BMC Genomics

RESEARCH

Open Access

Systematic detection of Oxford Nanopore adapter and barcode contamination in GenBank

Florencia Martino[1], Abraham J. Kandathil[1] and Steven J. Clipman[1*]

Abstract

Background Oxford Nanopore Technologies (ONT) workflows attach synthetic adapters and barcode oligonucleotides to DNA fragments for priming and multiplexing. Although these sequences are normally removed during basecalling or post-processing, incomplete trimming can embed non-biological fragments into genome assemblies deposited in public repositories.

Results A systematic sequence-similarity screening of the GenBank nucleotide search database identified 5,782 unique accessions containing 13,734 Oxford Nanopore Technologies (ONT) adapter and barcode fragments embedded in bacterial, viral, and eukaryotic assemblies. Most contaminating sequences correspond to rapid-ligation and ligation-based adapters, reflecting incomplete trimming in ONT workflows. Multiple ONT synthetic sequences were often detected within a single accession, consistent with protocols employing several adapters and barcodes. Compared with GC-matched randomized control sequences, the original ONT sequence catalogue yielded significantly more GenBank matches, supporting that the observed signal exceeds chance similarity.

Conclusions This ONT-specific contamination recurs across GenBank records and exhibits structured patterns across taxonomic groups and project types. Implementing ONT-aware screening during public database submissions and curation is critical to prevent the ongoing propagation of synthetic sequences and safeguard genomic data integrity.

Keywords Long-read sequencing, Bioinformatics, Synthetic oligonucleotides, Quality control, Genome assemblies, Sequencing artifacts, Adapter trimming, Public sequence repositories

Background

Curation of accurate genomic resources requires the complete removal of laboratory-derived artifacts from sequencing reads. Oxford Nanopore Technologies (ONT) workflows attach synthetic adapters and barcode oligonucleotides to DNA fragments to enable priming and multiplexing. These adapters are typically excised during basecalling or post-processing, yet failures in trimming can embed non-biological sequences into genome assemblies. Once deposited, such sequences may be mis-annotated as authentic genomic elements, undermining downstream comparative and functional analyses.

Contamination of sequence repositories by laboratory constructs has been recognized since the Sanger era and remains a persistent problem in high-throughput genomics [1, 2]. Large-scale surveys have documented millions of GenBank entries containing cloning vectors,

*Correspondence:
Steven J. Clipman
sclipman@jhmi.edu
[1]Division of Infectious Diseases, Johns Hopkins University School of Medicine, Baltimore, MD, USA

adapters, or other artifacts [3, 4], including over two million sequences in GenBank and RefSeq alone [3]. Illumina adapter contamination is well documented, with hundreds of adapter fragments reported within single bacterial assemblies [5, 6]. In contrast, contamination from ONT adapters and barcodes—whose chemistry and error profiles differ from short-read platforms—has not been systematically examined. Although tools such as Porechop can identify and trim ONT adapters [7], residual adapter and barcode sequences can remain in public datasets. Here we present one of the first genome-wide surveys of ONT-specific contamination in public nucleotide records, revealing that ONT-derived synthetic sequences recur across bacterial, viral, and eukaryotic assemblies.

Results

BLASTn searches requiring a minimum aligned length of 16 nucleotides, ≥ 95% identity, and ≥ 80% query coverage identified 5,782 unique accessions containing 13,734 distinct adapter/barcode–accession combinations (Supplementary table). These counts should be interpreted as screen-positive accessions identified by a query-driven search, not as a prevalence estimate based on a fixed denominator. Platform metadata explicitly indicated sequencing on ONT instruments for 2,025 of these accessions. Among the 5,583 positive records with assignable type, 77.6% correspond to complete chromosomes, 18.1% to transcript sequences, and 2.6% to plasmids (the remainder are mostly WGS/contigs and other genome fragments), indicating that contamination is observed across finished replicons, small-genome and cDNA-derived entries.

Residual fragments were not evenly distributed across adapter classes and accessions. SQK-NSK007 Y-Bottom adapter was the most frequently detected match, present in 20% of positive accessions, followed by the ligation bottom-strand adapter (19%) and the SQK-NSK007 Y-top adapter fragment (19%). Barcode sequences contributed a smaller overall fraction but were detected across independent projects, including multiplexed datasets. Co-occurrence is common: 81% of affected accessions contain at least two distinct ONT sequences, and 29% contain three or more. The most frequent pair, ligation bottom-strand adapter with SQK-NSK007 Y-Top, appears in more than 1,091 accessions, while the triad of ligation bottom-strand adapter, SQK-NSK007 Y-Top, and SQK-NSK007 Y-Bottom sequences occurs in over 1,019 accessions. These co-occurrence patterns involved combinations of sequences derived from ligation-based and rapid workflow catalogues. Taxonomic distribution was broad across the positive accessions, spanning bacterial, fungal, protozoan, metazoan, and phage-associated records. Within the bacterial subset, Enterobacterales was the most frequent order, and *Escherichia coli* alone was the most frequently represented bacterial species. Additional bacterial genera, including *Shigella*, *Streptomyces*, and *Photobacterium*, were also detected at lower frequencies, while adapter-associated hits were identified across a wide range of eukaryotic taxa, including fungi and protozoan parasites such as *Plasmodium*, *Cryptosporidium parvum*, and *Tritrichomonas foetus*, as well as in metazoan and bacteriophage-associated records, including members of Caudoviricetes.

Some detected matches involved motifs with known homology outside ONT workflows. The rapid ligation adapter-associated (RLA) core shares identity with the canonical right-end recognition site of the MuA transposase, which was also used in legacy fragmentation protocols and helps explain some sporadic pre-ONT hits. For this reason, isolated RLA-only matches were interpreted cautiously. By contrast, co-occurrence of RLA-compatible sequence with additional ONT-specific adapters and barcodes, particularly in records with explicit ONT metadata, was more compatible with an ONT-related origin than isolated RLA-only matches. RLA-compatible sequence was nevertheless common among explicitly ONT-sequenced accessions, and its frequent co-occurrence with Y-Top and ligation bottom-strand adapters in recent submissions supports the conclusion that adapter removal is incomplete in many modern long-read pipelines. A representative case can illustrate these patterns. The circular 2,066-bp *E. coli* plasmid sequenced with ONT (CP163936.1) contains a perfect 50 nt match to the rapid-ligation adapter, together with additional barcode 28, RBK004, and SQK-NSK007 adapter-derived sequences within the same terminal segment of the linearized record. Rapid-adapter, barcode, and sequencing-adapter components were detected within the same short terminal segment. Comparison against the nucleotide database further showed that CP163936.1 shares high sequence identity with numerous closely related small Enterobacterales plasmids across an otherwise continuous backbone, but these alignments consistently extend only through approximately the first 1,954 bp. The remaining terminal segment lay outside this widely shared backbone and contained the ONT-derived matches. Furthermore, homology in this terminal region was detected only in a second *E. coli* plasmid (CP165494.1), from the same BioProject generated by the same group using the same ONT-plus-Illumina workflow.

A similar terminal pattern is also seen in eukaryotic assemblies. In chromosome 3 of the fungus *Rhodotorula toruloides* (AP041615.1), short exact matches to Y-adapter components are present near both the beginning and end of the linearized chromosome record. A comparable terminal configuration is also seen in viral

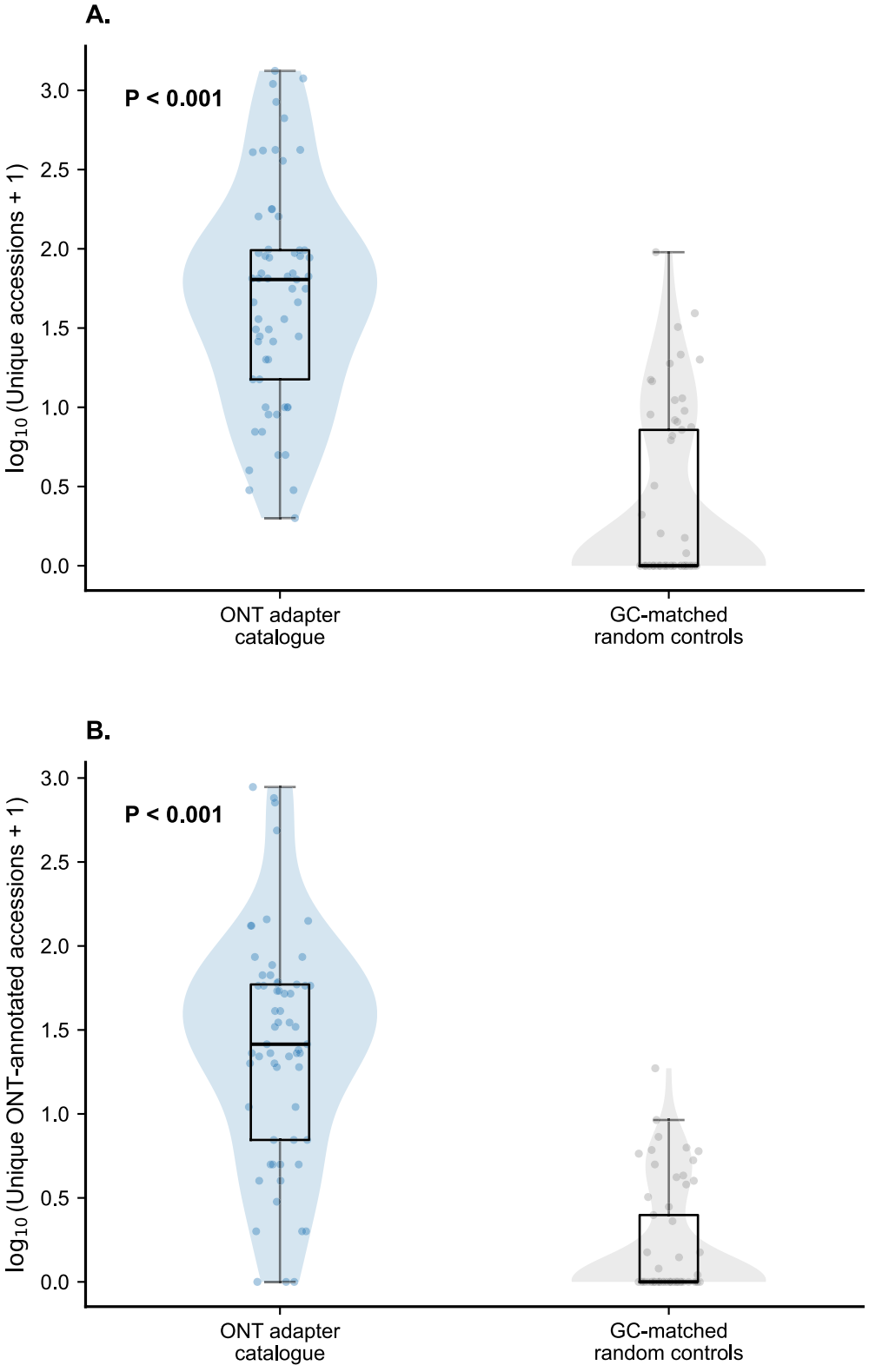

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Original ONT-derived sequences yield more GenBank matches than length- and GC-matched randomized controls. For each original ONT adapter or barcode sequence, 10 simulated control sequences matched for length and GC content were generated and screened using the same BLASTn-short workflow and post-search filters. To reduce redundancy while preserving a representative barcode subset, the control analysis for barcodes was restricted to 1 - 18. Panel (**A**) depicts aggregated per-adapter accession counts across all retained matches. Panel (**B**) depicts aggregated per-adapter accession counts restricted to records with explicit ONT-related metadata. Violin plots show the distribution across adapters, boxplots indicate the median and interquartile range, and points represent individual adapter-level values. Values are shown on a $\log_{10}$ scale after adding 1 to include zero-valued observations. P-values were calculated using one-sided paired Wilcoxon signed-rank tests

assemblies. In the SARS-CoV-2 genome OV192362.1, an exact 24 nt match to the ONT barcode 11 sequence is present within the final 90 bp of the linearized genome, together with an exact 24 nt match to an SQK-NSK007 Y-adapter component in the same terminal region. These ONT-derived motifs were localized near the ends of linearized genome records across bacterial, eukaryotic, and viral assemblies.

Given the short length of many adapter and barcode fragments, some incidental similarity to public records would be expected by chance. To evaluate this directly, we generated 10 simulated control sequences for each original ONT-derived sequence, matching its length and GC content while randomizing nucleotide order. When screened with the same BLASTn-short workflow and post-search filters, the original ONT catalogue consistently produced higher aggregated per-adapter accession counts than the simulated controls, both across all retained accessions and within the subset of records with explicit ONT metadata (both $P < 0.001$, one-sided paired Wilcoxon signed-rank test; Fig. 1). In both comparisons, the distributions for the original ONT catalogue were shifted upward relative to the simulated controls.

Discussion

Our analysis reveals that ONT-derived synthetic sequence in GenBank is recurrent and structured rather than random. The strongest support comes from the frequent co-occurrence of multiple distinct ONT sequence classes within the same accession, including combinations of rapid or ligation adapters, Y-adapter components, and, in multiplexed libraries, barcode sequences. This pattern was more informative than isolated matches alone, particularly for ambiguity-prone motifs such as the RLA-compatible sequence.

Comparison with length- and GC-matched randomized controls further supports this interpretation. Short synthetic fragments can generate incidental local similarity in very large public databases. When each ONT-derived sequence was compared with simulated sequences matched for length and GC content, the original ONT catalogue consistently produced higher adapter-level signal than the controls, both across all retained accessions and within the subset of records with explicit ONT metadata. Observed enrichment above this GC- and length-matched background is therefore compatible with recurrent retention of ONT-derived synthetic sequence in deposited assemblies.

The frequent detection of RLA-compatible sequences, together with their recurrent co-occurrence with other ligation-based adapters and barcodes, indicates that incomplete trimming during rapid and ligation workflows is the primary driver, consistent with previous evaluations of Nanopore library preparation methods that showed distinct adapter profiles and trimming performance [8]. Similar problems have been reported for other long-read platforms, where residual adapter sequences can persist into assemblies despite quality control, including PacBio HiFi reads and generic long-read datasets [9, 10]. These reproducible patterns confirm that the detected potential contamination is structured and predictable by kit chemistry rather than random noise.

ONT-derived matches were detected across bacterial, fungal, protozoan, metazoan, viral, and phage-associated records, suggesting that residual synthetic sequences are present in diverse classes of deposited assemblies. Although the detected fragments are short, matches meeting ≥ 95% identity and ≥ 80% query coverage are sufficiently specific to introduce artificial terminal sequence or misleading similarity into annotation, comparative genomics, primer or probe design, and pathogen surveillance workflows. These findings support the incorporation of ONT-aware screening into both pre-submission quality control and repository curation pipelines.

In several representative cases, these fragments were localized near contig or genome termini and co-occurred with multiple ONT-derived motifs within the same short region, consistent with residual adapter sequence retained during library preparation rather than dispersed or endogenous sequence. Although these fragments are short, matches meeting ≥ 95% identity and ≥ 80% coverage thresholds are sufficiently specific to interfere with annotation pipelines if left untrimmed. If retained in deposited assemblies, such fragments may propagate through public reference databases and introduce artificial terminal sequences or misleading similarity signals into comparative genomics, primer or probe design, and pathogen surveillance workflows.

These effects may be especially relevant for pathogen genomics, where public reference sequences are routinely reused for comparative analysis, genomic surveillance, and, in some settings, assay or primer design. In preparing data for publication, authors should screen both

reads and assemblies with ONT-aware tools. Repository curators should integrate the same catalogue into UniVec [11] and related contamination pipelines. Incorporating the full ONT adapters and barcodes catalogue into automated screening tools (e.g., NCBI FCS-adaptor [12]) and routine submission workflows can prevent synthetic sequences from propagating through public databases and distorting downstream analyses. Vigilant adapter detection and removal by both data submitters and repository curators is essential to preserve the integrity of long-read genomic resources and to prevent the propagation of synthetic sequences into public databases, which could lead to downstream misinterpretation in comparative and functional genomics analyses.

Conclusions

Residual Oxford Nanopore adapter and barcode sequences recur in GenBank records, spanning all major domains of life, and are most consistent with incomplete trimming during rapid and ligation workflows. Their recurrent co-occurrence within individual accessions, together with explicit ONT platform metadata in a substantial subset of records, supports that contamination is systematic rather than stochastic. Incorporating ONT-specific screening into submission and curation pipelines would reduce the continued propagation of synthetic sequences and improve the integrity of public genomic resources.

Methods

Adapter and barcode catalogue

We compiled a non-redundant set of 229 ONT adapter and barcode sequences from ONT chemistry documentation [13] and Porechop v0.2.4 trimming references [7]. Duplicate sequences were removed to generate a single non-redundant catalogue. Historical motifs, including the 24-nt ligation leader hairpin (AAGCAGTGGTATC AACGCAGAGTAC) identical to the Clontech SMARTer II template-switching primer, were retained to capture legacy cDNA protocols [14]. Sequences corresponding to 16S rRNA amplicon primers were retained in the catalogue for completeness but were excluded from all downstream results because they are not specific to ONT library preparation and would generate non-specific matches.

Database screening

Adapter and barcode sequences from the curated catalogue were queried against the NCBI nucleotide database using BLASTn (blastn-short; dust and soft-masking disabled), screening the full GenBank nucleotide search space available at the time of analysis and retaining hits with ≥ 16 nt aligned length, ≥ 95% nucleotide identity, and ≥ 80% query coverage. Metadata, including definition line, taxonomy, molecule type, topology, length, and last update, were retrieved through Entrez, with flatfile parsing when necessary. GenBank comments were scanned for platform indicators ("Oxford Nanopore," "MinION," "PromethION," "GridION," "Flongle") to identify ONT-sequenced records. Duplicate accessions were collapsed by unique version to avoid double-counting. The release year was retrieved during Entrez annotation and used in downstream interpretation to focus on records released from 2014 onward, the period beginning with the first publicly available ONT data. Final interpretation placed greater weight on recurrent co-occurrence of multiple ONT-derived sequences within the same accession and on explicit ONT platform metadata, whereas isolated ambiguity-prone motifs were interpreted more cautiously.

Length- and GC-matched randomized controls

Randomized length- and GC-matched control sequences were generated for each adapter and screened with the same BLASTn-short workflow and post-search filtering criteria used for the original ONT-derived queries (≥ 16 nt alignment length, ≥ 95% identity, and ≥ 80% query coverage), creating 10 replicates per sequence. Barcode-derived adapters 19–96 were excluded because they belong to an extended barcode set with high sequence similarity and limited usage in standard ONT workflows. Analysis was restricted to barcodes 1 - 18 as a representative subset to avoid redundancy and maintain computational tractability. Mean, standard deviation, median, and maximum unique accession counts were calculated across randomized replicates for each adapter. Paired one-sided Wilcoxon signed-rank tests were used to compare original adapters against length- and GC-matched controls for all retained accessions and for the subset of records explicitly annotated as ONT-derived. Searches were completed on March 21, 2026. All code, the full sequence catalogue, and results are available at https://github.com/MERIDAIN-Lab/ONT_NCBI_contamination.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12805-9.

Supplementary Material 1.

Authors' contributions

F.M. and S.J.C. jointly conceived and designed the study and interpreted the results. F.M. compiled and curated the Oxford Nanopore adapter catalogue, developed and executed the computational analyses, and drafted the manuscript. S.J.C. contributed to study design, data interpretation, funding acquisition, and manuscript writing, and provided overall supervision and critical revision. A.K. contributed through mentorship/advisory support and funding acquisition. All authors approved the final version of the manuscript and are jointly accountable for its accuracy and integrity.

Funding
This work was supported by the National Institutes of Health (R01DA058567 and DP2DA056130).

Data availability
The dataset(s) supporting the conclusions of this article are available in the Zenodo repository, DOI: https://doi.org/10.5281/zenodo.17409927. The project repository (code and catalogue) is available at https://github.com/MERIDAIN-Lab/ONT_NCBI_contamination.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 January 2026 / Accepted: 26 March 2026
Published online: 06 April 2026

References

1. Yoshikawa T, Sanders AR, Detera-Wadleigh SD. Contamination of sequence databases with adaptor sequences. Am J Hum Genet. 1997;60(2):463–6.
2. NCBI. National Center for Biotechnology Information, VecScreen contamination screening. 2025. Available from: https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/.
3. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. 2020;21(1):115.
4. Schäffer AA, Nawrocki EP, Choi Y, Kitts PA, Karsch-Mizrachi I, McVeigh R. VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. Bioinformatics. 2018;34(5):755–9.
5. Chrisman B, He C, Jung JY, Stockham N, Paskov K, Washington P, et al. The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families. Sci Rep. 2022;12(1):9863.
6. Moeller AH, Dillard BA, Goldman SL, Real MVF, Sprockett DD. Removal of sequencing adapter contamination improves microbial genome databases. BMC Genomics. 2024;25(1):1033.
7. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Bioinform Adv. 2023;3(1):vbac085.
8. Sauvage T, Cormier A, Delphine P. A comparison of Oxford nanopore library strategies for bacterial genomics. BMC Genomics. 2023;24(1):627.
9. Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022;23(1):157.
10. Ranjan P, Brown CA, Erb-Downward JR, Dickson RP. SNIKT: sequence-independent adapter identification and removal in long-read shotgun sequencing data. Bioinformatics. 2022;38(15):3830–2.
11. UniVec Database [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2025. Available from: https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.
12. Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biol. 2024;25(1):60.
13. Technologies ON. Chemistry Technical Document. 2024. Available from: https://nanoporetech.com/document/chemistry-technical-document#barcode-sequences.
14. Clontech Laboratories I. SMARTer® Ultra Low RNA Kit for Illumina® Sequencing User Manual: Takara Bio USA; 2014. Available from: https://www.mediray.co.nz/media/16233/br_clontech_rna_smarter_ultralow_tn_1114_633066_in.pdf.

Publisher's Note