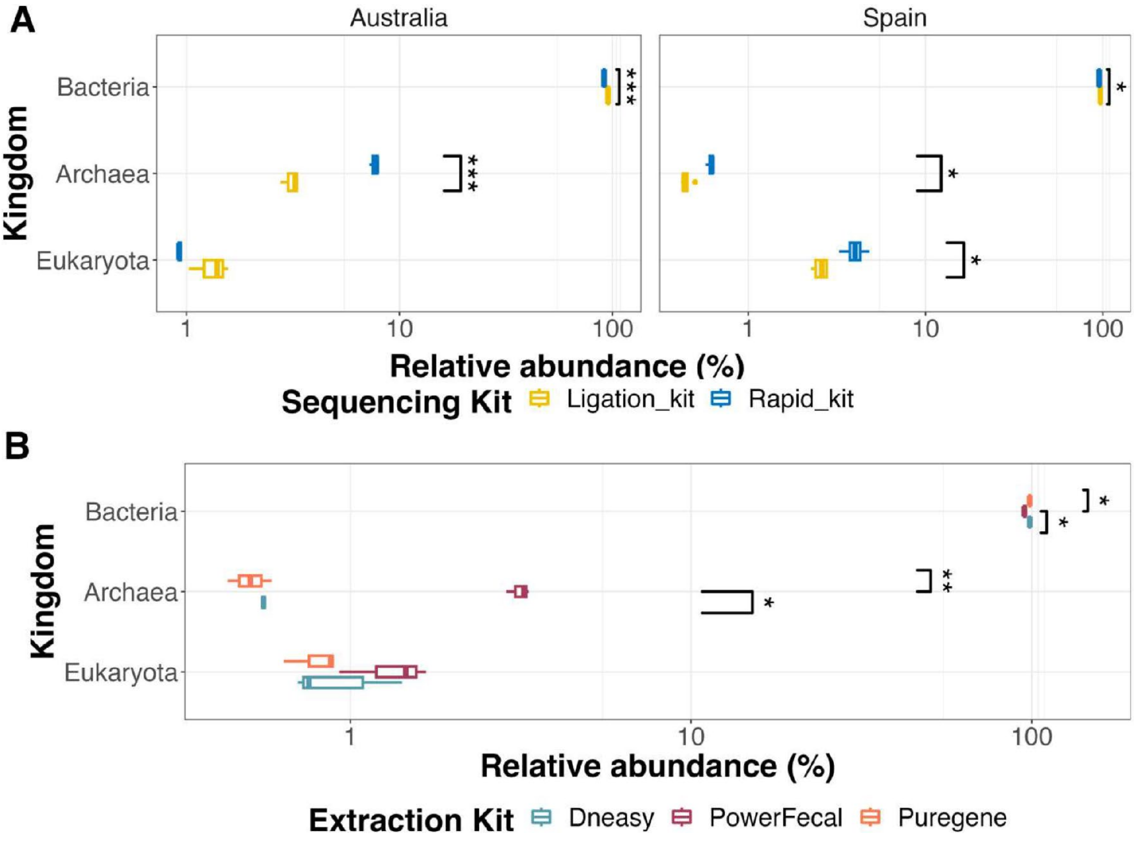

Fig. 6 Kingdom relative abundances of **(A)** sequencing and **(B)** extraction kits. The Australian dataset for sequencing kit analysis only included the data from the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. The DNA extraction kit dataset only used samples sequenced by the ligation kit SQK-LSK109. Data were basecalled under SUP mode. A t-test was used to compare the means of different extraction and sequencing protocols (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***

instruction was used in the Puregene protocol, the proportions of *Streptomyces* (1.78 ± 0.05%) and *Faecalibacterium* (0.54 ± 0.01%) were lower than the ones of DNeasy (*Streptomyces*: 2.56 ± 0.55%, $P > 0.05$; *Faecalibacterium*: 0.88 ± 0.31%, $P > 0.05$). The basecalling models also showed significant effects on genus profile characterization (Supplemental Fig. 17, Supplemental Fig. 20, and Supplemental Fig. 22; $P < 0.05$).

Metagenomic assembly statistics

The effects of library preparation kits on metagenome assembly were examined using the microbiome data sequenced by both ligation and rapid kits. Data were basecalled under SUP mode before analysis. The N50 of subsampled reads for assembly was calculated. Subsampled reads were assembled, followed by the quality evaluation. Higher contig numbers (89.33 ± 6.81) and assembly N50 values (24076.00 ± 6688.38 bp) were seen in the Australian rapid kit, compared to the ligation samples (Contig number: 20.33 ± 4.04, $P < 0.001$; N50: 19943.67 ± 2736.23 bp, $P > 0.05$). However, these two values were higher for the ligation protocol (Contig number: 349.75 ± 235.53; N50: 15750.75 ± 6160.70 bp), compared to the rapid kit (Contig number: 291.50 ± 174.57, $P > 0.05$; N50: 9370.50 ± 2710.62 bp, $P > 0.05$) in the Spanish samples. The read length N50 was negatively correlated with the contig numbers ($R = -0.84$ for both ligation and rapid kits; Fig. 9A). A positive correlation was seen between read length N50 and contig N50, with $R = 0.85$ for the rapid kit and $R = 0.68$ for the ligation kit (Fig. 9B).

Bacterial DNA methylation characterization

The effects of library preparation kits on bacterial DNA methylation characterization in metagenomic samples were investigated using the microbiome data sequenced by both ligation and rapid kits. Data were basecalled under the SUP mode before the analysis. The ligation kits showed significantly higher mapping proportions (from 1.48 ± 0.44% to 12.28 ± 3.42%) and sequencing coverages (from 9.85 ± 8.38x to 71.33 ± 0.78x) to the three bacterial reference genomes, compared to the rapid protocols (mapping proportions: from 0.88 ± 0.42% to 8.08 ± 0.54%, $P < 0.01$; sequencing coverages: from 8.10 ± 7.12x to 71.34 ± 0.74x, $P < 0.05$; Supplemental Fig. 24 and

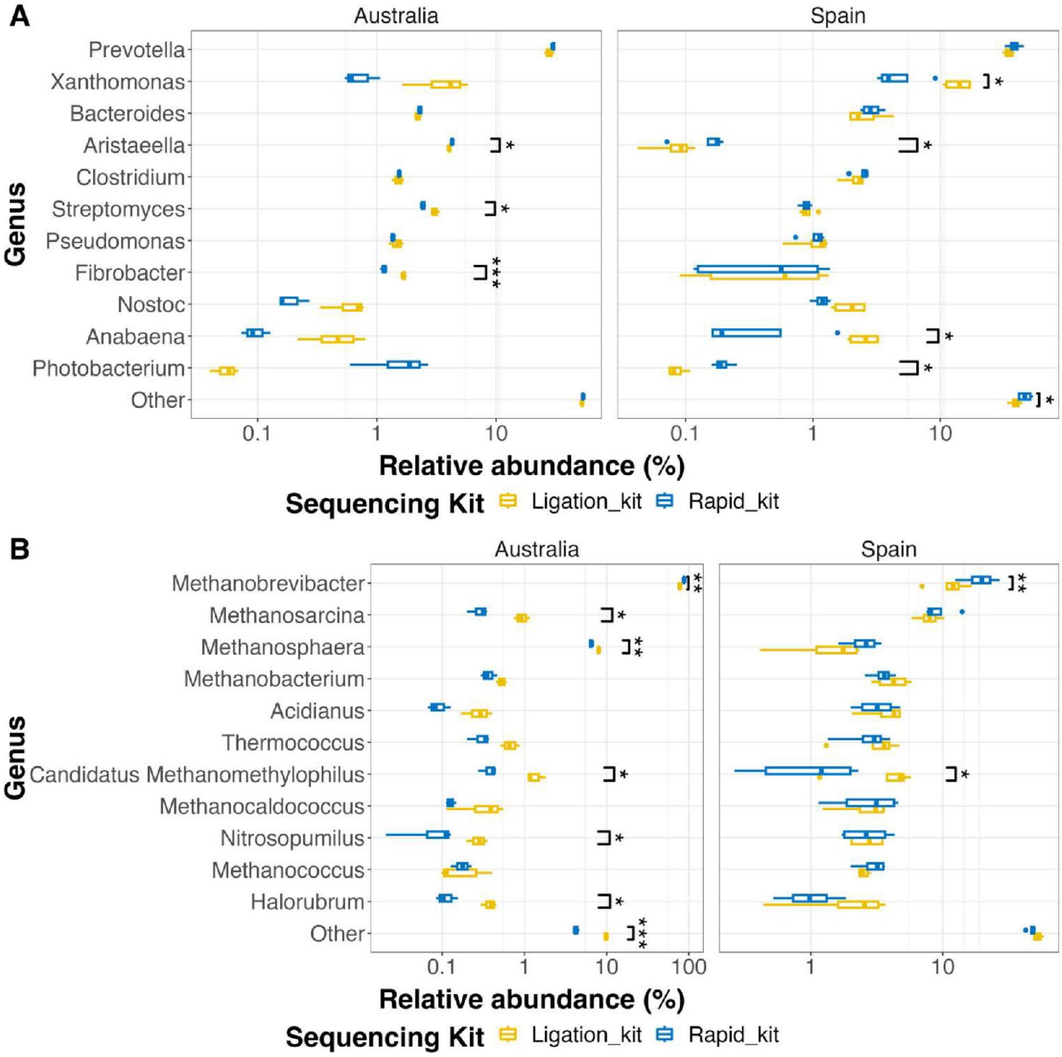

Fig. 7 **(A)** Bacterial and **(B)** archaeal genera relative abundances of sequencing kits. The Australian dataset for sequencing kit analysis only included the data from the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. Data were basecalled under SUP mode. A t-test was used to compare the means of different sequencing protocols (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***

Supplemental Fig. 25). After removing duplicated detected sites, the two sequencing protocols shared 15.6 to 61.0% unique modified sites across the three species (Fig. 10A and B, and Fig. 10C). These overlapped methylated position numbers were much higher than expected (Supplemental Tale 39; $P < 0.001$). In addition, the numbers of unique methylation sites were not significantly different between the two library preparation methods across three bacterial genomes (Fig. 10D).

Mock community analysis – read length N50 and relative abundance

To further investigate the biases of the ligation and rapid kits, two mock communities (mock 1 and mock 2) were constructed using the extracted DNA from three bacterial species, namely *Lactobacillus acidophilus* (34.5% GC), *Escherichia coli* (50% GC), *Bifidobacterium gallinarum* (64% GC). In mock 1, each species shared the same amount of DNA by weight (1: 1: 1). In mock 2, the DNA composition followed the ratio of *L. acidophilus*: *E. coli*: *B. gallinarum* = 1: 2: 3. The read length N50 significantly varied across different bacterial species and sequencing kits (Supplemental Fig. 26). The N50 values in *L. acidophilus* (from 1043.67 ± 59.03 bp to 2010.67 ± 94.31 bp) were significantly lower than the other species (from 4573.00 ± 122.11 bp to 8256.00 ± 14.93 bp; $P < 0.05$) in both the ligation and rapid kits. On the other hand, the N50 of the rapid kit (from 1043.67 ± 59.03 bp to 6053.33 ± 90.00 bp) were significantly lower than the ligation ones (from 1620.00 ± 94.00 bp to 8256.00 ± 14.93 bp;

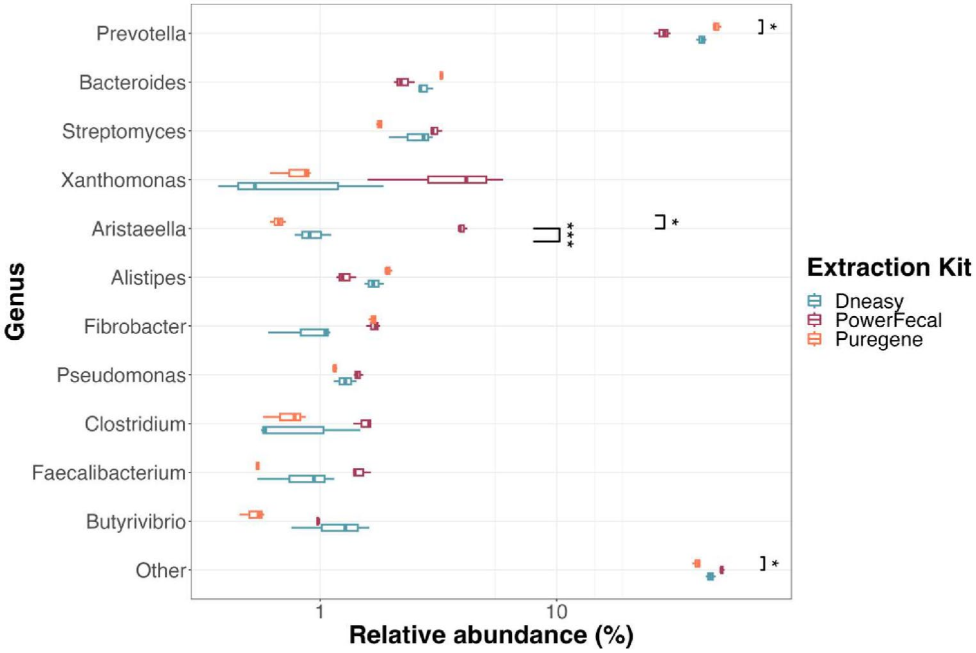

Fig. 8 Bacterial genus relative abundances of extraction kits. The DNA extraction kit dataset only used samples sequenced by the ligation kit SQK-LSK109 from the Australian dataset. Data were basecalled under SUP mode. An unpaired t-test was used to compare the means of different extraction protocols. *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***

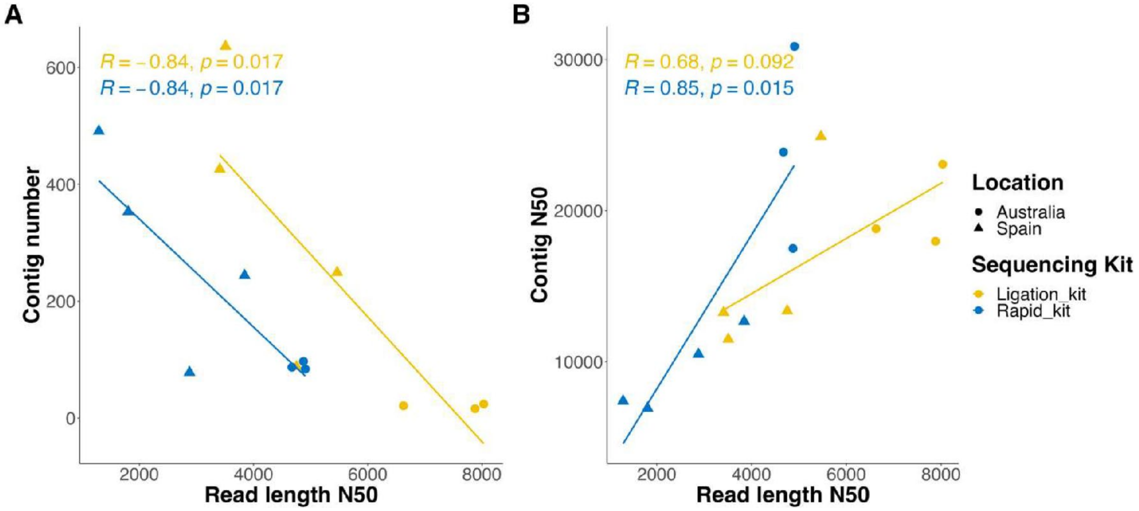

Fig. 9 Contig number, contig N50, and read length N50 analysis between two sequencing kits. The Australian dataset for sequencing kit analysis only included the data from the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. Data were basecalled under SUP mode. R is the Pearson correlation coefficient between read length N50 and contig number, and the Pearson correlation coefficient between read length N50 and contig N50

$P < 0.05$). These read length divergences may affect the relative abundance analysis using read-based classification approaches.

To exclude the potential effects of DNA fragmentation during the DNA extraction, the relative abundance was calculated through the division of the primary aligned base number of each species by the total primary aligned base number. The differential relative abundance was estimated by the DNA proportion of the mock community subtracted from the corresponding relative abundance. Both the sequencing kits showed sequencing biases to each bacterial species across various mock communities (Fig. 11). Despite the lower GC content (34.5%) of *L. acidophilus*, significantly higher proportions were observed in the ligation kit (mock 1: +12.06 ± 1.05%; mock 2: +9.81 ± 2.54%), while the ones in the rapid

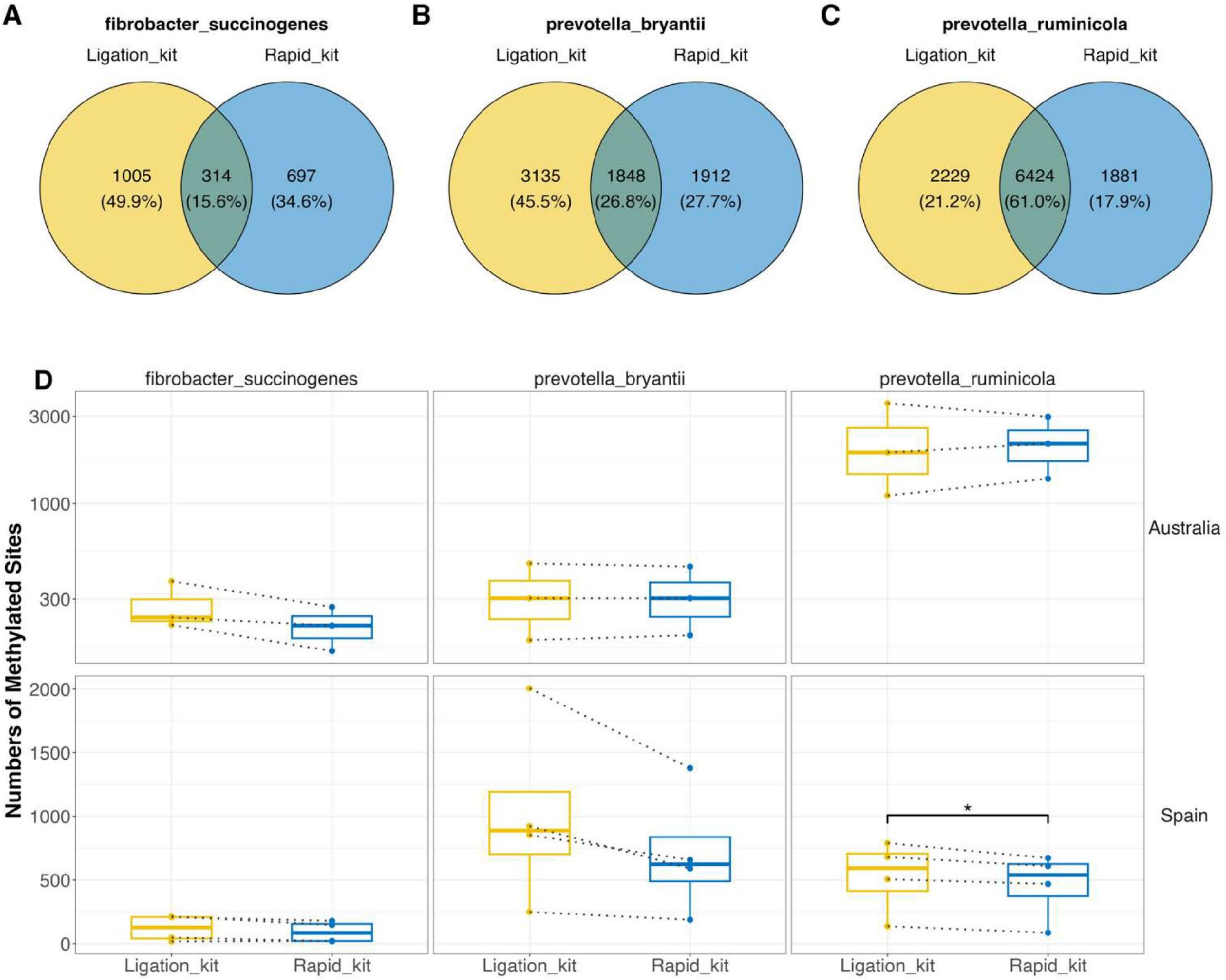

Fig. 10 The overlapping methylation sites in three bacterial genomes (**A**) *Prevotella ruminicola 23*, (**B**) *Prevotella bryantii strain TS1-5*, and (**C**) *Fibrobacter succinogenes subsp. succinogenes S85*, and (**D**) detected methylated numbers between two sequencing kits. The Australian dataset for sequencing kit analysis only included the data from the PowerFecal kit. The Australian dataset used the ligation kit SQK-LSK109 and the rapid sequencing kit SQK-RBK110.96; The Spanish dataset used the ligation sequencing kit SQK-LSK109 and the rapid sequencing kit SQK-RBK004. Data were basecalled under SUP mode. Dotted lines in boxplots link the same DNA sample. A t-test was used to compare the mean between two library preparation kits (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset)

kit showed significantly lower (mock 1: -8.01 ± 4.82%, $P < 0.05$; mock 2: -7.77 ± 1.01%, $P < 0.01$). Both sequencing kits showed notably increased percentages of *E. coli*, except for the one in the ligation kits of the mock 1 community (+ 0.14 ± 0.70%). Although the *B. gallinarum* DNA amount was higher in the mock 2 community, both sequencing kits showed decreased proportions in this bacterial species (ligation kit: -16.24 ± 2.17%; rapid kit: -9.82 ± 1.20%). In addition, the mock community divergences also indicated that none of the sequencing kits produced the expected profile. In mock 1, each species shared the same amount of DNA by weight (1: 1: 1). However, the relative abundance ratio from the ligation kit was around *L. acidophilus*: *E. coli*: *B. gallinarum* = 2: 2: 1, while the one from the rapid kit was approximately 1: 2: 1 (*L. acidophilus*: *E. coli*: *B. gallinarum*). In mock 2, the defined ratio was *L. acidophilus*: *E. coli*: *B. gallinarum* = 1: 2: 3. However, the ratios from the rapid kit and ligation kit were approximately 1: 6: 4 (*L. acidophilus*: *E. coli*: *B. gallinarum*) and 1: 2: 1 (*L. acidophilus*: *E. coli*: *B. gallinarum*), respectively.

Discussion

In this study, we identified a distinct recognition sequence (5'-TATGA-3') of the MuA transposase utilized by the ONT rapid kit. Biases in AT-rich regions were recognized in the rapid kit and the AT under-representation problems were minimized in the ONT ligation kit. Furthermore, the ligation kit outperformed the rapid kit by posing less bias at interaction site preference and a more

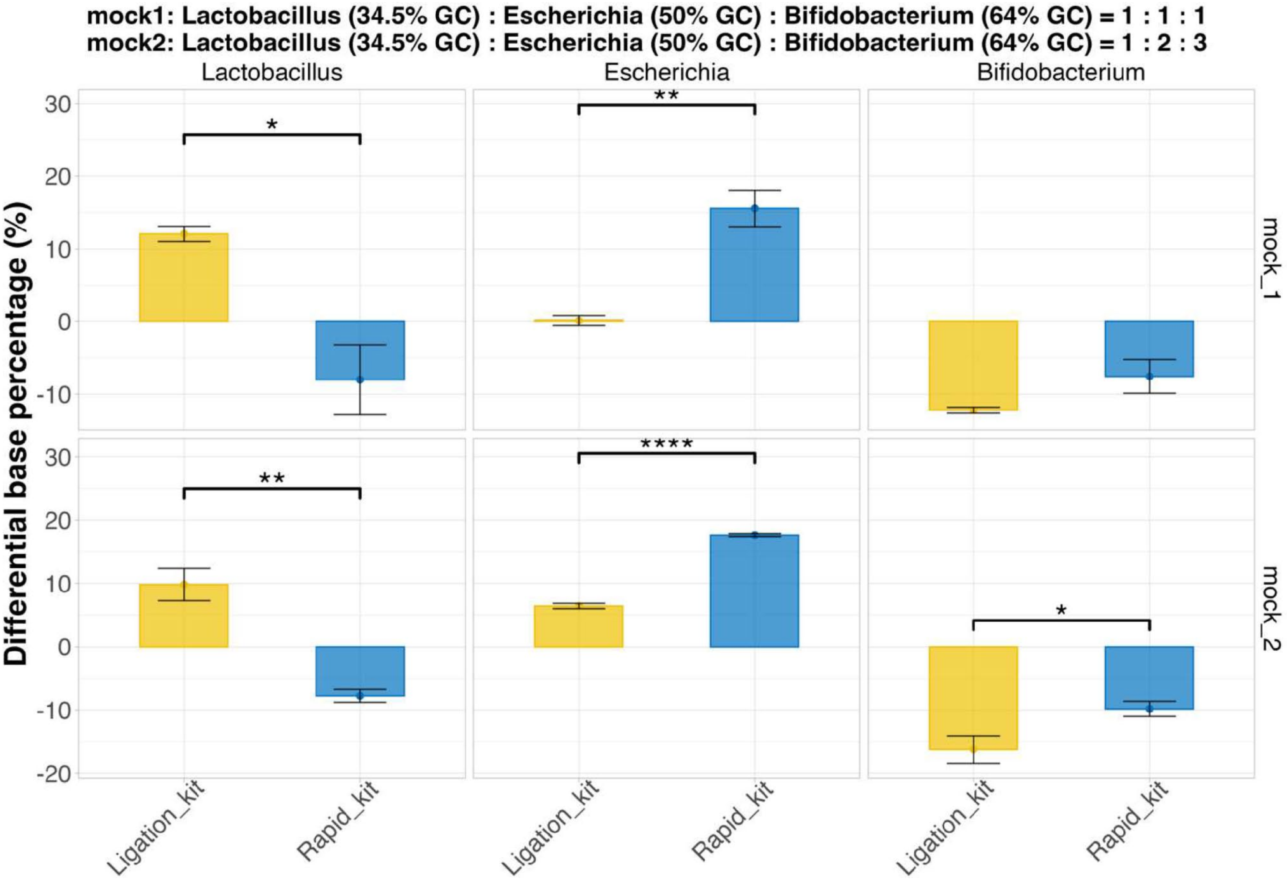

Fig. 11 The differential relative abundances of *Lactobacillus acidophilus* (34.5% GC), *Escherichia coli* (50% GC), and *Bifidobacterium gallinarum* (64% GC) of mock communities. The DNA ratio in mock 1 is *L. acidophilus*: *E. coli*: *B. gallinarum* = 1: 1: 1, while the ratio in mock 2 is *L. acidophilus*: *E. coli*: *B. gallinarum* = 1: 2: 3. This dataset used the ligation sequencing kit SQK-NBD114.96 and the rapid sequencing kit SQK-RBK114.24. The differential relative abundance was calculated by the DNA proportion of the mock community subtracted from the corresponding relative abundance, therefore if there is no bias all samples would cross the y-axis at 0. An unpaired t-test was utilized to compare the differential relative abundance between the two sequencing protocols. *P*-values < 0.05: *; *P*-values < 0.01: **; *P*-values < 0.001: ***

even sequencing coverage distribution. In the microbiome analysis, we observed notably different microbial profiles at the Kingdom and Genus levels when the samples were processed with different DNA extraction methods, Oxford Nanopore library preparation protocols, and basecalling models. Meanwhile, most of the assembly and methylation analysis results showed no significant divergence between sequencing kits. In addition, notable differential relative abundances of both Oxford Nanopore library preparation protocols across all bacterial species were observed in the mock community analysis.

A simple recognition motif (5'-AT-3') of the ligation kit was identified in this study, while the involvement of multiple enzymes in this method made the recognition sequence attribution challenging [13]. We also found lower AT contents around the read start sites and the following analysis observed a higher enzyme-DNA interaction frequency (0.07 ± 0.07) in high GC regions (60–80%) in the ligation kit. Additionally, the following sequencing depth analysis showed that regions with 25–30% GC had almost the same probability of the background frequencies. The under-representation of AT regions of human DNA was previously identified in the NEBNext Ultra II Library Prep Kit of Illumina without clear reason [8, 20]. A subsequent study found that increased temperature during enzymatic reactions caused exonuclease degradation in low-GC DNA fragments [13]. Interestingly, regardless of the version or barcode option, the ONT ligation sequencing kits utilize NEBNext Ultra II End-prep enzyme mix for the end-prep step, which is a similar approach to the NEBNext Ultra II Library Prep Kit for Illumina DNA library construction [5–7]. The enzyme-DNA interaction preference results of the ligation kits in this study were consistent with the findings of a study where they observed Taq DNA polymerase showed higher kinetic rates of dA-tailing in GC-rich terminal sequences [13]. Meanwhile, the improved sequencing coverage at 25–30% GC frames was also similar to the previous study using immobilized enzymes to modify the AT under-representation problems [13]. Generally, the major practical difference between Illumina and ONT sequencing is that ONT can sequence much longer DNA fragments with over 10 kb [1]. Therefore, we deduced the mitigation of sequencing depth decline at low-GC regions was due to the long reads from ONT which can spread more areas of the genome. However, we cannot identify the sequencing depth of the region with GC content lower than 26% due to the limitation of the ARS-UCDv1.2 reference genome where the lowest GC content is 26.11%. While the data generated from this study was from different individual animals than the reference genome, this is unlikely to be a major source of variation at the genome-wide scale.

We identified a 5-bp conserved recognition sequence, 5’-TATGA-3’, of MuA utilized in the ONT rapid kits by extracting the reference genome sequences corresponding to the mapped reads. In the subsequent analysis, we identified reduced insertion frequencies and coverage of MuA in GC-rich sequences. Transposases are enzymes capable of translocating the DNA fragments (transposons) to other regions of the genome [21]. Hence, these enzymes, including Tn5 and MuA, have been widely used in DNA sequencing because of their high efficiency in library preparation. For example, the ONT rapid kit can potentially reduce the library preparation time down to 60 min, in contrast to the ligation approach, which requires a longer preparation time. However, the sequencing biases introduced by transposases have been reported [9, 14]. For example, Tn5 was one of the popular transposases employed in Illumina library preparation kits [22]. The recognition motif of wild-type Tn5 was first identified as 5’-A-GNTYWRANC-T-3’ [23]. A similar recognition motif of Tn5 transposase was also identified in Nextera XT and Illumina Field DNA Prep kits [24]. Due to the potential effect of cleavage preference of the Tn5 enzymes, some reports identified a higher sequencing depth in AT-rich regions using Nextera XT kits, contributing to the negative effects in downstream genotyping and microbiome analysis [9, 14]. In contrast, some studies also indicated the impact of the Tn5 transposase on subsequent analysis was minor [25, 26]. In addition to Tn5, previous studies also identified that the MuA transposase strongly preferred trinucleotide CGG sites and produced a 5-bp duplication [11, 27]. Another study also demonstrated the consensus recognition motif of MuA was 5’-C(C/T)(G/C)(A/G)G-3’ [28]. However, these recognition motifs were distinct from our findings. In a recent study using the rapid sequencing kit to analyze the adeno-associated virus (AAV) single-stranded DNA, they also found some regions with higher MuA insertion probabilities [29]. These insertion biases were claimed to be weakly correlated to the GC ratios yet no correlation analysis was performed [29]. Likewise, a previous study using two *Aminobacter* and *Fusobacterium* with different GC backgrounds indicated that the rapid kit was not affected by the GC bias [30]. These findings from adeno-associated virus and bacterial samples were inconsistent with the sequencing bias results of MuA in our study. However, the nucleotide frequency fluctuations near the MuA-DNA interaction site in our study were nearly consistent with another report where they found that MuA showed a preference for pyrimidines (C or T) and purines (A or G) around the interaction sites [31].

We investigated whether DNA extraction, sequencing, and basecalling methods could generate different microbiome or bacterial methylation profiles. Generally, we observed microbial profile divergences from various sources, with the most variation introduced by DNA extraction kits, followed by library preparation protocols, and basecalling models. Additionally, we observed a significantly higher classification efficiency with the increased basecalling accuracy. We also found increased proportions of Gram-positive bacteria in data generated from the PowerFecal extraction kit, which incorporated both chemical and mechanical methods for cell lysis. Previous reports also found the beat-beating process could reveal higher abundances of Gram-positive bacteria [18, 19], which generally have thicker cell walls compared to the negative ones. In addition, significantly higher overall species diversity was seen in the PowerFecal data. While higher percentages of archaea were seen in the PowerFecal compared to other extraction methods, subsequent Alpha diversity analysis down to the Archaeal level showed a lower Shannon index in the PowerFecal protocol compared to the other DNA kits. This may be due to a notably higher proportion of archaeal reads in PowerFecal being classified as *Methanobrevibacter* genus. In the rumen, archaea are the sole methane synthesizer [32]. In this case, greenhouse gas research using a rumen microbiome matrix needs to consider the potential archaeal profile variations from different DNA isolation kits, because these divergences can bring biases to the subsequent analysis, such as methane prediction results.

In this study, rumen microbial profile variations were seen between the two Oxford Nanopore sequencing protocols, with species classifications tending towards bacteria in the ligation protocols and archaea in the rapid protocols. The alpha diversity analysis demonstrated the ligation protocols resulted in higher diversity than the rapid protocols in the Australian group, but the opposite pattern was observed in the Spanish dataset. From the sequencing bias analysis, we identified that the rapid kit showed higher interaction frequency and sequencing coverage at low GC content regions, compared to the ligation kit. In the microbiome relative abundance analysis, the bacterial genus *Xanthomonas* (65% GC) showed relative abundance discrepancies between the two library preparation protocols. These profile variations were consistent with the GC bias we identified where the transposase-based kit showed depleting coverage at sequences with high GC sequence regions and increased coverage in regions with low GC contents. Despite the expected higher abundance of the *Anabaena* genus (39% GC) in the rapid kits based on the enzyme-DNA interaction and sequencing coverage bias we identified, its abundance was depleted in the rapid kit as compared to the ligation kit. However, due to the limited sequencing depth of our microbiome samples, our study solely targeted the profile variation of the genus with high abundance. For low-abundance genera, a higher sequencing depth is necessary. In addition, other various external factors for the microbiome samples, such as uncharacterized species in the database, increased the difficulty of explaining the reasons behind these findings.

A distorted mock community profile was observed in both ONT ligation and rapid protocols in this study. The relative abundances of *B. gallinarum* (64% GC) in both sequencing kits were notably lower than expected, where the values of the ligation kits were even lower than the rapid kits, although our GC bias results showed the ligation kit showed a higher interaction frequency in high GC regions (60–80%). We also found the proportions of *L. acidophilus* (34.5% GC) were notably lower than the benchmark in the rapid kit in the mock community analysis, which was also inconsistent with the biases we found in the rapid kit. However, these mock community findings were similar to our microbiome results. In addition to distorted relative abundances, significantly different read length N50 values were also seen across all bacterial species, with the highest recorded in *B. gallinarum* and the lowest in *L. acidophilus*. Generally, long DNA molecules provide more information, such as structural variations [1], while short DNA fragments or a high GC content background may cause the loss of high AT content interaction motifs (in this case, 5’-TATGA-3’). Therefore, we speculated that the lower proportion of *L. acidophilus* in the rapid kit resulted from the motif-based nature of the transposase and the missing interaction motif due to short DNA; and the lower occurrence of recognition motif in the *B. gallinarum* reads due to the extremely high GC content could contribute to the depleted *B. gallinarum* proportion in the rapid kit. Multiple studies indicated DNA polymerases showing a strong preference for short DNA molecules during PCR amplification [33, 34]. In our further analysis, the read length N50 was negatively correlated to the differential relative abundance in the ligation kit (R=-0.76, $P < 0.001$). Interestingly, several DNA polymerases are involved during the ONT ligase-based library preparation. Therefore, although no amplification was involved during the library preparation, we speculated that the DNA polymerases also have a strong preference for short DNA fragments during end blunting and dA tailing, which caused the distorted mock profile from the ligation kit. However, the exact reason behind this mock profile variation needs to be further investigated. Still, we were able to identify that sequencing kit protocols can induce variation in microbiome analyses.

We identified the read length variations among different DNA extraction and sequencing protocols, with the longest DNA observed in the PowerFecal extraction and ligation sequencing kits. In addition, we also found that these varying read lengths finally affected the classification efficiency and assembly results, with decreased taxonomic classification performance in shorter reads. DNA extraction kits involving mechanical steps, such as grinding and beat beating, can increase DNA yields and meanwhile cause DNA shearing [35]. However, in our study, although bead beating was included, the longest reads were still observed in the PowerFecal extraction protocol compared with other extraction protocols without the mechanical lysis step. In addition, due to the insertion activity of transposase, the DNA molecules are shortened in the ONT rapid kit compared to the ONT ligation kit [16], which was consistent with the findings in our study. A previous study found that longer DNA reads from Oxford Nanopore sequencing increased the classified read proportions [36], which was also consistent with our results. In our subsequent assembly analysis, we found a positive correlation between read length N50 and contig N50 was seen in the assembly analysis. Generally, contig N50 is an important metric for evaluating the assembly performance because a higher N50 results in better gene function prediction and genome reconstruction [37]. Therefore, protocols preserving longer DNA are preferred for the microbiome study, in terms of taxonomic classification and metagenomic assembly. For example, DNA extraction protocols with less mechanical cell lysis and DNA library preparation protocols without DNA shearing steps (such as long-read sequencing protocols) allow the preservation of longer DNA.

The ligation-based protocols showed a higher number of N6-Methyladenosine (6mA) methylated sites, compared to the rapid protocols, although the difference was not significant. Previous studies demonstrated library preparation protocols can introduce biases in 5-Methylcytosine (5mC) characterization, due to PCR amplification and bisulfite conversion time [38–40]. Both Oxford Nanopore sequencing methods (ligation and rapid) do not have these enzymatic interactions with the methylated sites; rather the enzymes only interact with the ends of the reads. To date, no study has performed a comparison between the 6mA profile variations among different Oxford Nanopore sequencing protocols. Our study found a high correlation ($R = 0.82$) between DNA-enzyme interaction and sequencing coverage in the rapid kit, compared to the ligation kit. In addition, the observed overlapped methylation position number between the two sequencing kits was significantly higher than expected. Hence, we deduced the lower number of the detected 6mA sites in the rapid kit was due to the lower sequencing coverage, which in turn is a reflection of the slight bias in the microbiome sequence data due to the transposase recognition site bias. While missing methylation data can affect downstream analysis, such as epigenetic clocks which use methylation data to calculate age [41], the results here suggest that the effects would be minor. Still, it is prudent to include the sequencing kit used as a co-variate in any statistical analysis, as the bias induced by the transposase could alter the sequencing coverage, and thus affect the accuracy of the methylation calls due to a lower number of reads being used to calculate the methylated: non-methylated ratio. Nonetheless, except for the mock community, the samples used in this study were all bovine-related. The effects of Oxford Nanopore library preparation kit bias on other sample types, such as soil, sediments, and other animal-derived samples, need further investigation.

Conclusion

This study identified a distinct recognition motif (5'-TATGA-3') of MuA utilized in the rapid sequencing kit used to prepare libraries for Oxford Nanopore sequencing. Notable interaction and coverage biases were observed in the rapid kit due to the strong cleavage preference. Underrepresentation of AT contents at the sequence terminus was seen in the ligation kit. However, because of the potential long reads, the reduced coverage of low GC content areas was minimized in the ligation kit. There was significant variation in observed microbial species abundances associated with different library preparation methods, including in the methane-producing genus, *Methanobrevibacter*. Such variation was also observed in the further analysis using mock communities. The ligation kit produced longer microbial reads, resulting in increased performance in taxonomic classification, but did not significantly affect methylation distributions. Therefore, our findings indicated that a careful and consistent library preparation method selection is essential for quantitative microbiome studies, especially the bovine-related microbiome, due to the systematic bias induced by the enzymatic reactions in Oxford Nanopore library preparation.

Methods

Animal ethics and sampling

Animals for Australian bovine ear tissue samples were Brahman cattle (*Bos taurus indicus*) bought from a commercial farm (Queensland, Australia) and housed at the University of Queensland Biological Resources Animal Facility. Informed consent for all procedures was obtained from the owners before sampling. Animal ethics was obtained from the University of Queensland Animal Ethics Committee under animal ethics numbers 2022/AE000438 and 2021/AE000541. Animals for Australian microbiome samples were Holstein dairy cattle owned by the University of Queensland and bred at the University of Queensland Gatton Dairy. Animal ethics was obtained from the University of Queensland Animal Ethics Committee under animal ethics number 2021/AE000991. The Spanish animals were privately owned Holstein dairy cattle located in the Basque Country and Catalonia regions. Informed consent for all procedures was obtained from the owners prior to sampling. Handling of all animals was carried out under EU Directive 2010/63/EU for the protection of animals used for scientific purposes, and experimental protocols for ruminal sampling were approved by the corresponding Ethical Committee (Approval number NEIKER-OEBA-2017–004).

For the mammalian samples ear notches from cattle were used as the DNA source. Briefly, ear tissue samples were collected from 30 female cattle using TSU Sampling Units (Allflex Livestock Intelligence, Australia) and TSU Applicator (Allflex Livestock Intelligence, Australia) under the manufacturer's instruction. Once the samples were collected, they were stored at 4 °C until DNA extraction.

The rumen fluid sample from Australia was collected from a 3-year-old cannulated cow. Briefly, the rumen contents were collected through the cannula. Then rumen contents were passed through a sieve to remove large solids and obtain the rumen fluid, which was transferred into 50 mL tubes. The rumen fluid was distributed into 1.5 mL tubes and stored at −20 °C until DNA extraction. Ruminal fluid from Spanish samples were collected using an oral tube (18 mm diameter and 160 mm long) connected to a 1,000-mL Erlenmeyer flask and continued to a mechanical pump (Vacubrand ME 2SI, Wertheim, Germany), with all the material contacting the cow being carefully cleaned between cows. Each animal was moved to an individual stall for this process. The solid fraction of the ruminal content was discarded by filtering through 4 layers of sterile cheesecloth, while the outcoming liquid fraction was stored in 50 ml tubes. The tubes were instantly frozen using liquid nitrogen and then stored at − 80 °C until DNA extraction.

DNA extraction and sequencing – mammalian DNA

Bovine DNA was extracted using the Puregene kit (QIAGEN, Germany), following the instructions of the manufacturer with slight modifications. In brief, 3 µl Proteinase K (QIAGEN, Germany) and 24 µl 1 M dithiothreitol (DTT) were added in the cell lysis step, and DNA was eluted in a 56 µL DNA hydration solution.

DNA concentration and purity were measured using the Qubit™ 4 Fluorometer (Thermo Scientific, USA) and NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, DE) respectively. The ligation sequencing kit SQK-NBD114.24 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK) were used to prepare the DNA libraries according to the manufacturer's protocols with slight modifications with the incubation time. For the rapid kit, the Barcode-DNA incubation time was increased to a total of 30 min, and the incubation time after the addition of AMPure XP Beads or Rapid Adapter F was doubled. For the ligation kit, the end-prep reaction time was increased to a total of 40 min. The Adapter-DNA incubation time was increased to 30 min, and all incubation times with AMPure XP Beads were increased to 10 min. The incubation time after the addition of the Elution Buffer was 10 min. DNA libraries were loaded into compatible flow cells (SQK-NBD114.24 for R10.4.1 and SQK-RBK110.96 for R9.4.1), followed by DNA sequencing on PromethION P24 (ONT, UK). Guppy v6.5.7 was used for basecalling under the super accurate (SUP) mode. The sequencing was terminated when data reached around 0.3 Gb per sample.

DNA extraction and sequencing – microbiome DNA

The rumen microbiome DNA from the Australian sample was extracted from a single rumen fluid sample using three different DNA extraction kits, namely DNeasy Plant Mini Kit (QIAGEN, Germany), Puregene Kit (QIAGEN, Germany) for Gram-positive bacteria, and QIAamp PowerFecal Pro DNA Kit (QIAGEN, Germany). Each extraction method was performed in three technical replicates. The QIAamp PowerFecal Pro DNA Kit (QIAGEN, Germany) and Puregene Kit (QIAGEN, Germany) were used following the manufacturer's protocol. The RNase A (QIAGEN, Germany) and Proteinase K (QIAGEN, Germany) were added to the cell lysis step with an extra 3 h of incubation at 55 °C for the DNeasy Plant Mini Kit (QIAGEN, Germany). DNA concentration and purity were measured with Qubit™ 4 Fluorometer (Thermo Scientific, USA) and NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, DE) respectively. The library of DNA extracted from QIAamp PowerFecal Pro DNA Kit (QIAGEN, Germany) was prepared by both the ligation kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK) according to the manufacturer's protocols with the same modifications with the incubation time as mentioned above. The library for DNA extracted from the DNeasy Plant Mini Kit (QIAGEN, Germany) was constructed using the ligation sequencing kit SQK-LSK109 (ONT, UK), while the library preparation of DNA extracted from the Puregene Kit (QIAGEN, Germany) was performed using SQK-LSK109 (ONT, UK) combined with the barcoding kit EXP-NBD104 (ONT, UK). All DNA libraries were loaded on FLO-PRO002 flow cells (R9.4.1), followed by sequencing on PromethION P24 (ONT, UK). Sequencing was terminated when each sample reached 300 Mb.

Table 3 Summary of mock communities

Name	GC%	DNA ratio (mock 1)	DNA ratio (mock 2)
Lactobacillus acidophilus. DSM 20079	34.5	1	1
Escherichia coli isolate from chicken faecal	50.0	1	2
Bifidobacterium pullorum subsp. gallinarum DSM 20670	64.0	1	3

The Spanish microbiome data was prepared and sequenced at the Animal Breeding Department, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, CSIC, Madrid, Spain. The Spanish rumen microbiome data was sourced from four Holstein cows. Genomic DNA was extracted from 250 μL of each thawed and homogenized ruminal content sample, using the "DNeasy Power Soil" commercial kit (Qiagen, Valencia, CA, USA) following the manufacturer protocol with the following modifications: After the addition of C1 buffer, the incubation was made during 10 min at 60ºC and 700 rpm. Then a vortex step at max speed was performed during 20 min in a Multivortex V-32 (BioSan™). All centrifugation times prior to adding the supernatant onto an MB Spin Columns were increased to 5 min, whereas centrifugation times after adding Solution C5 were increased to 2 min. One sequencing library was prepared using the ligation sequencing kit SQK-LSK109 (ONT, UK), and one sequencing library was prepared using the rapid sequencing kit SQK-RBK004 (ONT, UK) for each animal, for a total of eight sequencing libraries. The resulting libraries were run on a GridION (ONT, UK) using FLO-MIN106 (R9.4.1). Sequencing was terminated when each sample reached 260 Mb.

To compare the effect of the basecalling algorithms, Guppy v6.5.7 was used to basecall all microbiome data under fast (FAST), high accuracy (HAC), and super accurate (SUP) basecalling modes. For all analyses that were not directly testing the effect of the basecalling algorithm, or where the basecalling algorithm was not included in the model, the SUP basecalled data was used.

DNA extraction and sequencing – mock community DNA

Three bacterial species with GC contents ranging from 34 to 64% (Table 3) were used for mock community construction. Bacterial suspensions from the stationary phase were collected for DNA extraction. The Puregene Kit (QIAGEN, Germany) was used to extract DNA from *Escherichia coli* following the manufacturer's protocol. DNA extraction of the other two bacterial species was conducted using the QIAamp PowerFecal Pro DNA

Kit (QIAGEN, Germany) following the manufacturer's protocol.

Two mock communities (mock 1 and mock 2) were constructed using the extracted DNA from three bacterial species, with varying proportions of DNA in each community (Table 3). The DNA library of each mock community was prepared using both the ligation sequencing kit (SQK-NBD114.96) and rapid sequencing kit (SQK-RBK114.24) according to the manufacturer's protocols with the same modifications with the incubation time as mentioned above. Each sequencing kit included three technical replicates. DNA libraries were loaded into flow cells (R10.4.1), followed by sequencing on the PromethION 2 Solo (ONT, UK). Sequencing was terminated when each replicate reached 1.2 Gb (100x coverage). ONT Dorado v0.8.0 was used to rebasecalled all data under the SUP model.

Recognition site enrichment analysis

To assess the recognition site bias of the transposase enzyme in the rapid sequencing kit SQK-RBK110.96 (ONT, UK), the enzyme-DNA interaction site (read start site) was first identified by mapping the ear tissue reads to the *Bos taurus* reference genome ARS-UCDv1.2 [42]. The resulting enzyme-DNA interaction sites were then assessed for the deviation from the expected nucleotide proportions, given the background genome.

To standardize the amount of data per sample, bovine fastq data (ear tissue) from the two different library preparation kits (The ligation sequencing kit SQK-NBD114.24 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK)) were subsampled to 300 Mb using Rasusa v0.7.1 [43] before the mapping to the *Bos taurus* reference genome ARS-UCDv1.2 [42]. Average read lengths of subsampled reads were calculated using a bash script. Minimap2 v2.26 [44] was used to align subsampled fastq files to *Bos taurus* reference genome ARS-UCDv1.2 [42] with output set to sam format.

The genome sequence surrounding the read start sites was extracted. SAMtools v1.13 [45] and BEDTools v2.30.0 [46] were used to extract the positions in the reference genome that corresponded to the start of the mapped position of primary mapped reads in the sam files produced by Minimap2. Forward-mapped and reverse-mapped read start positions were split into two different bed files, with plus and minus signs, respectively. Windows of 31-bp and 1001-bp were created around the start position (end position for the reverse complement mapped reads to consider the read orientation) in the bed files through custom bash scripts. The extraction of reference genome sequences to fasta files was conducted by SeqKit v2.4.0 [47] using the edited bed files, while those for reverse complement reads were extracted with extra flags "--complement --reverse -v". These extracted sequences represented the nucleotide sequences that the library preparation kit enzymes interacted with for the generation of each sequencing read.

To examine the sequence enrichment in the reads where the library preparation enzymes interacted with the genome, Weblogo v3.7.9 [48] was used. To control for the background nucleotide frequency, a custom bash script was coded for the calculation of the *Bos taurus* reference genome ARS-UCDv1.2. This background frequency was included in the Weblogo analysis.

GC bias analysis

For the GC analysis, 40% and 60% were used as the cutoffs to separate different GC regions, where areas below 40% were considered low and those above 60% were regarded as high [49]. GC bias in the read start site (enzyme-DNA interaction site) was investigated using the mammalian data from ear tissue. For this analysis, the interaction site was considered an indicator of whether the nucleotide that the enzymes from the ligation sequencing kits interacted with, or the binding site of the transposase for the rapid sequencing kits. Custom Python scripts were used to calculate the percentage of mapped nucleotides of each position around the start site of the reads. Each chromosome in the reference genome was separated into 10 kb windows and the GC content of each segment was calculated using BEDTools v2.30.0 [46]. The GC content distribution of each chromosome was analyzed using a custom R script. The sequencing depth of each position in chromosomes was calculated by SAMtools v1.13 [45] with the depth function. To account for the read mapping orientation the start of the forward and the end of the reverse strands were considered read start sites. The interaction frequencies and sequencing depths of each 10 kb window were calculated through custom Python scripts.

For the enzyme-DNA interaction bias analysis, interaction frequencies over 100 were removed to reduce the noise, because these values are extremely high and these were likely due to errors and structural variations (such as repetitive sequences that are similar to the sequences in other regions of the genome) [50]. GC contents (from 0 to 100%) were divided at 20% (0–20%, 30–40%,..., 80–100%) or 0.5% (0-0.5%, 0.5-1%,..., 99.5–100%) intervals. The relative interaction frequency of a GC region was calculated by dividing the region's interaction frequency by the total interaction frequency of all regions. The nucleotide where the enzymes from the ligation sequencing kits interacted with, or the binding site of the transposase for the rapid sequencing kits was considered as an interaction site. Background probabilities of each GC window were used to normalize the interaction frequencies. The formula for calculating the normalized interaction frequency is below:

$$M_j = log_2 \frac{1 + \left(I_j / \sum_{i=0}^{100} I_i\right)}{1 + \left(N_j / \sum_{i=0}^{100} N_i\right)}$$

where M_j is the normalized interaction frequency at j% GC ratio frames; I is the frequency of interaction at regions with GC at i% and j%; N is the frequency of regions with GC contents at i% and j%. The normalization was performed to remove the effect of the reference genome background GC distribution on enzyme-DNA interaction bias analysis.

The sequencing coverage in each 10 kb window was calculated and normalized by the length of each window. GC contents (0 to 100%) were divided at 20% (0–20%, 30–40%,..., 80–100%) or 0.5% (0-0.5%, 0.5-1%,..., 99.5–100%) intervals. Data above 100 kb were removed to reduce the noise, as these values are extremely higher than the expected data size (1 kb) and were likely due to errors and structural variations [50]. Background probabilities of each GC window were used to normalize the coverage frequencies. The formula to normalize the coverage was similar to the one calculating interaction frequency:

$$N_j = log_2 \frac{1 + \left(C_j / \sum_{i=0}^{100} C_i\right)}{1 + \left(N_j / \sum_{i=0}^{100} N_i\right)}$$

where N_j is the normalized coverage frequency at j% GC ratio frames; C is the frequency of coverage at GC contents at i% and j%; N is the frequency of regions with GC ratios at i% and j%. The normalization was performed to exclude the effect of background GC distribution of the reference genome on sequencing coverage analysis. The Pearson correlation coefficient was used to identify the correlation between normalized coverage and interaction frequency. An unpaired t-test was used to compare the normalized interaction frequency and coverage in low (below 40%) and high (above 60%) GC content regions between two library preparation kits (The ligation sequencing kit SQK-NBD114.24 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK)). The *P*-value threshold was 0.05.

Microbiome data analysis

Two independent datasets were used to assess the effects of the library preparation method, basecalling accuracy, and DNA extraction methods on the microbial distribution observed in microbiome analysis. In the Australian dataset, the same rumen fluid sample was extracted by three different DNA extraction kits, with three technical replicates per extraction kit. DNA extracted by the QIAamp PowerFecalPro (QIAGEN, Germany) kit was sequenced by both the ligation kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK), with three technical replicates per sequencing kit. In the Spanish dataset, the rumen fluid samples were collected from four cows (four biological replicates). The rumen fluid DNA from each cow was extracted using the DNeasy PowerSoil (QIAGEN, Germany) extraction kit, followed by the DNA library preparation using both the ligation sequencing kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK004 (ONT, UK) for each DNA sample.

After sequencing, Fast5 files were basecalled using Guppy v6.5.7 under fast (FAST), high-accuracy (HAC), and super accurate (SUP) modes. Adapters were trimmed by using Porechop_ABI v0.5.0 [51], followed by the application of Nanofilt v2.8.0 [52] for extracting reads above 100 bp and the corresponding Q scores. The minimum Q scores for FAST, HAC, and SUP were 8, 9, and 10 respectively. The N50 of raw and trimmed reads were calculated through SeqKit v2.4.0 [47]. Processed fastq files were taxonomically classified through Kraken2 v2.1.2 [53] under a customized Refseq database containing bacteria, fungi, archaea, and protozoa complete genomes from the NCBI.

Analysis of relative abundances, alpha diversity, and beta diversity was performed by vegan v2.6-4 [54] and phyloseq v1.40.0 [55] using all microbiome datasets to compare the microbial profile variations of extraction, sequencing, and basecalling protocols. For assembly statistical analysis, Australian and Spanish microbiome datasets sequenced by two library preparation kits and basecalled under SUP mode were used (Australia: the ligation sequencing kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK); Spain: the ligation sequencing kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK004 (ONT, UK)). To allow comparable analysis reads were subsampled to 0.19 Gb using Rasusa v0.7.1 [43], followed by assembly using Flye v2.9 with the '--meta' flag [56]. Quast v5.2.0 [57] was used to evaluate the assembly results. The N50 of subsampled reads was calculated using SeqKit v2.4.0 [47]. A t-test was used to compare the classified read percentage, read length N50, relative abundances, Shannon indexes, contig N50, contig numbers, and methylated site numbers among different DNA extraction, sequencing, and basecalling methods (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). The Holm-Bonferroni method was used to adjust the *P*-value for extraction and basecalling method comparison. Fitting Linear Models were used to evaluate the effects of basecalling models, sequencing kits, individual animals, technical replicates, and extraction kits on classification efficiency with the formulas below:

$$\boldsymbol{Australia} : classified \sim basecall\_mode + sequencing\_kit + extraction\_kit + technical\_replicate$$

$$\boldsymbol{Spain} : classified \sim basecall\_mode + sequencing\_kit + animal\_id$$

where *classified* is the percentage of classified reads, the *basecall_mode* is the model used for basecalling (set as numeric: 1 for FAST, 2 for HAC, and 3 for SUP), the *sequencing_kit* is the sequencing protocol used for library preparation (set as factors), the *animal_id* is the unique identity for each animal (set as factors), the *technical_replicate* is the unique identity for each technical replicate (set as factors), and the *extraction_kit* is the protocol for DNA extraction (set as factors). Linear Models were fit using the *lm* function in RStudio. Analysis of variance (ANOVA) was used to compare different Fitting Linear Models. The Pearson correlation coefficient was used to identify the correlations between microbiome read length N50 and classified read proportions, microbiome read length N50 and contig number, and microbiome read length N50 and contig N50. The *P*-value threshold was 0.05.

Bacterial DNA methylation analysis

Australian and Spanish rumen microbiome datasets sequenced by two library preparation kits were used in this analysis (Australia: the ligation kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK110.96 (ONT, UK); Spain: the ligation sequencing kit SQK-LSK109 (ONT, UK) and the rapid sequencing kit SQK-RBK004 (ONT, UK)). Only reads basecalled with SUP mode were incorporated. To allow comparable analysis, each of the datasets was subsampled to 0.19 Gb using Rasusa v0.7.1 [43], after trimming and filtering. The positions of N6-Methyladenosine (6 mA) of the genomes were characterized using mCaller [58] and stored in a tsv file. The mapping proportions and sequencing coverage were calculated using SAMtools v1.13 [45]. Three bacterial reference genomes were used in the analysis with the corresponding Refseqs downloaded from NCBI (Table 4). A t-test was also used to compare the methylated site numbers between the two sequencing methods (An unpaired t-test for the Australian dataset and a paired t-test for the Spanish dataset). A Fisher's Exact Test was performed to evaluate whether the observed overlap methylated position number between the two sequencing kits was higher than expected, with the alternative hypothesis as "greater". The total number of adenosine and thymine in the reference genome was selected as the background number of methylated 6mA positions for the Fisher's Exact Test. The *P*-value threshold was 0.05.

Mock community data analysis

Mock community data were rebasecalled under the SUP model using ONT Dorado v0.8.0 with the adapter trimming function enabled. Reads shorter than 100 bp were removed using Nanofilt v2.8.0 [50]. The alignment to three reference genomes, *L. acidophilus* (GCF_003047065.1), *E. coli* (GCF_000008865.2), and *B. gallinarum* (GCF_004135085.1), was conducted using Minimap2 v2.26 [42], followed by the extraction primary mapped read into new fastq files by SAMtools v1.13 [45]. The primary mapped base number and N50 were calculated through a customized bash script.

To exclude the potential effects of DNA fragmentation during the DNA extraction, the relative abundance of each bacterial species was calculated through the division of the primary mapped base number of each species by the total mapped base number. The calculation of differential relative abundance was conducted by the DNA proportion of the mock community subtracted from the corresponding relative abundance. An unpaired t-test was utilized to compare the differential relative abundance between the two sequencing protocols. The *P*-value threshold was 0.05.

Table 4 Bacterial reference genomes for methylation analysis

Name	GC%	Genome size	Refseq assembly number
Prevotella bryantii strain TS1-5	38.0	3.4 Mb	GCF_022024195.1
Prevotella ruminicola 23	47.5	3.4 Mb	GCF_018389405.1
Fibrobacter succinogenes subsp. succinogenes S85	48.0	3.8 Mb	GCF_000146505.1

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11649-z.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
We would like to thank Mr. Mariano Parra from the University of Queensland for helping to collect the rumen fluid samples. We thank Mr. David Chau for providing bacterial species for our mock community construction. We also would like to thank Dr. James P. Copley for providing spirit leading during this work. We would like to thank the members of the Center of Animal Science (Queensland Alliance for Agriculture and Food Innovation, University of Queensland) for their comments and feedback as this work was developed.

Author contributions

Conception and design were by Z.C, C.T.O, O.G.R, and E.M.R. Sample collection and experiments were by Z.C, C.T.O, O.G.R, M.G.R., H.J.L, L.T.N, S.J.M and E.M.R. Data analysis and visualization were by Z.C, C.T.O, and E.M.R. Draft writing was by Z.C. Draft reviewing was by Z.C, C.T.O, O.G.R, M.G.R, H.J.L, L.T.N, S.J.M and E.M.R.

Funding
This study was supported by the LESTR Low Emission Saliva Test for Ruminants project, funded under the Meat & Livestock Australia, grant number P.PSH.2010.

Data availability

The raw sequence data of this study were archived under the NCBI SRA BioProject, accession number: PRJNA1117518 (https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA1117518). All scripts are available in the GitHub repository: https://github.com/SimonChen1997/ONT-sequencing-bias-analysis.

Declarations

Ethics approval and consent to participate

Animal use in this study was approved by the University of Queensland Animal Ethics Committee under animal ethics numbers 2021/AE000991, 2022/AE000438 and 2021/AE000541 for samples sourced from Australia (2021/AE000991 for the Australian microbiome samples; 2022/AE000438 and 2021/AE000541 for the Australian bovine ear tissue samples), and by the Basque Institute for Agricultural Research and Development Ethics Committee (Neiker-OEBA-2017–004) on 28 March 2017, in accordance with Spanish Royal Decree 53/2013 for the protection of animals used for experimental and other scientific purposes. Informed consents for all procedures in this study were obtained from the owners of the animals before sampling.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 August 2024 / Accepted: 28 April 2025
Published online: 19 May 2025

References

1. Hu TS, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. Hum Immunol. 2021;82:801–11.
2. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, Peto T, Crook D, Stoesser N. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genomics. 2023;9:000910.
3. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genom. 2022;2:100129.
4. Stoddart DJ, White J. Polynucleotide modification methods. Oxford Nanopore Technologies PLC. 2021. https://patents.google.com/patent/US11186857B2/en. Accessed 7 May 2024.
5. Lubiene J, Berezniakovas A, Lubys A. Enzyme composition for dna end repair, adenylation, phosphorylation. Thermo Fisher Scientific. 2014. https://patents.google.com/patent/US20150087557A1/en. Accessed 7 May 2024.
6. Xu M-Q, Fang Y, Zhang A, Sun L. Application of immobilized enzymes for Nanopore library construction. New England Biolabs Inc. 2020. https://patents.google.com/patent/US20220090056A1/en. Accessed 7 May 2024.
7. Gormley NA, Smith GP, Bentley D, Rigatti R, Luo S. Method of preparing libraries of template polynucleotides. Illumina Cambridge Limited. 2016. https://patents.google.com/patent/US7741463B2/en. Accessed 7 May 2024.
8. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. Genome Biol. 2011;12:R18.
9. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, Hisatsune J, Sugai M, Takehiko I, Hayashi T. Comparison of the sequencing bias of currently available library preparation kits for illumina sequencing of bacterial genomes and metagenomes. DNA Res. 2019;26:391–8.
10. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, Graham MR, Sharma MK. Comparison of sample preparation methods used for the Next-Generation sequencing of Mycobacterium tuberculosis. PLoS ONE. 2016;11:e0148676.
11. Coyote-Maestas W, Nedrud D, Okorafor S, He YG, Schmidt D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. Nucleic Acids Res. 2020;48:1010–1010.
12. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the illumina sequencing system. Nat Methods. 2008;5:1005–10.
13. Zhang AH, Li SH, Apone L, Sun XL, Chen LX, Ettwiller LM, Langhorst BW, Noren CJ, Xu MQ. Solid-phase enzyme catalysis of DNA end repair and 3' A-tailing reduces GC-bias in next-generation sequencing of human genomic DNA. Sci Rep. 2018;8:15887.
14. Lan JH, Yin YX, Reed EF, Moua K, Thomas K, Zhang QH. Impact of three illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol. 2015;76:166–75.
15. Wolpe JB, Martins AL, Guertin MJ. Correction of transposase sequence bias in ATAC-seq data with rule ensemble modeling. Nar Genom Bioinform. 2023;5:lqad054.
16. Soares LMM, Hanscom T, Selby DE, Adjei S, Wang W, Przybylski D, Thompson JF. DNA read count calibration for single-molecule, long-read sequencing. Sci Rep. 2022;12:17257.
17. Ross EM, Moate PJ, Bath CR, Davidson SE, Sawbridge TI, Guthridge KM, Cocks BG, Hayes BJ. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. Bmc Genet. 2012;13:53.
18. Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajilic-Stojanovic M, Kekkonen RA, Palva A, de Vos WM. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell Lysis. J Microbiol Meth. 2010;81:127–34.
19. Tourlousse DM, Narita K, Miura T, Sakamoto M, Ohashi A, Shiina K, Matsuda M, Miura D, Shimamura M, Ohyama Y, et al. Validation and standardization of DNA extraction and library construction methods for metagenomics-based human fecal Microbiome measurements. Microbiome. 2021;9:95.
20. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, MacInnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics. 2012;13:1.
21. Hickman AB, Dyda F. Mechanisms of DNA transposition. Microbiol Spectr. 2015;3:MDNA3–0034.
22. De G, Gross SM, Li J-S, Morrell N, Slatter A, Shen K, Snow S. Tagmentation using immobilized transposomes with linkers. Illumina Cambridge Limited. 2018. https://patents.google.com/patent/US20180245069A1/en. Accessed 7 May 2024.
23. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn5/IS50 target recognition. P Natl Acad Sci USA. 1998;95:10716–21.
24. Gunasekera S, Abraham S, Stegger M, Pang S, Wang PH, Sahibzada S, O'Dea M. Evaluating coverage bias in next-generation sequencing of Escherichia coli. PLoS ONE. 2021;16:e0253440.
25. Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE. Evaluation of a transposase protocol for rapid generation of shotgun High-Throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microb. 2011;77:8071–9.
26. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang XQ, Shendure J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010;11:R119.
27. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA. 2012;3:3.
28. Haapa-Paananen S, Rita H, Savilahti H. DNA transposition of bacteriophage Mu - A quantitative analysis of target site selection in vitro. J Biol Chem. 2002;277:2843–51.
29. Radukic MT, Brandt D, Haak M, Muller KM, Kalinowski J. Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. Nar Genom Bioinform. 2021;3:lqab029.
30. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A, Hansen LH. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Gigascience. 2020;9:giaa008.
31. Alzbutas G, Askolin S, Gagilas J, Gliebutė S, Haakana H, Juhila J, Kavanagh I, Lubys MKL-L, Morkūnas A et al. J,: MuA transposase enzyme enables fast and easy DNA library preparation for next generation sequencing. 2013. https://www.gene-quantification.de/qpcr-ngs-2013/posters/P013-qPCR-NGS-2013.pdf. Accessed 14 April 2024.
32. Moss AR, Jouany JP, Newbold J. Methane production by ruminants: its contribution to global warming. Ann Zootech. 2000;49:231–53.
33. Shagin DA, Lukyanov KA, Vagner LL, Matz MV. Regulation of average length of complex PCR product. Nucleic Acids Res. 1999;27:e23–i–e23-iii.
34. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques. 2012;52:87–94.
35. Yu ZT, Morrison M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. Biotechniques. 2004;36:808–12.
36. Govender KN, Eyre DW. Benchmarking taxonomic classifiers with illumina and nanopore sequence data for clinical metagenomic diagnostic applications. Microb Genomics. 2022;8:000886.
37. Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, Pignatelli M, Moya A. Comparison of different assembly and annotation tools on analysis of

simulated viral metagenomic communities in the gut. BMC Genomics. 2014;15:37.

38. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, Reik W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018;19:33.
39. Morrison J, Koeman JM, Johnson BK, Foy KK, Beddows I, Zhou WD, Chesla DW, Rossell LL, Siegwald EJ, Adams M, Shen H. Evaluation of whole-genome DNA methylation sequencing library preparation protocols. Epigenet Chromatin. 2021;14:28.
40. Zhou L, Ng HK, Drautz-Moses D, Schuster SC, Beck S, Kim C, Chambers JC, Loh M. Systematic evaluation of library Preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Rep. 2019;9:10383.
41. Di Lena P, Sala C, Nardini C. Estimage: a webserver hub for the computation of methylation age. Nucleic Acids Res. 2021;49:W199–206.
42. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. De Novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9:giaa021.
43. Hall MB. Rasusa: randomly subsample sequencing reads to a specified coverage. J Open Source Softw. 2022;7:3941.
44. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
45. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of samtools and BCFtools. Gigascience. 2021;10:giab008.
46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
47. Shen W, Le S, Li Y, Hu FQ. SeqKit: A Cross-Platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
48. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004;14:1188–90.
49. Techa-Angkoon P, Childs KL, Sun YN. GPRED-GC: a gene prediction model accounting for 5’-3’ GC gradient. BMC Bioinformatics. 2019;20:482.
50. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. PLoS ONE. 2021;16:e0257521.
51. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford nanopore technology sequencing reads for downstream trimming. Bioinform Adv. 2022;3:vbac085.
52. De Coster W, D’Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
53. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20:257.
54. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O’Hara B, Simpson G, Solymos P, Stevens H, Wagner H. Vegan: Community ecology package. 2015. https://github.com/vegandevs/vegan. Accessed 7 May 2024.
55. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of Microbiome census data. PLoS ONE. 2013;8:e61217.
56. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL. Pevzner PA: MetaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020;17:1103–10.
57. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
58. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE. Single-molecule sequencing detection of 6-methyladenine in microbial reference materials. Nat Commun. 2019;10:579.

Publisher’s note