RESOURCE ARTICLE OPEN ACCESS

Unlocking River Biofilm Microbial Diversity: A Comparative Analysis of Sequencing Technologies

Meri A. J. Anderson[1,2] | Amy C. Thorpe[1] | Susheel Bhanu Busi[1] | Hyun Soon Gweon[2] | Jonathan Warren[3] | Kerry Walsh[3] | Daniel S. Read[1]

[1]UK Centre for Ecology & Hydrology (UKCEH), Benson Lane, Crowmarsh Gifford, Wallingford, UK | [2]School of Biological Sciences, University of Reading, Reading, UK | [3]Environment Agency, Bristol, UK

Correspondence: Meri A. J. Anderson (merand@ceh.ac.uk) | Daniel S. Read (daniel.read@ceh.ac.uk)

Received: 12 June 2025 | **Revised:** 13 October 2025 | **Accepted:** 23 October 2025

Handling Editor: Holly Bik

Funding: M.A.J.A. was supported by NERC SCENARIO PhD studentship [NE/S007261/1]. This work was supported by the Environment Agency under research project SC220034. S.B.B. was supported by the BBSRC Institute Strategic Programme: Decoding Biodiversity (DECODE) grant. A.C.T., J.W., K.W. and D.S.R. were supported by the Natural Environment Research Council (NERC) [grant numbers NE/X015947/1 and NE/X015777/1]. The views expressed in this paper are the views of the authors and do not represent the position of the Environment Agency or other employer organisations.

Keywords: bacterial community composition | epilithic biofilms | long-read sequencing | Short-read sequencing

ABSTRACT

Freshwater ecosystems are under increasing pressure from pollution, habitat degradation and climate change, highlighting the need for reliable biomonitoring approaches to assess ecosystem health and identify the causes of biodiversity and ecosystem service loss. Characterisation of freshwater microbiomes has the potential to be an important tool for understanding freshwater ecology, ecosystem health and ecosystem function. High-throughput sequencing technologies, such as Illumina short-read and Pacific Biosciences long-read sequencing, are widely used for microbial community analysis. However, the relative performance of these approaches for monitoring freshwater microbiomes has not been well explored. In this study, we compared the performance of long- and short-read sequencing approaches to assess archaeal and bacterial diversity in 42 river biofilm samples across seven distinct river sites in England by targeting the 16S ribosomal RNA gene. Our findings demonstrated that longer reads generated by PacBio sequencing provide a higher taxonomic resolution, enabling the classification of taxa that remained unassigned in the short-read Illumina datasets. This enhanced resolution is particularly beneficial for biodiversity assessments because it improves species-level identification, which is crucial for ecological monitoring. Despite this, both sequencing methods produced comparable bacterial community structures regarding taxon relative abundance, suggesting that the sequencing approach does not profoundly affect the comparative assessment of community composition. However, while Illumina offers higher throughput and cost efficiency, PacBio's ability to resolve complex microbial communities highlights its potential for studies requiring precise taxonomic identification.

1 | Introduction

DNA sequencing has transformed how we study the living world, opening new opportunities for understanding biodiversity and ecosystem function (Shendure et al. 2017; Goodwin et al. 2016). This technology has become an increasingly important tool in environmental research, particularly for studying complex multi-kingdom microbial communities (Thompson et al. 2017). By revealing the breadth of microbial diversity in environmental samples, DNA sequencing can

https://doi.org/10.1111/1755-0998.70075

help researchers and environmental regulators make more informed decisions regarding ecosystem management and conservation (Taberlet et al. 2012; Porter and Hajibabaei 2018). Environmental DNA (eDNA) monitoring has emerged as a transformative method and an essential tool for biomonitoring, with diverse applications, including pathogen detection (Farrell et al. 2021), tracking invasive species (Thomas et al. 2020), monitoring endangered or cryptic species (Ota et al. 2020), assessing biodiversity (Keck et al. 2022) and identifying habitat connectivity (Littlefair et al. 2023). The effectiveness of these approaches relies on several factors, including the ability to classify and identify DNA markers, including specific taxa of interest, such as pathogens and rare or invasive species, at the highest taxonomic resolution. Accurate and reliable sequencing technologies are pivotal for environmental monitoring because they provide a molecular lens through which researchers can detect and quantify biodiversity, track ecosystem changes and health, monitor invasive species and their potential impacts, evaluate conservation efforts and unravel complex ecological interactions with unprecedented precision and sensitivity.

Short-read sequencing platforms, such as Illumina, have become widespread owing to their availability, cost-effectiveness and high-throughput capabilities (Bentley et al. 2008; Satam et al. 2023). The <600 bp reads (up to 2×300 bp) generated by Illumina technology are particularly effective for analysing hypervariable regions of the 16S rRNA gene (Yang et al. 2016). However, analyses using these shorter reads can struggle to resolve complex genomic regions and repetitive sequences (van Dijk et al. 2018). In contrast, long-read sequencing platforms, such as Pacific Biosciences (PacBio), can generate reads averaging 10–25 Kb (Hon et al. 2020), offering improved resolution of structural variants and complex genomic regions (Rhoads and Au 2015; Logsdon et al. 2020). Although these longer reads can span multiple repeat and hypervariable regions simultaneously and potentially provide more accurate taxonomic classifications (Callahan et al. 2019), they are typically more expensive and have lower throughput (Amarasinghe et al. 2020).

The contrasting properties of short- and long-read sequencing technologies can substantially influence ecological inferences derived from molecular data. Illumina short-read platforms provide high-throughput, accurate sequences that are well suited for detecting dominant taxa but may under-represent rare or low-abundance species due to limited read length and amplification bias (Wang et al. 2022). In contrast, PacBio long-read sequencing generates extended fragments that capture more genetic information, revealing cryptic diversity and providing higher taxonomic resolution (van Dijk et al. 2018). Consequently, choosing a sequencing platform has become a critical methodological consideration that can fundamentally alter the ecological interpretation of molecular biodiversity data. Previous comparative studies using Illumina and PacBio sequencing technologies have revealed significant variations in methodology and performance across different research contexts. However, most comparative studies have been conducted on model organisms or within well-characterised ecosystems, limiting their applicability to diverse ecological contexts (Ferrarini et al. 2013; Zhang et al. 2020; Galata et al. 2021; Barthélémy et al. 2024). Although Gao et al. (2024) described the utility of long-read data for characterising deep-sea surface sediments, existing comparative analyses often fail to comprehensively address how different sequencing technologies might differentially represent complex ecological interactions and biodiversity gradients. These limitations create a significant research opportunity to develop a more nuanced understanding of how sequencing technology performance varies across different ecological contexts, particularly in understudied aquatic ecosystems.

Our study compared short-read (Illumina, ca. 235 bp) and long-read (PacBio, ca. 1600 bp) sequencing to analyse epilithic river biofilm bacterial communities using 16S rRNA gene sequencing. We hypothesised that long-read sequencing would provide greater taxonomic resolution and a more comprehensive understanding of biodiversity than short-read sequencing, particularly for distinguishing closely related bacterial species (Johnson et al. 2019; Tedersoo et al. 2018). To test this, we sequenced DNA from 42 biofilm samples using short- and long-read sequencing methods to compare community characteristics, including the overlap and uniqueness of bacterial taxa detected using each approach. By addressing the current knowledge gaps in the performance and application of different sequencing approaches, this study enhances our understanding of the optimal use of sequencing technologies for environmental monitoring (Ruppert et al. 2019).

2 | Methods

2.1 | Sample Collection

Epilithic biofilm samples were collected from rivers across England as part of the Environment Agency's River Surveillance Network (RSN) monitoring program, following the standard sampling method described by Kelly et al. (2020). The total sampling campaign encompassed 2101 river biofilm samples collected from 861 sites between 2021 and 2023; a detailed description of the methods for sample collection and short-read sequencing is available in the Environment Agency report (Environment Agency 2024). This study focused on a subset of 42 samples collected from seven sites twice a year in 2021, 2022 and 2023 (Figure 2). Land cover maps from UKCEH (Morton et al. 2021) were used to describe dominant land cover types for the upstream catchment of each site (Table S3). Samples were obtained by scraping five stones or macrophytes into a tray containing 50 mL of river water. The upper surfaces were brushed with a clean toothbrush to remove biofilms. 5 mL of the biofilm suspension was removed using a pipette and preserved in 5 mL of DNA preservation buffer (3.5 M ammonium sulphate, 17 mM sodium citrate and 13 mM EDTA). Following collection, samples were concentrated by centrifugation at 3000× ***g*** for 15 ± 2 min at 5°C ± 2°C, frozen and transported on dry ice to the UK Centre for Ecology & Hydrology (UKCEH), Wallingford, for subsequent analysis. Over a three-month period prior to biofilm sample collection, up to five water chemistry samples were taken; this data was used to calculate the water chemistry means. Water quality measurements included nitrate, phosphorus and dissolved oxygen levels, along with water temperature and pH. Full water chemistry data are available in Table S3, Figure S1.

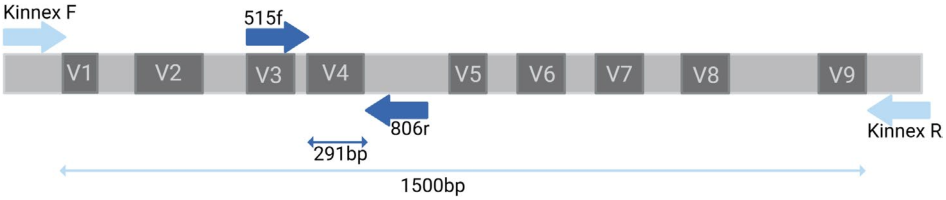

FIGURE 1 | Primer positions on the 16S rRNA gene, showing the overlap of the Illumina 16SV4 primers within the 16S gene sequenced by the Kinnex primers.

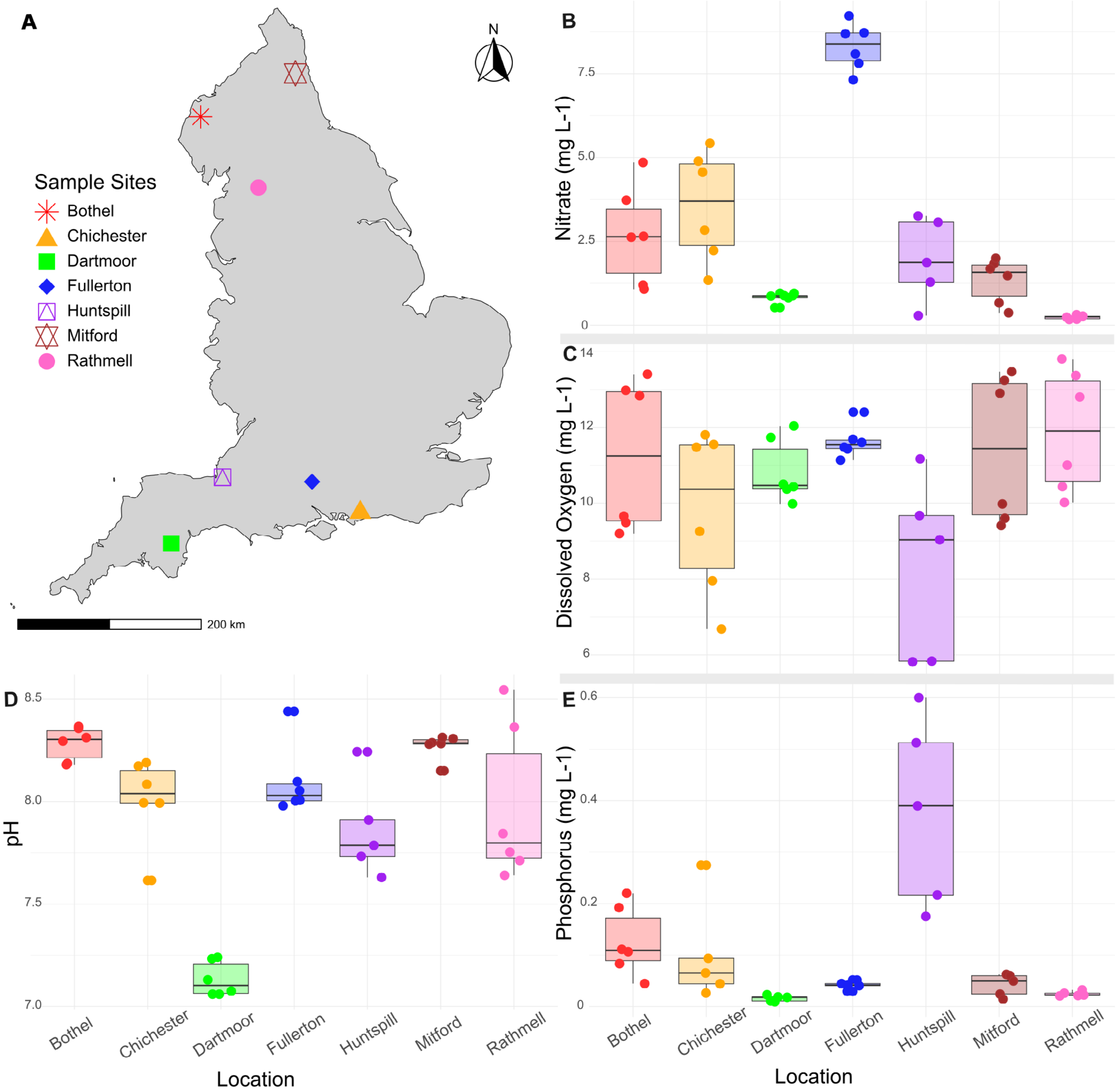

FIGURE 2 | (A) Map of the seven sites across England from which the biofilm samples were collected. Water chemistry data from the seven sites. Nitrate (mg L^{-1}) (B), dissolved oxygen (mg L^{-1}) (C), pH (D) and phosphorus (mg L^{-1}) (E) levels at each sample point for each location. Water chemistry means were calculated using five chemistry samples collected over a three-month period prior to biofilm sample collection.

2.2 | DNA Extraction

DNA was extracted from 100 μL of biofilm suspension using the Quick-DNA Faecal/Soil Microbe Kit (Zymo Research, California, United States), with modifications to optimise the DNA yield (Newbold et al. 2025). The extraction protocol was adapted as follows: 500 μL of DNA/RNA Shield (Zymo Research) was added to each sample as lysis buffer. The samples were mechanically disrupted using a TissueLyser II (Qiagen, Germany) at 20 Hz for 20 min. 20 μL of recombinant Proteinase K (Roche, Switzerland) was added to the lysate and incubated at 65°C for 20 min. The purified DNA was eluted in 100 μL of elution buffer. A negative extraction control, without sample material, was used to monitor potential contamination. DNA concentration was quantified using the Qubit dsDNA High-Sensitivity kit (Life Technologies Limited) according to the manufacturer's protocol. The extracted DNA was stored at 4°C until PCR amplification. A detailed step-by-step extraction, PCR and library preparation protocol is available at https://doi.org/10.17504/protocols.io.j8nlk8em6l5r/v1.

2.3 | Sequencing

Two methods of sequencing were used to amplify the 16S rRNA gene (Figure 1). Illumina NextSeq for short-read sequencing (291 bp) and Pacific Biosciences Sequel II for long-read sequencing (1500 bp).

2.3.1 | Pacific Biosciences

The V1-V9 region of the 16S rRNA gene was amplified by Novogene using the Kinnex protocol (Srinivas et al. 2025) (primer sequences in Table S1, PCR amplification in Table S2). The PCR products of the barcoded V1–V9 amplicons were detected by agarose gel electrophoresis prior to processing on a PacBio Sequel II sequencing platform (Novogene, Cambridge, UK).

2.3.2 | Illumina

The V4 region of the 16S rRNA gene was amplified using specific primers (Table S1) modified to include Illumina adaptor sequences. In a UV-sterilised laminar flow hood, a master mix was prepared containing 0.5 μL of 2000 units mL^{-1} Q5 high-fidelity DNA polymerase, 10 μL of 5× reaction buffer, 10 μL of 5× high GC enhancer (New England Biolabs, UK), 1 μL of a 10 mM dNTP mix (Bioline, UK), 0.1 μL of each 100 μM forward and reverse primer pair (Table S1) and 26.3 μL of molecular grade water. The master mix (48 μL) was dispensed into each well of a 96-well plate, and 2 μL of template DNA was added per sample. Negative PCR controls were also included. The thermocycling conditions are presented in Table S2. Successful amplification was verified by 1.5% agarose gel electrophoresis using GelRed nucleic acid staining. PCR products were purified using a MultiScreen PCR filter plate, resulting in 35 μL of eluted product.

The second PCR step employed a dual-indexing approach to enable sample demultiplexing. Indexing primers were prepared using an Opentrons liquid-handling robot, each consisting of a forward (i5) or reverse (i7) Illumina adaptor sequence, an i5 or i7 Nextera index and an Illumina pre-adaptor sequence. The second PCR mix contained 0.25 μL of Q5 DNA polymerase, 5 μL of reaction buffer, 5 μL of high GC enhancer, 0.5 μL of dNTPs, 5 μL of the indexing primers (pre-prepared in the plate), 7.25 μL of molecular grade water and 2 μL of purified PCR product from the first PCR step. The cycling protocol is presented in Table S2. Amplification was confirmed using agarose gel electrophoresis.

The second-step PCR product was normalised using the NGS Normalisation kit (Norgen Biotek, Canada) to achieve a concentration of approximately 5 ng μL^{-1}. The samples were pooled by plate and quantified using a Qubit High-Sensitivity Assay Kit. The amplicon library was prepared by diluting and pooling the samples, followed by concentration and purification using gel extraction. The final libraries were quantified, diluted to 1000 pM and sent to Illumina Cambridge for sequencing on a NextSeq 2000 with a P1 flow cell and 40% PhiX control.

3 | Data Analysis

Amplicon sequence reads were processed using the DADA2 pipeline (Callahan et al. 2019) implemented in R [version 4.4.2] (R Core team 2024). Short-read and long-read sequences underwent distinct processing workflows, with full analysis scripts for short-read sequences available at https://github.com/amycthorpe/amplicon_seq_processing_biofilms. For short-read analyses, raw sequences were demultiplexed, and adaptor sequences were trimmed using the Illumina FASTQ generation pipeline. Primers were removed using the 'trimLeft' parameter, and the quality profiles of the forward and reverse reads were examined. Reads were truncated when quality scores fell below Q30 and filtered using stringent criteria, including removing reads with ambiguous bases and a maximum expected error threshold of 2. The DADA2 algorithm learned error rates from a 100 million base subset, with visualisations confirming the alignment of the estimated rates with the observed data. Reads were then dereplicated into unique sequences based on the error rate model, and the core sample inference algorithm was used to identify true sequence variants. Paired forward and reverse reads were aligned and merged, requiring a minimum of 12-base overlap. Chimeric sequences were identified and removed, resulting in an amplicon sequence variant (ASV) abundance table. Long-read analyses followed a modified protocol (https://benjjneb.github.io/LRASManuscript/LRASms_fecal.html), with primers removed and reads trimmed to a minimum length of 1200 bp and a quality threshold of three (filterAndTrim(nops2, filts2, minQ = 3, minLen = 1200, maxLen = 1600, maxN = 0, rm.phix = FALSE, maxEE = 2)). Subsequent demultiplexing generated a sequence table with sample-specific counts. Taxonomy was assigned to each ASV using the naive Bayesian classifier (Wang et al. 2022) with a minimum bootstrap confidence of 60 against the SILVA v138.1 reference database (Quast et al. 2012) for Illumina and PacBio 16S rRNA gene sequences.

The sequences were rarefied to a uniform sequencing depth by examining rarefaction curves and identifying the sequencing depth at which the richness plateaued (Figure S6). Both long- and short-read data were rarefied to 3000 reads per sample to conserve the majority of samples. All negative extraction and

PCR controls and a small number of samples (five) did not meet the rarefaction depth and were therefore removed. Sequences assigned as mitochondria and chloroplasts were removed from the datasets. These accounted for 31% and 32% of the total reads in the short- and long-read data, respectively.

Downstream analyses were performed using R [version 4.4.2]. Counts, taxonomy and metadata files were loaded into Microeco (Liu et al. 2021) for processing and visualisation. To assess differences in the number of ASVs between sequencing technologies, we used a Wilcoxon signed-rank test followed by a linear mixed-effects model to evaluate the direction of the effect. Taxonomic assignment proportions across ranks were compared using the Mann–Whitney U test. Differences in the relative abundance of taxa between sequencing platforms were tested using a Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons.

To examine community composition similarities between sample types and sequencing platforms, Principal Coordinates Analysis (PCoA) based on Bray–Curtis dissimilarity at the genus level was performed, along with a Procrustes analysis to assess concordance between ordinations. Prior to analysis, ASV identifiers were replaced with their corresponding taxonomic names from the kingdom to genus level to ensure meaningful comparison of community structure between sequencing methods. Additionally, Deming regression was used to investigate the agreement between the number of reads assigned to each phylum across sequencing types. All statistical analyses were performed in R, with significance thresholds set at $p < 0.05$.

To compare short- and long-read ASVs in the sequence space and identify overlapping short- and long-read ASVs, we used NCBI BLAST (Camacho et al. 2009), with the following parameters: ‘–id 80 –query-cover 90 –subject-cover 90 –more-sensitive –outfmt 100’. The output was filtered to retain sequences with at least 90% identity and a minimum length of 235 bp for subsequent analysis. This length threshold was chosen as it represents 10% below the median short-read length. Short-read ASVs that did not map to long-read ASVs were considered unique sequences.

To assess potential primer bias, the long-read sequences were trimmed using the specific primer sequences employed in the short-read protocol to isolate the corresponding amplicon region. These trimmed long-read sequences were then processed using the same DADA2 pipeline as the original long-read data. All subsequent analyses were conducted in R using the same workflow to ensure consistency and comparability.

The relevant code for the analysis and figures, and raw data can be found at: https://doi.org/10.5281/zenodo.17432155 and https://www.ncbi.nlm.nih.gov/bioproject/1353123.

4 | Results

4.1 | River Surveillance Network Biofilms

At each sampling point (Figure 2a), epilithic biofilms were collected to assess bacterial biodiversity, complemented by the concurrent collection of relevant water chemistry and nutrient data (Figure 2b–e) to understand how biotic and abiotic factors shape the community composition of bacteria within the RSN. Water chemistry measurements, including pH, conductivity and nutrient concentrations (e.g., nitrate, phosphate), showed variation across sites, with each parameter displaying a broad distribution (Table S3). However, potential physicochemical drivers of sequencing technology differences were not further explored due to the limited number of sites ($n = 7$), which constrained our ability to draw statistical inferences.

4.2 | Short- and Long-Read Taxonomic Compositions Are Similar

To compare the two sequencing approaches (short Illumina vs. long PacBio reads), we assessed community composition and the relative abundance of assigned taxa. While both methods recovered similar overall compositions at the phylum and genus levels, significant differences were observed in the relative abundance of several taxa based on a Wilcoxon rank-sum test with Bonferroni correction (Figure 3a,b). At the phylum level, Actinobacteriota ($p = 0.0005$), Myxococcota ($p < 0.0001$), Gemmatimonadota ($p = 0.00012$) and Chloroflexi ($p = 0.014$) were significantly more abundant in the short-read dataset. At the genus level, one notable difference was a significantly higher abundance of Ferruginibacter ($p = 0.0038$) in the long-read dataset.

An ordination analysis to assess the similarity between short-read and long-read results plot (Figure 3c) showed partial overlap between the two sequencing methods, indicating shared community composition. To further quantify the similarity between the two datasets, a Procrustes analysis was conducted (Figure 3d). The results showed a strong concordance between short-read and long-read sequencing compositions ($R^2 = 0.973$, $p = 0.0001$), indicating that, despite methodological differences, both approaches captured comparable community structures. The Procrustes error plot visually represents the alignment between the datasets, with minimal deviation in most cases.

Furthermore, we assessed the number of reads assigned to bacterial phyla across sequencing methods to determine whether either approach exhibited taxonomic biases. We found a strong correlation ($R^2 = 1.02$, Deming regression) between short-read and long-read sequencing abundances (Figure 3e). The relationship between short and long reads is based on a Deming regression analysis, which yielded a slope of 1.02 (95% confidence interval (CI): 0.98–1.06), indicating that, for most phyla, short-read sequencing and long-read sequencing are highly comparable.

However, small phylum-specific trends were evident. Verrucomicrobiota, for example, exhibited higher read counts in the long-read dataset, whereas Cyanobacteria were relatively more abundant in the short-read dataset. These deviations suggest that certain bacterial phyla may be differentially represented depending on the sequencing method, potentially due to differences in primer binding efficiency, amplified fragment length, error profiles or taxonomic classification accuracy between methods.

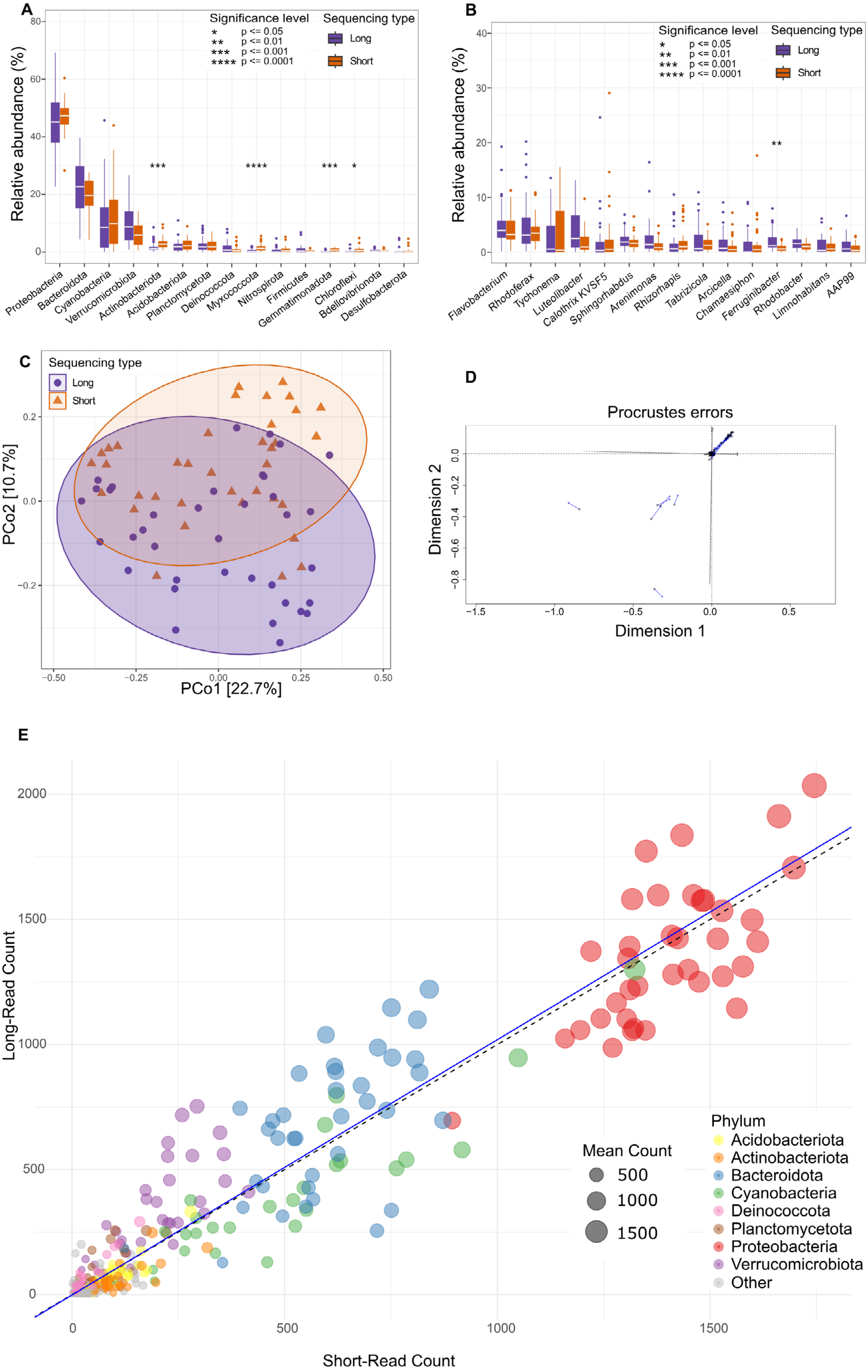

FIGURE 3 | Legend on next page.

FIGURE 3 | Comparison of bacterial taxon composition between short- (orange) and long-read (purple) sequences. Comparison of the top 15 abundant taxa at the phylum (A) and genus (B) levels for short- and long-read sequencing. Significant levels indicated by * show significant differences between long- and short-read relative abundances for each taxa. PCoA of paired samples for short- and long-read sequencing (C). (D) Procrustes error plot for paired short- and long-read samples ($p = 0.00001$). (E) Scatter plot showing the relationship between short- and long-read abundance for each phylum across all the samples. The top eight phyla are colour-coded, and the circle size is proportional to the mean number of reads per phylum in each paired sample. The dashed black line represents the line of perfect fit (1:1), and the blue line depicts the Deming regression line, with a slope of 1.02.

4.3 | Improved Taxonomic Resolution With Long Reads

The number of ASVs was 62% higher in long-read sequencing (10,480 ASVs) than in short-read sequencing (6507 ASVs) after rarefaction; the difference between pre- and post-rarefaction was not of note (Figure S5 and Table S4). At the per-sample level, paired analysis (Figure 4a) revealed a significant difference in the number of ASVs produced by each sequencing method (Wilcoxon signed-rank test: $V = 550.5$, $p = 0.0027$), with the short-read method producing significantly more ASVs.

A linear mixed-effects model (LMM) was used to assess the effect of sequencing type on the number of observed ASVs, while accounting for paired samples (Figure 4b). The model included sequencing type as a fixed effect and sample identity as a random effect to account for variations among samples. The model showed that short-read sequencing detected significantly more ASVs than long-read sequencing, with an estimated increase of 65.38 ASVs (±18.94 SE, $t = 3.45$, $p = 0.00144$) per sample. The random effect of sample identity had a variance of 10,155 (SD = 100.77), indicating a substantial between-sample variability. The residual variance was 6636 (SD = 81.46), reflecting the within-sample variation after accounting for sequencing type.

We assessed the proportion of taxonomic assignments across various ranks to determine whether the sequencing type affected taxonomic resolution. At both the kingdom and phylum levels, there were no significant differences in the percentage of taxa assigned between the short- and long-read sequencing methods; 100% of the taxa were assigned at these ranks using both approaches. At the class level, 98.9% of the taxa were assigned using short-read sequencing, compared to 99.6% with long-read sequencing—a small but statistically significant difference ($p = 0.02$; Mann–Whitney U test). More pronounced differences were observed at lower taxonomic ranks. Long-read sequencing resulted in significantly higher proportions of taxonomic assignment at the order (4.4% increase, $p < 0.0001$), family (5.8% increase, $p < 0.0001$) and genus (11.6% increase, $p < 0.0001$) levels compared to the short-read method (Figure 4c).

4.4 | Unique Features of Sequencing Type

Given the substantial taxonomic overlap between the sequencing methods, we evaluated the sequence similarity of the short- and long-read ASVs to determine which were truly unique. Using NCBI BLAST (Camacho et al. 2009), we found that on average 83.8% (58%–96%) of short-read ASVs aligned with long-read ASVs (Figure S4a) within the expected 350–750 bp region of the long-read 16S sequences, corresponding to the V4 primers used for short-read amplicon sequencing (Figure S3).

To test whether sequence identity cutoffs in BLAST influence mapping of short-read ASVs to long-read sequences, we examined the alignments at different identity thresholds. At a relaxed identity cutoff (90%), over 95% of short-read ASVs had a corresponding long-read match, whereas at a stringent 100% identity cutoff, this proportion dropped to ~60% (Figure S4a). Notably, short-read ASV length (230 bp vs. 253 bp) did not affect mapping success, as both lengths exhibited comparable match percentages (Figure S4c). These lengths were selected based on the minimum and median short-read ASV lengths.

Importantly, we analysed the composition of unique short-read ASVs that had no matches in the long-read dataset and found that they were distributed across multiple phyla. Interestingly, these short-read ASVs were predominantly associated with taxa of low prevalence (percentage found in all samples, average 5.6%) and abundance (percentage occurring in total ASVs, average 0.004%) (Figure 5b). The reverse was also true for unique long-read ASVs that had no matches to the short-read dataset. These long-read ASVs are mainly associated with taxa of low prevalence (average 3.5%) and abundance (average 0.006%) (Figure 5a).

A comparison of taxonomic annotations revealed that 977 taxa were shared between the short-read and long-read datasets, while 84 taxa (~7% of the whole dataset) were unique to short-read sequencing and 136 taxa (~11% of the whole dataset) were unique to long-read sequencing (Figure 5c). These counts were based on full taxonomic annotations from each dataset.

5 | Discussion

The dynamic spatial and temporal complexity of river ecosystems creates habitats that drive the remarkable diversity and ecological richness of aquatic microbial communities. Epilithic river biofilms represent intricate ecological niches where many environmental parameters may influence the structure and composition of microbial communities (Shibabaw et al. 2021). Consequently, the accurate characterisation of these microbial communities requires molecular approaches that can capture the subtle taxonomic and functional diversity inherent in these dynamic systems. To comprehensively investigate the microbial landscape across seven distinct river sites in England, we used two sequencing technologies, Illumina short-read and PacBio

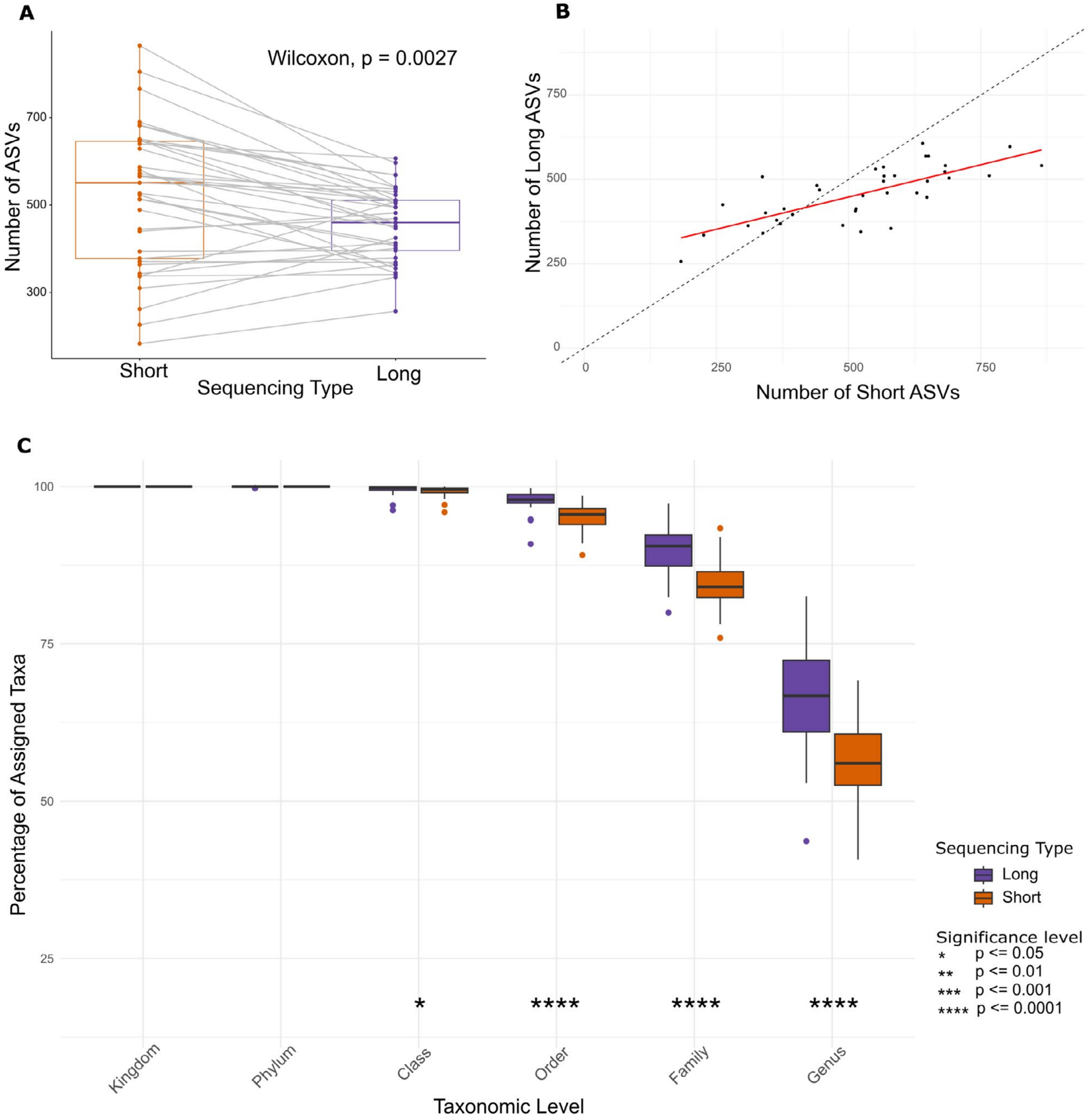

FIGURE 4 | Comparison of taxonomic assignment and ASV detection between long-read (purple) and short-read (orange) sequencing methods. (A) Paired comparison of ASV counts per sample, analysed using the Wilcoxon signed-rank test ($V=550.5$, $p=0.0027$). (B) Relationship between ASV counts detected by short- and long-read sequencing, where the dashed black line represents a 1:1 ratio (perfect agreement), and the red line represents the fitted linear model (LLM). Short-read sequencing detected significantly more ASVs per sample than long-read sequencing, with an estimated increase of 65.38 ASVs (±18.94 SE, $t=3.45$, $p=0.00144$) per sample. (C) Percentage of assigned taxa at different taxonomic levels for long- and short-read sequencing, Mann–Whitney *U* test to test for significant difference between sequencing types, *p* values are represented by *.

long-read sequencing, targeting the 16S ribosomal RNA (rRNA) gene. This approach enabled a comparative assessment of the microbial community structure, providing the opportunity to distinguish the capabilities of each sequencing platform in resolving complex bacterial assemblages embedded within river biofilm environments.

Our analysis revealed that, for 16S rRNA-based taxonomic assessments of river biofilms, the choice of sequencing method (Illumina short-read or PacBio long-read) did not significantly influence the relative abundance of taxa within bacterial communities. Despite the disparity in read length and taxonomic resolution, both sequencing platforms produced broadly comparable abundance profiles across major taxonomic groups. This suggests that short-read sequencing, although limited in its ability to resolve taxa at deeper levels, still captures reliable patterns in community structure. Consequently, it remains a practical and informative approach for studies focused on broad-scale surveys or relative abundance patterns. These findings are consistent with previous studies (e.g., Butt et al. 2022;

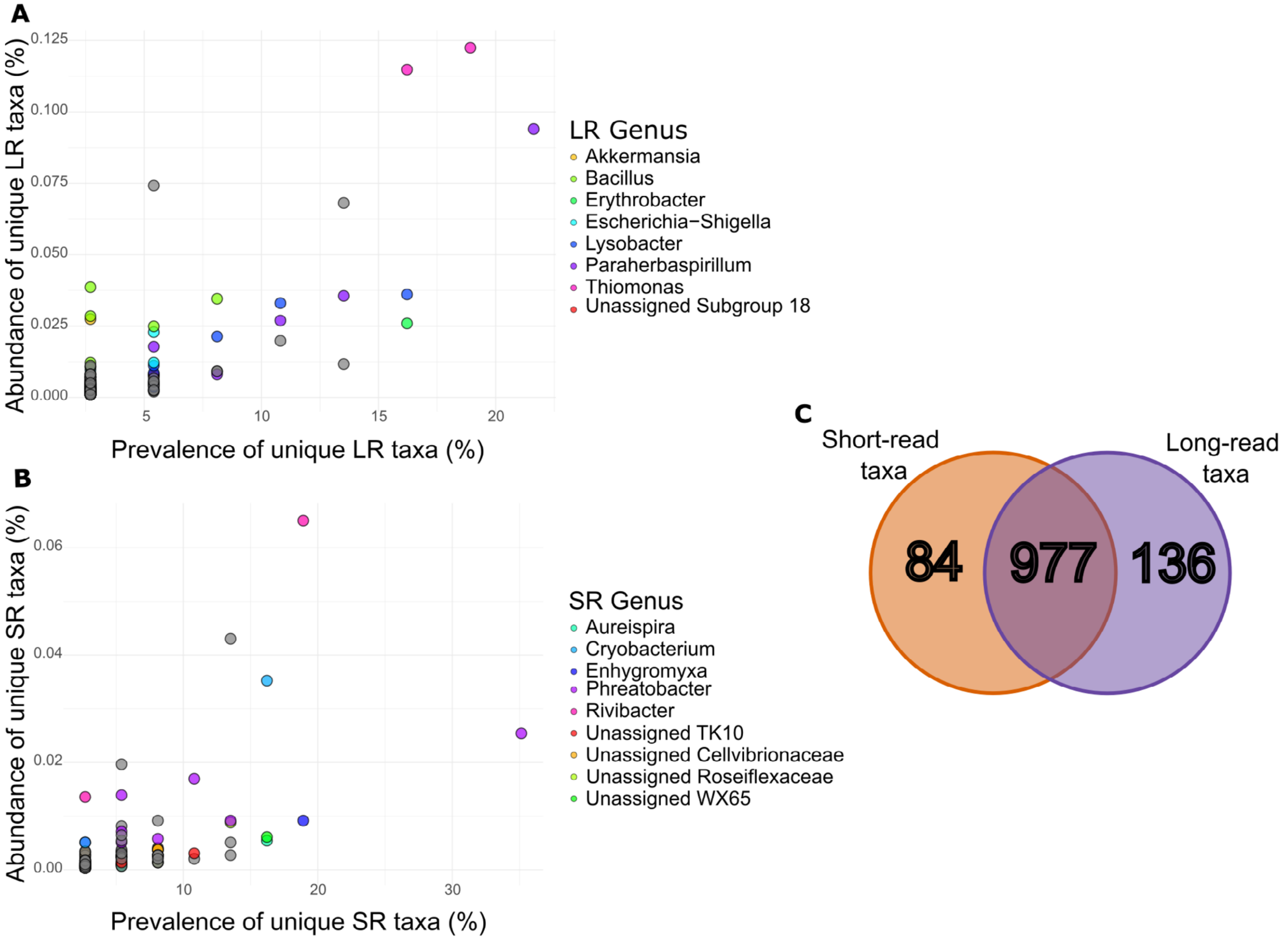

FIGURE 5 | The prevalence (the % of total samples detected in) and abundance (the % of total reads detected in) of (A) long-read (LR) ASVs that did not have $a \geq 98\%$ match to any short-read ASVs in the whole dataset and (B) short-read (SR) ASVs that did not have $a \geq 98\%$ match to any long-read ASVs in the whole dataset, coloured by the top genera. (C) Venn diagram showing the number of unique taxa in the short-read (orange) and long-read (purple) data, and the overlap of the datasets.

Buetas et al. 2024), which reported largely concordant diversity measures between sequencing platforms, and support the use of either method depending on the specific ecological question being asked.

Although long-read sequencing improved taxonomic resolution, particularly at finer levels, it did not significantly alter per-sample richness or diversity estimates (e.g., Chao1 or Shannon; Figure S2). It must, however, be noted that the environmental gradients are the likely selecting factors reflected in the Procrustes analysis, which may be a stronger driver compared to technical variabilities in read length. Interestingly, while the total number of unique ASVs across the dataset was higher in the long-read data, individual samples contained significantly more ASVs in the short-read data. This likely reflects greater sequencing depth in the short-read dataset, allowing detection of more low-abundance variants per sample. In contrast, the higher resolution of long-read data can split similar sequences into more distinct ASVs across the dataset, inflating the total count. This inflation of total ASV count in the long-read data highlights how sequencing platform choices must be considered when interpreting ASV-based metrics. Further research will be needed to determine whether the higher per-sample ASV counts observed in the Illumina dataset reflect genuine biological diversity, such as improved detection of rare taxa resulting from greater sequencing depth. It is also possible that these higher counts are partially inflated by technical artefacts inherent to short-read data, including residual sequencing errors or undetected chimeras that denoising algorithms like DADA2 may not completely resolve (Haas et al. 2011).

Importantly, our findings demonstrate that long-read sequencing provides superior taxonomic resolution compared to short-read methods for analysing microbial communities within river biofilms. This aligns with a report by Gao et al. (2024), who also reported an increased taxonomic resolution of ASVs, including precision at the species and strain levels. However, we did not perform species-level taxonomic comparisons, as only a small proportion of ASVs could be confidently assigned to species across both sequencing methods. This limited resolution likely stems from incomplete reference databases for freshwater microbes, combined with the challenges of accurate species-level classification using current algorithms, especially for short-read data. Consequently, we focused our comparisons on higher taxonomic ranks where assignments were more robust and informative.

PacBio's ability to generate longer contiguous sequences facilitated a more accurate taxon classification, particularly at deeper taxonomic levels, such as genera and species. This was also recently shown by Buetas et al. (2024), who found that all species were correctly identified using PacBio sequencing compared to Illumina short reads using a mock community. This advantage is particularly evident in complex microbial assemblages, where short-read sequences often fail to resolve ambiguities owing to their limited length and reliance on overlapping fragments for their assembly. To retrieve as much taxonomic information as possible we used lower bootstrap values which may have contributed to inflated assignment rates. Irrespective of this, our analyses revealed that many sequences that remained unclassified in the short-read data matched successfully identified taxa in the long-read dataset. This improved classification makes the long-read dataset inherently more robust and reliable for the comprehensive assessment of biodiversity. This increase in the number of classified taxa is important for biodiversity assessments, as it allows for a more comprehensive understanding of community composition and structure. Such detailed insights are essential for applications such as biomonitoring and ecological research, where understanding the full spectrum of biodiversity is a priority.

Both methodologies, as highlighted previously by existing research, have advantages and disadvantages (Buetas et al. 2024; Eisenhofer et al. 2024; Gao et al. 2024). For example, Buetas et al. (2024) highlighted that Illumina provided an 8-fold higher throughput and lower cost than PacBio. Although in their study and ours, we identified a higher number of ASVs at the per-sample level in the Illumina data, and the overall increase in ASVs was not exponential in relation to the cost. This was similarly reported by Cook et al. (2024), who reiterated the need for deeper sequencing with long reads to achieve parity with the short-read methodology. This is further supported by the observation of unique taxa within the Illumina samples compared to the PacBio samples in our study. Similar to the findings of Buetas et al. (2024), the unique ASVs identified by the short-read method were of relatively low abundance (~0.004%) and prevalence (~5.6%). We found the same trend in the long-read dataset, where unique ASVs had an average abundance of ~3.5% and prevalence of just ~0.006%. In terms of taxonomic assignments, 84 taxa were unique to short-read sequencing, 136 were unique to long-read sequencing, and 977 were shared between both methods. These unique taxa represented a small proportion of the total detected diversity—approximately 7% for short-read and 11% for long-read data. The higher number of unique taxa in the long-read dataset likely reflects the increased sequence length, which enhances the ability to resolve subtle differences between closely related organisms. Despite these differences, the substantial overlap in taxonomic composition demonstrates that both methods capture broadly consistent community profiles. From a biomonitoring perspective, this highlights the utility of long-read sequencing for enhancing taxonomic resolution without compromising comparability with established short-read approaches. Based on our findings, even at the genus level, there was a large inconsistency between the PacBio and Illumina taxonomic assignments (Figure S4b). This is likely due to the current state of databases, which are varied and not yet fully standardised or validated. For example, bespoke databases were developed by Lo et al. (2023) for aquatic pathogens. Similarly, others have reported that despite PacBio sequencing annotating more reads to the species level, the vast majority were taxonomically unassigned because of the possible underrepresentation of species in databases (Pasolli et al. 2019). With the increase in long-read methods for biomonitoring, database selection will have a critical impact. Our findings are in concordance with other reports, as highlighted above; therefore, the availability of curated databases in the future will play a major role (Sierra et al. 2020).

Collectively, our findings comparing the utility of short- and long-read methods for biomonitoring across the RSN revealed the accuracy and possible pitfalls of each sequencing technology. We acknowledge several limitations exist, including but not limited to primer bias, differences in the amplified 16S rRNA gene regions and potential error/correction rates across the two methodologies. Primer bias, in particular, can affect which taxa are preferentially amplified, potentially skewing community composition. To mitigate this, we trimmed the long-read sequences to match the same region amplified by the short-read primers and re-analysed the data using the DADA2 pipeline with the same downstream processing, where ordinations revealed that the trimmed long reads overlapped across both the original long-read dataset and the short-read dataset (Figures S8, S9). Furthermore, taxonomic assignment and ASV richness in the trimmed dataset were similar to those from the short-read data (Figure S7). These findings confirm that primer bias and the sequenced regions potentially contribute to observed differences in community composition. However, this bias is an expected feature of amplicon sequencing, and our study was specifically designed to assess the impact of sequencing technology, rather than primer performance, on taxonomic resolution and community profiling. PCR primer sets can introduce variability in taxonomic recovery, as different regions of the 16S rRNA gene (e.g., V4 versus other variable regions) capture distinct portions of microbial diversity. Consequently, primer choice can influence apparent community composition and relative abundance patterns. These effects are well documented in microbial ecology studies using short-read sequencing approaches (Apprill et al. 2015; Klindworth et al. 2013). The existing workflows pertaining to error corrections are still in their infancy with long reads and these areas require future research to understand the nuances of user choices in influencing outcomes of eDNA analysis (Bylemans et al. 2025).

Despite this, we showed an increased resolution of taxonomic assignment using PacBio long-read sequence data, while simultaneously highlighting the possible inadequacy of sequencing depth using this platform, which can currently be achieved more cost-effectively using short-read technologies such as Illumina. It is important to reiterate that the optimal sequencing strategy depends on the research question, and that the specific goals or priorities of the study may dictate which method is chosen as appropriate. Overall, our data provide critical insights into the current molecular biomonitoring landscape and may serve as a valuable resource for future comparisons and subsequent benchmarks, particularly in environmental and ecological contexts.

Author Contributions

Physicochemical data and samples were collected by the Environment Agency. D.S.R., M.A.J.A., K.W. and J.W. conceptualised the study. M.A.J.A. and A.C.T. performed sample processing, DNA extraction and sequencing. M.A.J.A. and S.B.B. designed and performed the bioinformatic analysis. M.A.J.A. performed statistical analysis and figure production. M.A.J.A. and S.B.B. drafted the manuscript. All authors read and revised the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://www.ncbi.nlm.nih.gov/bioproject/1353123, reference number https://doi.org/10.5281/zenodo.17432155.

References

Amarasinghe, S. L., S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil. 2020. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." *Genome Biology* 21, no. 1: 30. https://doi.org/10.1186/s13059-020-1935-5.

Apprill, A., S. McNally, R. Parsons, and L. Weber. 2015. "Minor Revision to V4 Region SSU rRNA 806R Gene Primer Greatly Increases Detection of SAR11 Bacterioplankton." *Aquatic Microbial Ecology* 75: 129–137. https://doi.org/10.3354/ame01753.

Barthélémy, D., E. Belmonte, L. D. Pilla, et al. 2024. "Correction: Barthélémy Et al. Direct Comparative Analysis of a Pharmacogenomics Panel With PacBio Hifi Long-Read and Illumina Short-Read Sequencing." *Journal of Personalized Medicine* 14, no. 10: 1028. https://doi.org/10.3390/jpm14101028.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456, no. 7218: 53–59. https://doi.org/10.1038/nature07517.

Buetas, E., M. Jordán-López, A. López-Roldán, et al. 2024. "Full-Length 16S rRNA Gene Sequencing by PacBio Improves Taxonomic Resolution in Human Microbiome Samples." *BMC Genomics* 25, no. 1: 310. https://doi.org/10.1186/s12864-024-10213-5.

Butt, S. L., H. M. Kariithi, J. D. Volkening, et al. 2022. "Comparable Outcomes From Long and Short Read Random Sequencing of Total RNA for Detection of Pathogens in Chicken Respiratory Samples." *Frontiers in Veterinary Science* 9: 1073919. https://doi.org/10.3389/fvets.2022.1073919.

Bylemans, J., T. Everts, R. Brys, and R. P. Duncan. 2025. "From Anarchy to Clarity, Data Pre-Processing and Statistical Choices Influence Quantitative Environmental DNA (eDNA) Analyses." *Methods in Ecology and Evolution* 16: 1322–1333. https://doi.org/10.1111/2041-210X.70064.

Callahan, B. J., J. Wong, C. Heiner, et al. 2019. "High-Throughput Amplicon Sequencing of the Full-Length 16S rRNA Gene With Single-Nucleotide Resolution." *Nucleic Acids Research* 47, no. 18: e103. https://doi.org/10.1093/nar/gkz569.

Camacho, C., G. Coulouris, V. Avagyan, et al. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10, no. 1: 421. https://doi.org/10.1186/1471-2105-10-421.

Cook, R., A. Telatin, S.-Y. Hsieh, et al. 2024. "Nanopore and Illumina Sequencing Reveal Different Viral Populations From Human Gut Samples." *Microbial Genomics* 10, no. 4: e001236. https://doi.org/10.1099/mgen.0.001236.

Eisenhofer, R., J. Nesme, L. Santos-Bay, et al. 2024. "A Comparison of Short-Read, HiFi Long-Read, and Hybrid Strategies for Genome-Resolved Metagenomics." *Microbiology Spectrum* 12, no. 4: e0359023. https://doi.org/10.1128/spectrum.03590-23.

Environment Agency. 2024. "Molecular Data Generation and Preliminary Analysis of River Microbial Biofilm Communities." Environment Agency, Bristol.

Farrell, J. A., L. Whitmore, and D. J. Duffy. 2021. "The Promise and Pitfalls of Environmental DNA and RNA Approaches for the Monitoring of Human and Animal Pathogens From Aquatic Sources." *Bioscience* 71, no. 6: 609–625. https://doi.org/10.1093/biosci/biab027.

Ferrarini, M., M. Moretto, J. A. Ward, et al. 2013. "An Evaluation of the PacBio RS Platform for Sequencing and De Novo Assembly of a Chloroplast Genome." *BMC Genomics* 14, no. 1: 670. https://doi.org/10.1186/1471-2164-14-670.

Galata, V., S. B. Busi, B. J. Kunath, et al. 2021. "Functional Meta-Omics Provide Critical Insights Into Long- and Short-Read Assemblies." *Briefings in Bioinformatics* 22, no. 6: bbab330. https://doi.org/10.1093/bib/bbab330.

Gao, J., Z. Wang, W. Deng, et al. 2024. "Improved Resolution of Microbial Diversity in Deep-Sea Surface Sediments Using PacBio Long-Read 16S rRNA Gene Sequencing." *M Sphere* 9: e0077024. https://doi.org/10.1128/msphere.00770-24.

Goodwin, S., J. D. McPherson, and W. R. McCombie. 2016. "Coming of Age: Ten Years of Next-Generation Sequencing Technologies." *Nature Reviews Genetics* 17, no. 6: 333–351. https://doi.org/10.1038/nrg.2016.49.

Haas, B. J., D. Gevers, A. M. Earl, et al. 2011. "Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-Pyrosequenced PCR Amplicons." *Genome Research* 21, no. 3: 494–504. https://doi.org/10.1101/gr.112730.110.

Hon, T., K. Mars, G. Young, et al. 2020. "Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes." *Scientific Data* 7, no. 1: 399. https://doi.org/10.1038/s41597-020-00743-4.

Johnson, J. S., D. J. Spakowicz, B.-Y. Hong, et al. 2019. "Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis." *Nature Communications* 10, no. 1: 5029. https://doi.org/10.1038/s41467-019-13036-1.

Keck, F., R. C. Blackman, R. Bossart, et al. 2022. "Meta-Analysis Shows Both Congruence and Complementarity of DNA and eDNA Metabarcoding to Traditional Methods for Biological Community Assessment." *Molecular Ecology* 31, no. 6: 1820–1835. https://doi.org/10.1111/mec.16364.

Kelly, M. G., S. Juggins, D. G. Mann, et al. 2020. "Development of a Novel Metric for Evaluating Diatom Assemblages in Rivers Using DNA Metabarcoding." *Ecological Indicators* 118: 106725. https://doi.org/10.1016/j.ecolind.2020.106725.

Klindworth, A., E. Pruesse, T. Schweer, et al. 2013. "Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and Next-Generation Sequencing-Based Diversity Studies." *Nucleic Acids Research* 41, no. 1: e1. https://doi.org/10.1093/nar/gks808.

Littlefair, J. E., J. S. Hleap, V. Palace, M. D. Rennie, M. J. Paterson, and M. E. Cristescu. 2023. "Freshwater Connectivity Transforms Spatially Integrated Signals of Biodiversity." *Proceedings of the Royal Society B: Biological Sciences* 290, no. 2006: 20230841. https://doi.org/10.1098/rspb.2023.0841.

Liu, C., Y. Cui, X. Li, and M. Yao. 2021. "Microeco: An R Package for Data Mining in Microbial Community Ecology." *FEMS Microbiology Ecology* 97, no. 2: fiaa255. https://doi.org/10.1093/femsec/fiaa255.

Lo, L. S. H., X. Liu, H. Liu, M. Shao, P.-Y. Qian, and J. Cheng. 2023. "Aquaculture Bacterial Pathogen Database: Pathogen Monitoring and Screening in Coastal Waters Using Environmental DNA." *Water Research X* 20: 100194. https://doi.org/10.1016/j.wroa.2023.100194.

Logsdon, G. A., M. R. Vollger, and E. E. Eichler. 2020. "Long-Read Human Genome Sequencing and Its Applications." *Nature Reviews Genetics* 21, no. 10: 597–614. https://doi.org/10.1038/s41576-020-0236-x.

Morton, R. D., C. G. Marston, A. W. O'Neil, and C. S. Rowland. 2021. "Land Cover Map 2020 (Land Parcels, GB)." NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/0e99d57e-1757-451f-ac9d-92fd1256f02a.

Newbold, L. K., J. D. Taylor, A. C. Thorpe, J. Warren, K. Walsh, and D. S. Read. 2025. "DNA Extraction Methodology Has a Limited Impact on Multitaxa Riverine Benthic Metabarcoding Community Profiles." *Environmental DNA* 7, no. 3: e70102. https://doi.org/10.1002/edn3.70102.

Ota, W. M., C. Hall, J. Malloy, and M. A. Clark. 2020. "Environmental DNA Monitoring: Better Tracking of Endangered, Rare, Cryptic, and Invasive Species." *Journal of Science Policy & Governance* 17, no. 1: 1038126. https://doi.org/10.38126/JSPG170117.

Pasolli, E., F. Asnicar, S. Manara, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes From Metagenomes Spanning Age, Geography, and Lifestyle." *Cell* 176, no. 3: 649–662. https://doi.org/10.1016/j.cell.2019.01.001.

Porter, T. M., and M. Hajibabaei. 2018. "Scaling Up: A Guide to High-Throughput Genomic Approaches for Biodiversity Analysis." *Molecular Ecology* 27, no. 2: 313–338. https://doi.org/10.1111/mec.14478.

Quast, C., E. Pruesse, P. Yilmaz, et al. 2012. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41, no. Database issue: D590–D596. https://doi.org/10.1093/nar/gks1219.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rhoads, A., and K. F. Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics* 13, no. 5: 278–289. https://doi.org/10.1016/j.gpb.2015.08.002.

Ruppert, K. M., R. J. Kline, and M. S. Rahman. 2019. "Past, Present, and Future Perspectives of Environmental DNA (eDNA) Metabarcoding: A Systematic Review in Methods, Monitoring, and Applications of Global eDNA." *Global Ecology and Conservation* 17: e00547. https://doi.org/10.1016/j.gecco.2019.e00547.

Satam, H., K. Joshi, U. Mangrolia, et al. 2023. "Next-Generation Sequencing Technology: Current Trends and Advancements." *Biology* 12, no. 7: 997. https://doi.org/10.3390/biology12070997.

Shendure, J., S. Balasubramanian, G. M. Church, et al. 2017. "DNA Sequencing at 40: Past, Present and Future." *Nature* 550, no. 7676: 345–353. https://doi.org/10.1038/nature24286.

Shibabaw, T., A. Beyene, A. Awoke, M. Tirfie, M. Azage, and L. Triest. 2021. "Diatom Community Structure in Relation to Environmental Factors in Human Influenced Rivers and Streams in Tropical Africa." *PLoS One* 16, no. 2: e0246043. https://doi.org/10.1371/journal.pone.0246043.

Sierra, M. A., Q. Li, S. Pushalkar, et al. 2020. "The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community." *Genes* 11, no. 8: 878. https://doi.org/10.3390/genes11080878.

Srinivas, M., C. J. Walsh, F. Crispie, et al. 2025. "Evaluating the Efficiency of 16S-ITS-23S Operon Sequencing for Species Level Resolution in Microbial Communities." *Scientific Reports* 15, no. 1: 2822. https://doi.org/10.1038/s41598-024-83410-7.

Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg. 2012. "Environmental DNA." *Molecular Ecology* 21, no. 8: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x.

Tedersoo, L., A. Tooming-Klunderud, and S. Anslan. 2018. "PacBio Metabarcoding of Fungi and Other Eukaryotes: Errors, Biases and Perspectives." *New Phytologist* 217, no. 3: 1370–1385. https://doi.org/10.1111/nph.14776.

Thomas, A. C., S. Tank, P. L. Nguyen, J. Ponce, M. Sinnesael, and C. S. Goldberg. 2020. "A System for Rapid eDNA Detection of Aquatic Invasive Species." *Environmental DNA* 2, no. 3: 261–270. https://doi.org/10.1002/edn3.25.

Thompson, L. R., J. G. Sanders, D. McDonald, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551, no. 7681: 457–463. https://doi.org/10.1038/nature24621.

van Dijk, E. L., Y. Jaszczyszyn, D. Naquin, and C. Thermes. 2018. "The Third Revolution in Sequencing Technology." *Trends in Genetics* 34, no. 9: 666–681. https://doi.org/10.1016/j.tig.2018.05.008.

Wang, S., X. Su, H. Cui, et al. 2022. "Microbial Richness of Marine Biofilms Revealed by Sequencing Full-Length 16S rRNA Genes." *Genes* 13, no. 6: 1050. https://doi.org/10.3390/genes13061050.

Yang, B., Y. Wang, and P.-Y. Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17, no. 1: 135. https://doi.org/10.1186/s12859-016-0992-y.

Zhang, J., L. Su, Y. Wang, and S. Deng. 2020. "Improved High-Throughput Sequencing of the Human Oral Microbiome: From Illumina to PacBio." *Canadian Journal of Infectious Diseases and Medical Microbiology* 2020: 1–13. https://doi.org/10.1155/2020/6678872.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** men70075-sup-0001-DataS1.zip.