Minor Library Preparation Modification Significantly Reduces Barcode Crosstalk in ONT Multiplexed Sequencing

Qing Dai[1], Claudia K. Gunsch[1], and Joshua A. Granek[2,*]

[1]Civil and Environmental Engineering, Pratt School of Engineering, Duke University, Durham NC, 27708, USA
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham NC, 27708, USA
[*]joshua.granek@duke.edu

ABSTRACT

Barcode crosstalk is a potential source of error for multiplexed high-throughput sequencing. It results in reads misassigned to the wrong sample, mimicking cross-contamination. We identify, quantify, and present a solution to barcode crosstalk in the ONT library preparation. Post-Ligation Pooling improves read misassignment more than ten-fold with a minor modification. Our results highlight the problem of misassignment and a simple solution to improve data quality.

Main

Index barcodes enable multiplexing in high-throughput sequencing experiments. By allowing multiple samples to be sequenced together, per-sample costs are lowered and batch effects can be eliminated. However, barcoding introduces the potential for read misassignment: reads attributed to the wrong sample. Misassignment has two sources: demultiplexing error and barcode crosstalk (also termed barcode hopping or index misassignment). Demultiplexing errors occur during data processing when barcode sequence errors prevent a correct match to the expected barcode; the primary consequence is loss of usable data. Barcode crosstalk is more pernicious because is masquerades as cross-contamination; it occurs during library-preparation when an incorrect barcode is ligated to a nucleic-acid fragment.

Mitigating barcode crosstalk is essential for accuracy and reproducibility, because even low-level events can generate false positives, inflate diversity estimates, and distort quantitative analyses[1,2]. Demultiplexing error can be mitigated by algorithmic improvements and error correction[3,4]. The physical and biochemical contributors to barcode crosstalk, such as adapter carryover, bead cleanup efficiency, and library pooling strategies, remain less systematically studied than demultiplexing error. Mitigation of barcode crosstalk requires careful re-design of experiments. One strategy, unique dual index (UDI) allows crosstalk-derived reads to be bioinformatically detected and excluded. Use of UDIs reduces misassignment from 0.1–6% to <0.01%[5] on Illumina patterned flow cells. However this strategy is not simple to implement on other platforms.

We observed apparent cross-contamination in our shotgun metagenomic data generated using Oxford Nanopore Technologies (ONT) sequencers from built-environment microbiome samples: negative controls (molecular biology–grade water blank and no-input empty blank) and positive controls contained sequences from each other and p-trap water sample (Figure 1A). After ruling out common contamination sources, including "kitome" and "splashome"[1], we considered barcode crosstalk.

In the standard ONT ligation protocol, samples are pooled before adapter ligation, immediately after index barcoding. We developed the PLP protocol, in which libraries are pooled after final adapter ligation. PLP greatly reduced apparent cross-contamination, eliminating misassigned DNA Control Strand (DCS; ONT's standard positive control, which is an amplicon from the *lambda* phage genome) reads, nearly eliminating reads in blanks, reducing reads in controls by one to two orders of magnitude, and increasing positive control on-target reads (DCS: from ~84% to 94%; *PhiX* from ~20% to ~50%; Figure 1B; Tables S1 and S2).

Our environmental data show that PLP substantially reduces barcode crosstalk in a real-world context sensitive to cross-contamination. To further quantify crosstalk-induced misassignment we designed a simple system of four defined DNA samples (details in *Methods*). This experiment compared the standard ONT ligation protocol, PLP, ONT's recently released Short Fragment Buffer (SFB) cleanup, and a combined SFB+PLP protocol.

To avoid confounding protocol performance with sample or run-specific effects, we used a combinatorial, Latin-square–inspired design stratified by input amount. For each protocol comparison, the four genomes were split into complementary two-sample groups; within each run, two groups were prepared with the two protocols being compared, pooled at the adapter-cleanup step, and sequenced on the same flow cell (Table S3). Sequencing both library types on the same flow cell ensured that differences

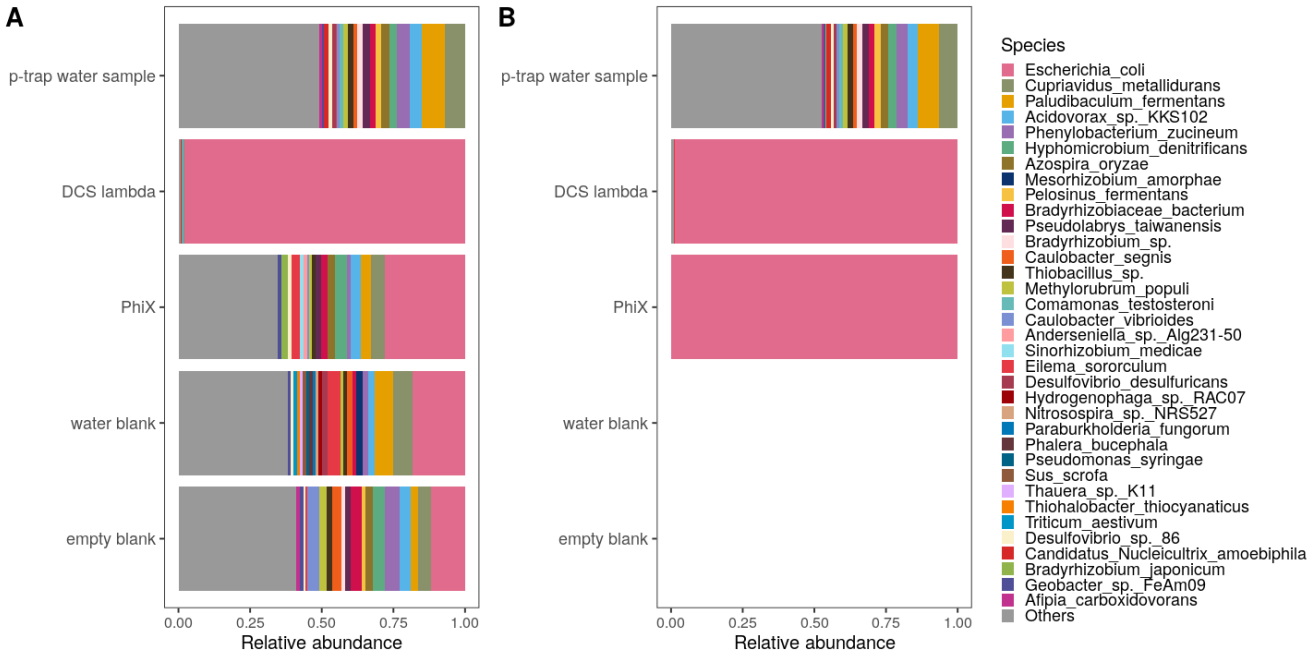

Figure 1. Taxonomic classification of libraries prepared with (A) standard protocol, and (B) modified PLP protocol. Note that the positive controls (DCS *lambda* and *PhiX*) are classified as *Escherichia coli* because both are bacteriophages found in the *E. coli* genome.

reflected the library preparation rather than run-to-run variability.

Across all experiments, barcode crosstalk was markedly lower with modified library-prep strategies (Figure 2; Table S4). Under the standard protocol, we observed the highest misassignment rate (up to 0.5–0.6%). PLP alone reduced barcode crosstalk by one to two orders of magnitude, increasing demultiplexing precision from ~99% to ~99.96%, and recall from 98.6% to ≥99.9%, with misassigned reads falling from 0.362% to 0.023%. ONT's SFB cleanup also improved performance (misassignment reduced from 0.362% to 0.064%), and the combined SFB+PLP protocol gave the best results (Figure 2, Tables S8, S7). Notably, PLP-based workflows improved the fidelity without compromising data yield.

Mechanistically, seems to result when DNA fragments and residual free barcodes share a common ligation environment. In the standard workflow (Figure 3, left), samples are pooled before final adapter attachment, allowing residual barcodes to ligate onto sample DNA during pooled adapter ligation. In our PLP workflow (Figure 3, right), samples remain physically isolated through all ligation steps and are pooled only after final adapter ligation, removing the shared ligation environment. This re-ordered pooling strategy is a simple change that requires only minimal deviations from the standard protocol and can be implemented with existing kits without purchasing additional reagents.

Our findings extend earlier reports of index misassignment across different sequencing platforms. Illumina-based pooled tagged-amplicon workflows can exhibit tag-jumping artifacts >2%[1], and even a small index-hopping rate of ~0.3–0.9% on Illumina patterned flow cells can yield >8% index-hopped molecules in low-input single-cell libraries[6]. In our ONT dataset, the standard protocol's misassignment rate was of similar magnitude, a level unacceptable for many applications. In contrast, the PLP+SFB method reduced ONT misassignment to ~0.01-0.02%, approaching Illumina's UDI workflows (<0.01%[5]). This improvement is especially important in low-biomass contexts, where trace contamination can overwhelm the signal[7] (Figure 1A).

More broadly, by substantially reducing barcode misassignment, the PLP and SFB+PLP workflows increase the suitability for applications where cross-sample contamination is a concern. ONT sequencing is increasing used in new applications as varied as simultaneous long-read profiling of somatic mutations and T-cell receptor repertoires[8] and IgA-coating indexing[9] With the PLP protocol (with or without SFB wash), libraries prepared from only a few nanograms of DNA can be multiplexed with greater confidence in intrinsically rare signals such as those in single-cell sequencing or environmental DNA detection in low-biomass microbiome samples. Our results suggest that simple protocol-level changes can help bring ONT multiplexing performance on par with best-in-class indexed short-read workflows and may support more robust multiplexed ONT sequencing in clinical, environmental, and discovery-oriented studies.

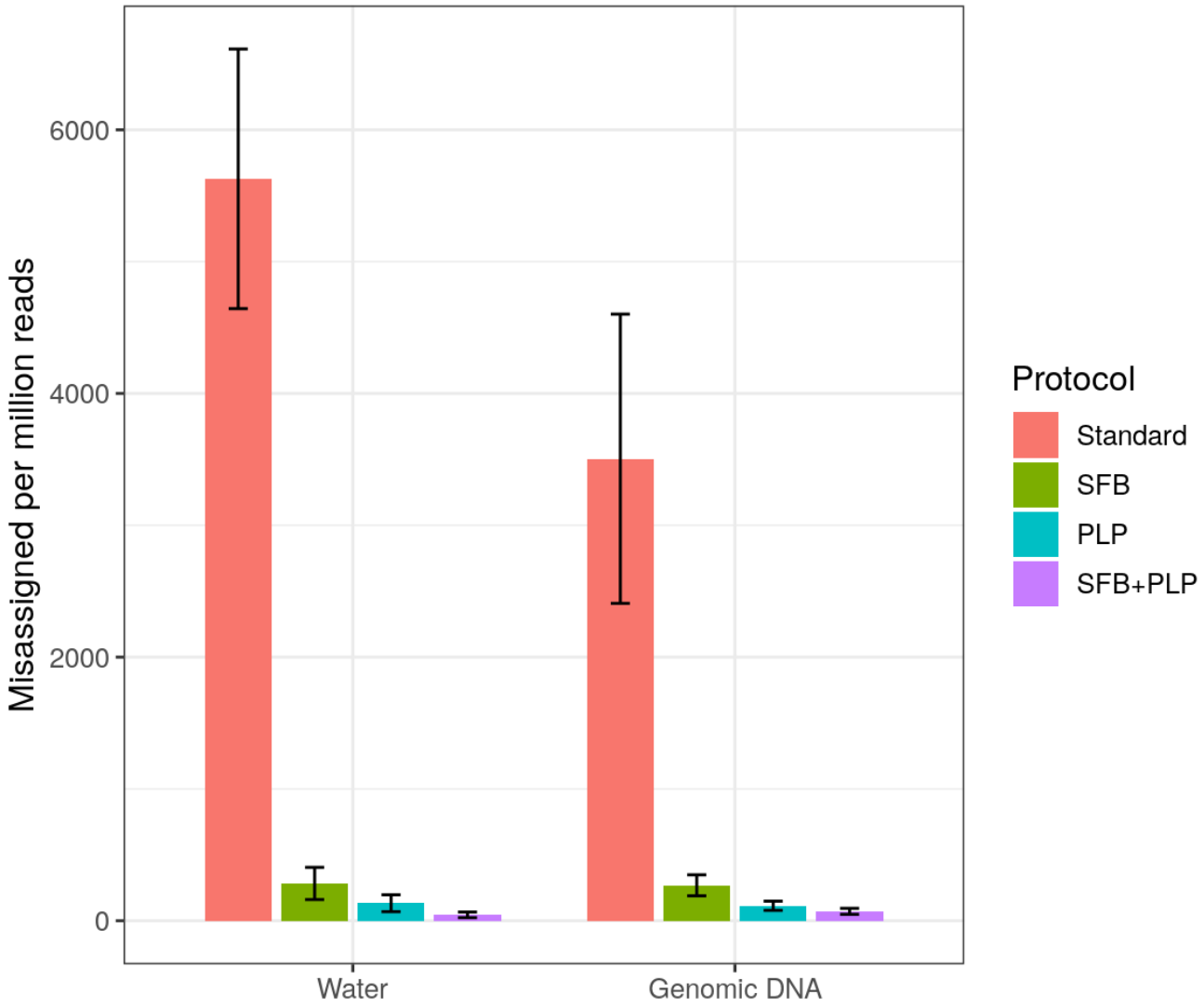

Figure 2. Misassigned reads per million mapped reads across different treatment.

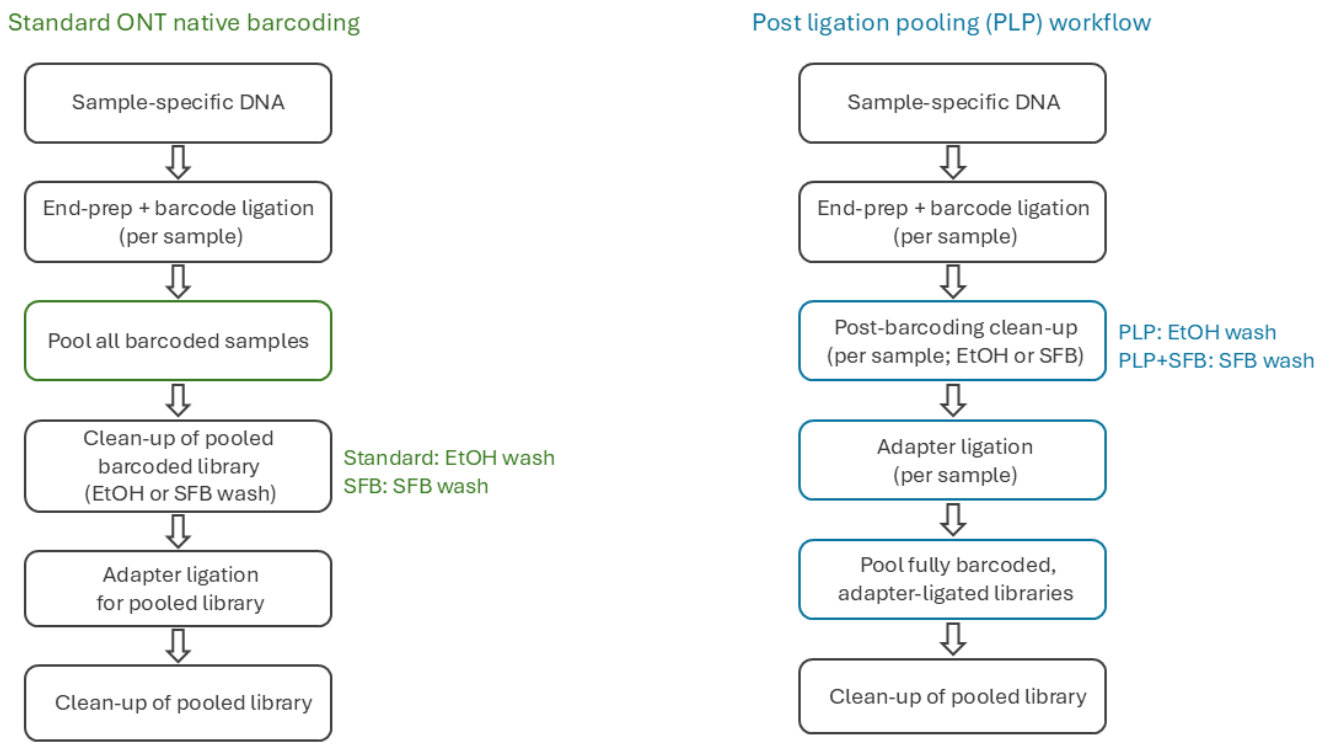

Figure 3. Comparison of pooling strategy in ONT native barcoding workflows

Methods

Sequencing and Data Preprocessing

The Native Barcoding Kit (ONT, SQK-NBD114.96) was used for all DNA library preparation, with protocol modifications discussed below.

POD5 files were basecalled and demultiplexed with Dorado v1.0.2[10] using a custom pipeline (https://gitlab.oit.duke.edu/granek-lab/projects/dorado-pipelines).

Metagenomic Experiment and Analysis

Genomic DNA extracted from a p-trap water sample using QIAGEN DNeasy PowerSoil Pro Kits. Positive controls were DCS DNA (ONT) and *phiX* DNA (New England Biolabs, ΦX174 RF I DNA). Negative controls were a water blank (molecular biology–grade water substituted for DNA sample) and an "empty blank" (nothing was substituted for sample DNA during the library preparation). For each of these samples (p-trap, DCS *lambda*, *phiX*, water blank, and empty blank), we prepared two technical libraries from the same DNA: one using the PLP protocol and one using the standard protocol. The PLP and standard libraries were then sequenced on two separate Flongle flow cells, with 10 ng input DNA per sample (except negative controls) in all cases.

The Taxonomic-Profiling-Minimap-Megan pipeline[11] (https://github.com/PacificBiosciences/pb-metagenomics-tools/tree/master/Taxonomic-Profiling-Minimap-Megan), with minor adaptations for running on a cluster through SLURM, was used for taxonomic profiling. Crosstalk of positive control reads in the metagenomic data was quantified S2) using a custom reference consiting of the DCS *lambda* ampicon and *PhiX174* genome sequences.

Defined Sample Experiment and Analysis

Defined Sample Experimental Design

To evaluate barcode crosstalk under different library preparation conditions, three ATCC bacterial genomic DNA (ATCC 8482 *Phocaeicola vulgatus*, ATCC 23114 *Mycoplasma hominis* and ATCC 6940 *Corynebacterium striatum*) and the positive control DNA (DCS, *lambda* amplicon DNA provided in the ONT kit) were used as samples with known sequence, allowing barcode-to-sample traceability for downstream validation.

Each treatment condition included two DNA samples and one blank (molecular biology–grade water) control and were run on both Flongle and MinION flowcells. To minimize batch effects resulting from variation in flow cell performance (e.g., available pores, active channel counts, or runtime stability), we implemented a combinatorial, Latin-square–inspired design stratified by input amount (Table S3). For each protocol comparison (standard vs PLP, SFB vs PLP, and SFB vs PLP+SFB), the four genomes were partitioned into two complementary two-species mixtures, with different pairings used for the Flongle and 20 ng MinION runs. Within each run, the paired mixtures were prepared using the two protocols being compared, pooled after the adapter-cleanup step, and sequenced on the same flow cell.

Library Preparation and Sequencing

The standard preparation follows ONT's Native Barcoding Kit protocol prior to Oct. 2025 (barcode, pool, ethanol wash, adapter ligation). The SFB preparation follows the ONT's protocol released in Oct. 2025 (the standard protocol, except the ethanol wash is replaced with an SFB wash. PLP moves the pooling step after adapter ligation. PLP+SFB also moves the pooling step after adapter ligation and uses SFB instead of Figure 3) Detailed step-by-step procedures are provided as Supplementary Information.

For each comparison, paired experiments were performed and sequenced on same flow cells: standard vs PLP, SFB vs PLP, and SFB vs SFB + PLP. In total, three MinION and three Flongle flow cells were used. Each Flongle library was prepared from 3 ng of input DNA per sample (12 ng in total), while each MinION library used 20 ng of input DNA per sample (80 ng in total).

Data Analysis

Minimap2.28[12] was used to map FASTQs to a "hybrid" reference, which was generated by concatenating the ATCC 8482, ATCC 23114, ATCC 6940 genomes and the DCS *lambda* amplicon sequence. The reference sequences of ATCC genomic DNA can be access from their official database[13] (https://github.com/ATCC-Bioinformatics/genome_portal_api). The DCS genomic DNA are available from ONT website https://a.storyblok.com/f/196663/x/f69b1ef376/dcs_reference.txt. The phiX reference genome used here is downloaded from NCBI Reference Sequence: NC_001422.1 (https://www.ncbi.nlm.nih.gov/nuccore/9626372). Samtools1.22.1[14] was used to extract mapping counts from the resulting BAM files. Only demultiplexed reads were considered in calculating metrics (i.e. unclassified reads were excluded). Unmapped reads are included in calculation of precision. R was used for data analysis and figure generation.

Code Availability

All analyses are fully reproducible using code available in the accompanying GitLab repository (https://github.com/QingDAI0225/barcode_crosstalk_data/tree/main), publicly available Apptainer images, and publicly available data (see below).

Data Availability

The data have been deposited with links to BioProject accession number PRJNA1377585 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

References

1. Schnell, I. B., Bohmann, K. & Gilbert, M. T. P. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.* **15**, 1289–1303, DOI: 10.1111/1755-0998.12402 (2015).
2. van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L. & Guschanski, K. Index hopping on the illumina hiseqx platform and its consequences for ancient dna studies. *Mol. Ecol. Resour.* **20**, 1171–1181, DOI: 10.1111/1755-0998.13009 (2019).
3. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded oxford nanopore reads with deep convolutional neural networks. *PLOS Comput. Biol.* **14**, e1006583, DOI: 10.1371/journal.pcbi.1006583 (2018).
4. Khezri, A. *et al.* Mysterymaster: scraping the bottom of the barrel of barcoded oxford nanopore reads. *BMC Bioinforma.* **26**, DOI: 10.1186/s12859-025-06266-2 (2025).
5. MacConaill, L. E. *et al.* Unique, dual-indexed sequencing adapters with umis effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, DOI: 10.1186/s12864-017-4428-5 (2018).
6. Farouni, R., Djambazian, H., Ferri, L. E., Ragoussis, J. & Najafabadi, H. S. Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell rna-sequencing. *Nat. Commun.* **11**, DOI: 10.1038/s41467-020-16522-z (2020).
7. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, DOI: 10.1186/s12915-014-0087-z (2014).
8. Penter, L. *et al.* Integrative genotyping of cancer and immune phenotypes by long-read sequencing. *Nat. Commun.* **15**, DOI: 10.1038/s41467-023-44137-7 (2024).
9. Steimle, A. *et al.* Gut microbial factors predict disease severity in a mouse model of multiple sclerosis. *Nat. Microbiol.* **9**, 2244–2261, DOI: 10.1038/s41564-024-01761-3 (2024).
10. GitHub - nanoporetech/dorado: Oxford Nanopore's Basecaller — github.com. https://github.com/nanoporetech/dorado/. [Accessed 11-12-2025].
11. Portik, D. M., Brown, C. T. & Pierce-Ward, N. T. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinforma.* **23**, DOI: 10.1186/s12859-022-05103-0 (2022).
12. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574, DOI: 10.1093/bioinformatics/btab705 (2021).
13. Nguyen, S. V. *et al.* The atcc genome portal: 3, 938 authenticated microbial reference genomes. *Microbiol. Resour. Announc.* **13**, DOI: 10.1128/mra.01045-23 (2024).
14. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, DOI: 10.1093/gigascience/giab008 (2021). Giab008, https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf.

Acknowledgements

This work was supported by the Engineering Research Centers Program of the National Science Foundation under NSF Cooperative Agreement No. EEC-2133504. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Author contributions statement

QD conceived the experiments, QD conducted the experiments, QD analysed the results, and all authors (QD, CKG, and JAG) reviewed and revised the manuscript.

Additional information

Competing interests

The authors declare no competing interests.