Genomic Language Model Mitigates Chimera Artifacts in Nanopore Direct RNA Sequencing

Yangyang Li[1†], Ting-You Wang[1†], Qingxiang Guo[1], Yanan Ren[1], Xiaotong Lu[1], Qi Cao[1,2], Rendong Yang[1,2*]

[1]Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, 60611, IL, USA.
[2]Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, 675 N St Clair St, Chicago, 60611, IL, USA.

*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu;
Contributing authors: yangyang.li@northwestern.edu; tywang@northwestern.edu; qingxiang.guo@northwestern.edu; ynren1020@gmail.com; xiaotong.lu@northwestern.edu; qi.cao@northwestern.edu;
†These authors contributed equally to this work.

Abstract

Chimera artifacts in nanopore direct RNA sequencing (dRNA-seq) introduce substantial inaccuracies, complicating downstream applications such as transcript annotation and gene fusion detection. Current basecalling models are unable to detect or mitigate these artifacts, limiting the reliability and utility of dRNA-seq for transcriptomics research. To address this challenge, we present DeepChopper, a genomic language model specifically designed to identify and remove adapter sequences from base-called dRNA-seq long reads with single-base precision. Operating independently of raw signal or alignment information, DeepChopper effectively eliminates chimeric read artifacts, significantly enhancing the accuracy of crucial downstream analyses. This improvement in reliability unlocks the full potential of nanopore dRNA-seq, establishing it as a more robust tool for diverse transcriptomics applications.

Introduction

Long-read RNA sequencing technologies are revolutionizing transcriptomic research by providing unparalleled resolution for detecting complex splicing and gene fusion events often missed by conventional short-read RNA-seq methods. Among these technologies, Oxford Nanopore Technologies (ONT) dRNA-seq stands out by sequencing full-length RNA molecules directly, preserving native RNA modifications and allowing a more accurate and comprehensive analysis of RNA biology. This approach bypasses the inherent limitations of cDNA-based sequencing methods, such as artifacts arising from reverse transcription, template switching, and Polymerase Chain Reaction (PCR) amplification [1, 2].

Despite these advantages, a critical question remains: Does ONT dRNA-seq introduce technical artifacts? A previous study has suggested that dRNA-seq might generate chimera artifacts, leading to multi-mapped reads [3]. These artifacts may result from ligation during library preparation or chimeric reads produced by software missing the open pore signal, potentially confounding downstream analyses such as transcriptome assembly, quantification, and detection of alternative splicing and gene fusion events. Detecting these chimera artifacts is challenging because long-read aligners often produce chimeric alignments from such artifacts that are indistinguishable from those derived from true gene fusion events. Importantly, chimeric read artifacts frequently contain internal adapter sequences [3], which could theoretically serve as a distinguishing feature to differentiate them from true gene fusion-derived reads. However, ONT dRNA-seq basecallers, trained in RNA, struggle to properly call these DNA-based adapter sequences under an RNA model [4]. As a result, current adapter detection tools [5–7] cannot exploit this feature to eliminate chimeric read artifacts, leaving the issue unresolved (Extended Data Table 1).

To address these challenges, we developed DeepChopper, a Genomic Language Model (GLM) for long-read sequence analysis. Leveraging recent advances in Large Language Model (LLM) that can interpret complex genetic patterns [8], DeepChopper processes long genomic contexts with single-nucleotide resolution. This capability enables precise identification of ONT adapter sequences within base-called long reads, facilitating the detection and removal of chimeric read artifacts in dRNA-seq data. Through analysis of both existing and newly generated dRNA-seq data, including those using the most recent RNA004 chemistry, we uncovered the prevalence of chimera artifacts—a critical issue previously overlooked in the long-read sequencing field. We demonstrated that these artifacts significantly impact transcriptomic analysis by complicating gene fusion detection, transcript annotation, and alternative splicing studies. By both identifying and addressing this problem, our work enhances the reliability and precision of dRNA-seq data, substantially improving its utility in transcriptomic research.

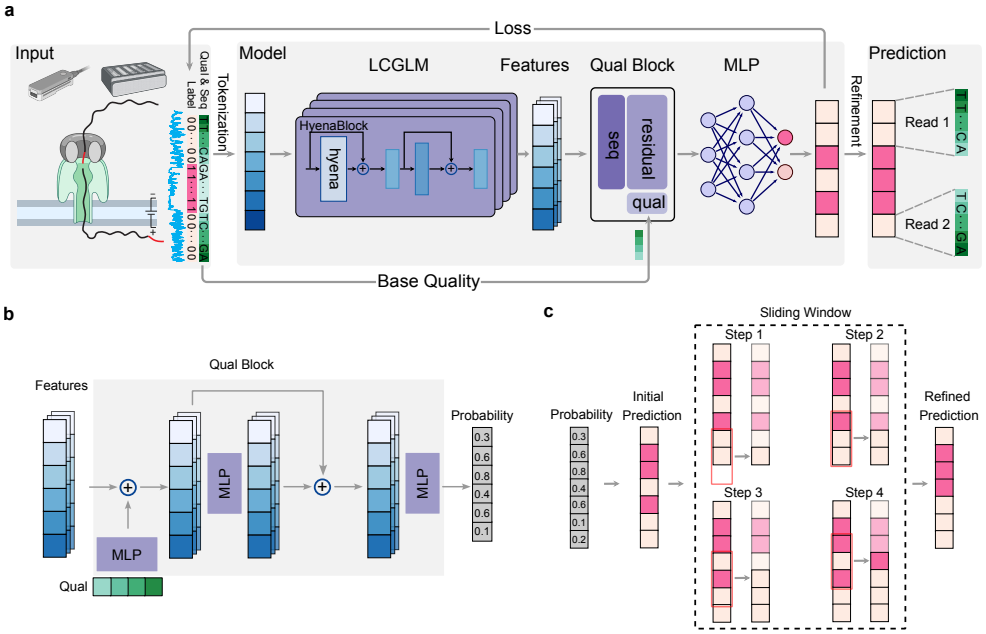


Fig. 1 DeepChopper architecture and methodology. (a) Overview of the DeepChopper model. Raw sequences are first tokenized into vectors and processed by HyenaDNA to generate embedding features. These features are integrated with base quality information in the quality block to produce per-token probability scores. A refinement strategy further optimizes the predictions. Created with BioRender.com. (b) Architecture of the quality block. The block combines a multilayer perceptron (MLP) (purple) with a residual connection to process both embedding features and sequence base quality scores (green vector). The output provides per-token probabilities indicating whether each base belongs to an adapter sequence. (c) Illustration of the sliding window refinement method. The model's initial predictions are inferred from probability. Then the predictions are processed using a sliding window approach (red rectangle) to refine predictions. The dashed rectangle highlights the first four steps of this refinement process, where each step refines the prediction for a single base position in terms of the majority vote.

Results and Discussion

DeepChopper Architecture and Training

DeepChopper leverages the long-context genomic language model (LCGLM) HyenaDNA [9], which excels at capturing long-range dependencies (Fig. 1a). process sequencing base quality information, DeepChopper extends its framework by incorporating a dedicated quality block, which is a neural network comprising multiple MLPs with residual connections [10] (Fig. 1b, See Methods for details). This addition enables the effective utilization of sequencing base quality, a crucial feature for improving prediction accuracy. By combining broad contextual understanding with nucleotide-level precision, this hybrid architecture allows DeepChopper to accurately identify and process adapters sequences. Reads containing internal adapters are split into multiple records, with 3′ end adapters simultaneously trimmed (Fig. 1a), thereby preserving authentic biological sequences and eliminating chimera artifacts.

To further refine the prediction accuracy, DeepChopper implements a post-processing stage using a sliding window and majority vote approach, as illustrated in Fig. 1c. The model applies sliding window with a stride of 1 across the read, analyzing the distribution of predicted adapter positions within each window (See Methods for details). This refinement process ensures that the predicted adapter sequences are contiguous and biologically plausible by smoothing irregularities and eliminating isolated false positives, while preserving genuine adapter signals that demonstrate consistent patterns across adjacent windows. By applying this iterative refinement strategy, DeepChopper is able to optimize the final predictions and improve the overall accuracy of the adapter detection.

Comparing to existing general-purpose GLM, DeepChopper is specifically optimized for long-read sequence analysis at single-nucleotide resolution. This fine-grained resolution provides a critical advantage for genomic analysis tasks requiring precise base-level predictions. While DNABERT [11] is limited to input sequences of approximately 512 bp, DNABERT2 [12] to 10,000 bp, and Nucleotide Transformer [13] to 6,000 bp, DeepChopper supports input lengths up to 32 kilobases—sufficient to encompass most complete mRNA transcripts. This 32K nucleotide input limit was selected based on both technical and biological considerations. Technically, the constraint reflects the architectural design and context window limitations of the underlying HyenaDNA model. Biologically, the vast majority of human mRNA transcripts fall well within this range, with a median length of approximately 1.5–2 kb, and the majority well below the maximum input threshold [14]. Moreover, Nanopore sequencing studies report that the typical maximum aligned length of native RNA reads is around 21 kb [15], further supporting the adequacy of this design.

The combination of extended input capacity and single-nucleotide tokenization enables DeepChopper to accurately identify non-reference elements, such as ONT adapter sequences, with base-pair precision—an essential capability for detecting chimera artifacts in dRNA-seq data. In addition, DeepChopper's lightweight architecture, consisting of only 4.6 million parameters, makes it computationally efficient and scalable for large-scale dRNA-seq analysis. This is in contrast to models like Evo [16], which require billions of parameters and significantly more computational resources.

To train DeepChopper for identifying adapter sequences within dRNA-seq long reads, we utilized data from six human cell lines: HEK293T, A549, HCT116, HepG2, K562 and MCF-7 provided by the Singapore Nanopore Expression Project (SG-NEx) [17]. We curated a training set of 480,000 long reads and a validation set of 60,000 ones initially deemed free of adapters and inserted putative adapter sequences, derived from the raw dRNA-seq data, into these reads to create instances containing internal and $3'$ end adapters (See Methods for details). An independent test set comprising 60,000 long reads was held out for performance evaluation.

DeepChopper Benchmarking and Model Optimization

We conducted comprehensive benchmarking of DeepChopper against existing ONT adapter trimming tools including Pychopper [5], Porechop [6], and Porechop_ABI [7], though it should be noted that none of these existing tools were specifically designed for dRNA-seq data analysis. Performance evaluation was carried out using the synthetic

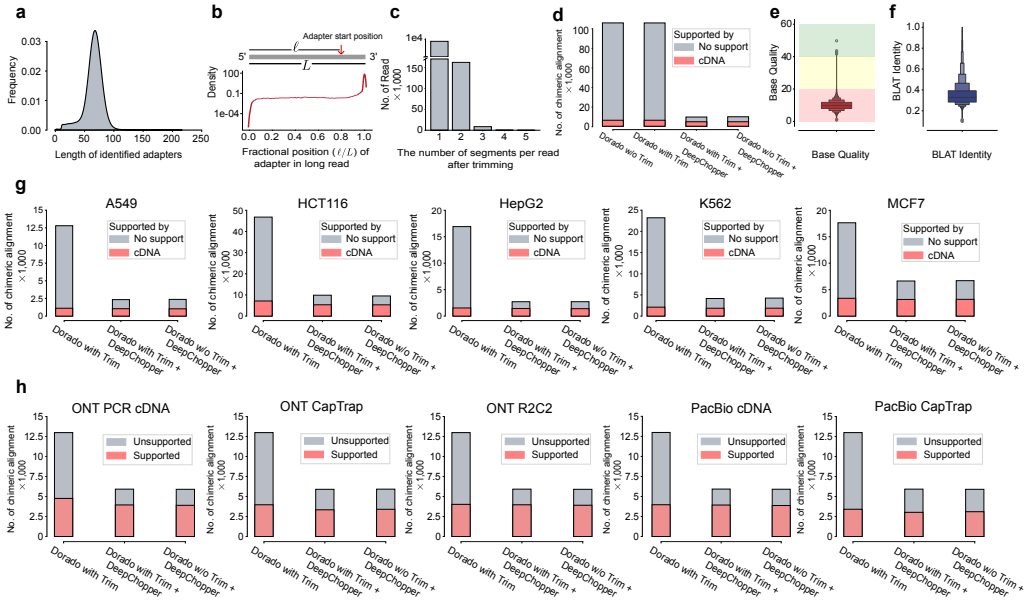


Fig. 2 Detection of chimeric read artifacts in dRNA-seq data using DeepChopper and validation with orthogonal sequencing platforms (a) Length distribution of predicted adapters by DeepChopper in VCaP dRNA-seq data. (b) Distribution of relative adapter position along read length in VCaP dRNA-seq data. Grey rectangle represents a long read from 5′ to 3′. Relative position is calculated as the ratio of the length before DeepChopper predicted adapter start position to the total read length. (c) Distribution of segments per read after trimming: 1 segment indicates 3′ end adapter trimmed, while 2 or more indicate internal adapters trimmed. (d) Chimeric alignments (in thousands) for VCaP dRNA-seq reads processed by Dorado with and without adapter trimming, Dorado with adapter trimming followed by DeepChopper, and DeepChopper. DeepChopper-involved methods greatly reduce chimeric alignments not supported by direct cDNA sequencing. (e) Distribution of base qualities from identified internal adapters by DeepChopper. Background colors indicate quality levels: green (high), yellow (medium), and red (low). (f) BLAST-like alignment tool (BLAT) identity distribution of the internal adapter sequences mapping against human reference genome. (g) The number of chimeric alignments (in thousands) for A549, HCT116, HepG2, K562, and MCF7 cell lines processed by Dorado with adapter trimming, Dorado with adapter trimming followed by DeepChopper, and DeepChopper. DeepChopper-involved methods consistently reduce chimeric alignments not supported by cDNA sequencing across all cell lines. (h) Chimeric alignments from dRNA-seq of the WTC11 cell line, evaluated for support using additional ONT and Pacific Biosciences (PacBio) sequencing data with different protocols. DeepChopper-involved methods reduce unsupported chimseric alignments across all methods compared to Dorado with adapter trimming.

testing dataset (n = 60,000 reads), enabling rigorous assessment of precision, recall, and F1-score metrics. As shown in Extended Data Fig. 1, all existing tools demonstrated negligible performance metrics when processing dRNA-seq adapter sequences, indicating fundamental incompatibility with the dRNA-seq protocol. In contrast, DeepChopper achieved exceptional accuracy in identifying both terminal and internal adapters, with recall, precision, and F1 scores consistently exceeding 0.99. These results highlight DeepChopper's unique capability to address the specific complexities inherent to dRNA-seq reads and underscore the critical need for purpose-built solutions in this domain.

To further evaluate DeepChopper's classification accuracy, we assessed its performance at single-nucleotide resolution. As shown in Extended Data Table 2, the model exhibited sharply bimodal probability distributions for base-level classification, stratified by ground truth labels. For true adapter bases (Extended Data Fig. 2a), the model consistently assigned high probabilities (> 0.8) to the adapter class while suppressing probabilities for the non-adapter class (< 0.2). Conversely, for true non-adapter bases (Extended Data Fig. 2b), the model reliably predicted high probabilities for the non-adapter class and low probabilities for the adapter class. The minimal presence of intermediate probability values suggests that DeepChopper makes highly confident predictions with low ambiguity between classes. This decisive classification behavior reflects the model's robust feature representation, well-calibrated decision boundaries, and strong generalization to previously unseen data.

We further conducted an ablation experiment to assess the contribution of the quality block component in the model architecture. As shown in Extended Data Table 2, inclusion of the quality block led to a marked improvement in performance, with the F1 score increasing from 0.97 to 0.99. This module leverages per-base quality scores to better distinguish genuine adapter sequences from similar motifs that may occur naturally within reads. The enhanced performance suggests that incorporating sequencing quality information enables the model to more effectively filter out spurious signal and improve classification robustness. This experiment was performed using the same training methodology as previous evaluations but with a newly generated, focused dataset of 100,000 reads (See Methods for details).

Chimeric Read Artifact Detection in Cancer dRNA-seq Data

To assess DeepChopper's ability to detect chimera artifacts in real data, we generated an independent dRNA-seq dataset using the prostate cancer VCaP cell line, which was excluded from model training. This dataset provides a robust framework for evaluating chimera artifacts in genuine dRNA-seq samples, ensuring that DeepChopper's performance generalizes beyond the training data. We conducted dRNA-seq of VCaP cells using ONT's SQK-RNA002 chemistry, consistent with that used in the SG-NEx project. Using a MinION sequencer with four R9.4 flow cells, we generated 9,177,639 long reads in FASTQ format, with base-calling performed using SG-NEx's Dorado software [18].

DeepChopper processes input data through a three-stage pipeline: (1) FASTQ-to-Parquet conversion for efficient input/output, (2) adapter prediction using a neural network, and (3) post-processing to trim and segment reads based on predicted adapter positions. To improve runtime, we implemented core functions in Rust, enabled GPU-based inference, and parallelized key components across the pipeline.

We benchmarked DeepChopper's computational performance using read subsamples ranging from 0.1M to 9M reads. As shown in Extended Data Fig. 3a, total runtime scaled linearly with input size, with the full dataset requiring 5 hours to process. Adapter prediction was the most time-consuming step, especially at scale. Memory usage also increased with input size, peaking at 70 GB for CPU-based inference, while GPU-based execution maintained lower memory demands (Extended Data Fig. 3b). FASTQ conversion and post-processing remained efficient and lightweight across all

input sizes, demonstrating DeepChopper's scalability for high-throughput dRNA-seq analysis (See Methods for details).

Applying DeepChopper to the full VCaP dataset increased usable read yield by 3%, resulting in 9,357,913 adapter-trimmed reads. It identified 8,218,172 adapter sequences across 7,990,102 reads (87% of total), most measuring 70 bp—consistent with the expected length of the RMX adapter used in ONT's SQK-RNA002 dRNA-seq kit (Fig. 2a) [19]. Analysis of adapter locations revealed that 7,777,624 reads had adapters at the 3′ end, while 212,478 contained internal adapters (Fig. 2b), indicating that chimeric artifacts are common in VCaP dRNA-seq data.

Further examination showed that chimera artifacts could arise from the joining of multiple long reads, with the most frequent pattern involving two reads joined by a single internal adapter (Fig. 2c). To validate these findings, we analyzed minimap2 [20] chimeric alignments and compared them to a matched VCaP direct cDNA-seq dataset, which we generated as part of this study. Chimeric reads fully supported by cDNA sequencing were considered bona fide events (See Methods for details). Notably, we also evaluated whether ONT's Dorado basecaller trimming feature could mitigate these artifacts. However, we found that Dorado alone—regardless of trimming—was insufficient to eliminate spurious chimeric alignments. In contrast, DeepChopper reduced unsupported chimeric alignments by 91% and increased the fraction of cDNA-supported chimeric events from 5% to 47%, whether applied before or after Dorado trimming (Fig. 2d). These results underscore DeepChopper's ability to distinguish true biological chimeras from technical artifacts.

To further verify the artifactual nature of internal adapters, we analyzed their base quality scores and aligned them to the human reference genome using BLAT [21]. Adapter regions identified within chimera artifacts exhibited significantly lower base quality (Fig. 2e) and poor sequence identity to the reference genome (Fig. 2f), supporting their non-human and non-biological origin.

Finally, we evaluated post-processing performance as a function of the sliding window size used for segmentation. Using a 1M-read subsample, we tested window sizes of 11, 21, 31, 41, and 51 bp. Smaller windows yielded slightly higher cDNA support percentages (47.5% for 11 bp vs. 43.3% for 51 bp; Extended Data Fig. 4a), but increased fragmentation of reads into 4+ segments (Extended Data Fig. 4b). A 21 bp window provided the optimal balance, maintaining high support while minimizing over-segmentation. Based on these results, 21 bp was selected as the default setting, and DeepChopper allows users to adjust this parameter for dataset-specific optimization (See Methods for details).

Multi-sample Validation Across Platforms and Species

To further evaluate DeepChopper's performance beyond the VCaP dataset, we performed multi-sample validation across diverse biological systems and sequencing platforms. We began by analyzing dRNA-seq data from the SG-NEx project, comparing chimeric alignments before and after DeepChopper trimming. Across these samples, DeepChopper reduced unsupported chimeric alignments by 62% to 84%, while preserving cDNA-supported chimeric alignments without noticeable reduction (Fig. 2g). These results reinforce the widespread presence of chimera artifacts in

dRNA-seq and the effectiveness of DeepChopper in selectively removing them without compromising true biological signals.

We next applied DeepChopper to the human WTC11 induced pluripotent stem cell line using data from the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) [22]. This dataset includes cDNA-based long-read sequencing generated with multiple protocols (PCR-cDNA, CapTrap, R2C2) across ONT and PacBio platforms, providing a robust benchmarking resource. DeepChopper selectively eliminated only those chimeric alignments not supported by any cDNA-based method (Fig. 2h), further demonstrating its precision in distinguishing genuine chimeras from technical artifacts.

To assess cross-species generalizability, we extended the analysis to the F121-9 mouse embryonic stem cell line, also from the LRGASP dataset. DeepChopper reliably removed artifactual chimeric reads not supported by any orthogonal cDNA-based sequencing platform (Extended Data Fig. 5), confirming its applicability to both human and non-human transcriptomes.

Importantly, across all datasets, DeepChopper consistently outperformed ONT's Dorado adapter trimming alone—even when applied as a post-processing step—underscoring its distinct and additive utility in chimera artifact correction.

Chimera Artifact Analysis in RNA004 Chemistry

Recently, ONT released a new SQK-RNA004 chemistry for dRNA-seq, but it remains unclear whether chimera artifacts persist with this update. To investigate, we generated new data from the VCaP cell line using this updated chemistry. Although DeepChopper was trained exclusively on RNA002 adapter patterns, we applied it in a zero-shot setting to assess its performance on RNA004 reads.

DeepChopper reduced chimeric alignments by 21% compared to Dorado base-called and adapter-trimmed reads, increasing the proportion of cDNA-supported chimeric alignments from 25% to 30% (Extended Data Fig. 6a). Similar results were observed when DeepChopper was applied after Dorado's adapter trimming, demonstrating compatibility with standard preprocessing pipelines. Additionally, internal adapter-like sequences identified by DeepChopper exhibited low base quality scores (mean Q-score: 7.8) and poor alignment identity to the human genome (mean BLAT identity: 0.38), supporting their classification as artifacts (Extended Data Fig. 6b).

While DeepChopper's reduction of chimeric alignments in RNA004 (21%) is lower than in RNA002 (91%), this difference is expected. It reflects both the model's training on RNA002-specific patterns and inherent improvements in RNA004 chemistry, which was designed to reduce artifact formation [23]. Nonetheless, the model's ability to generalize across sequencing chemistries without retraining highlights its robustness and practical utility for emerging platforms.

Impact on Downstream Transcriptome Analysis

To investigate factors contributing to chimera artifact formation, we examined gene expression levels and transcript lengths associated with chimeric read artifacts. Genes involved in these artifacts showed significantly higher expression than the general

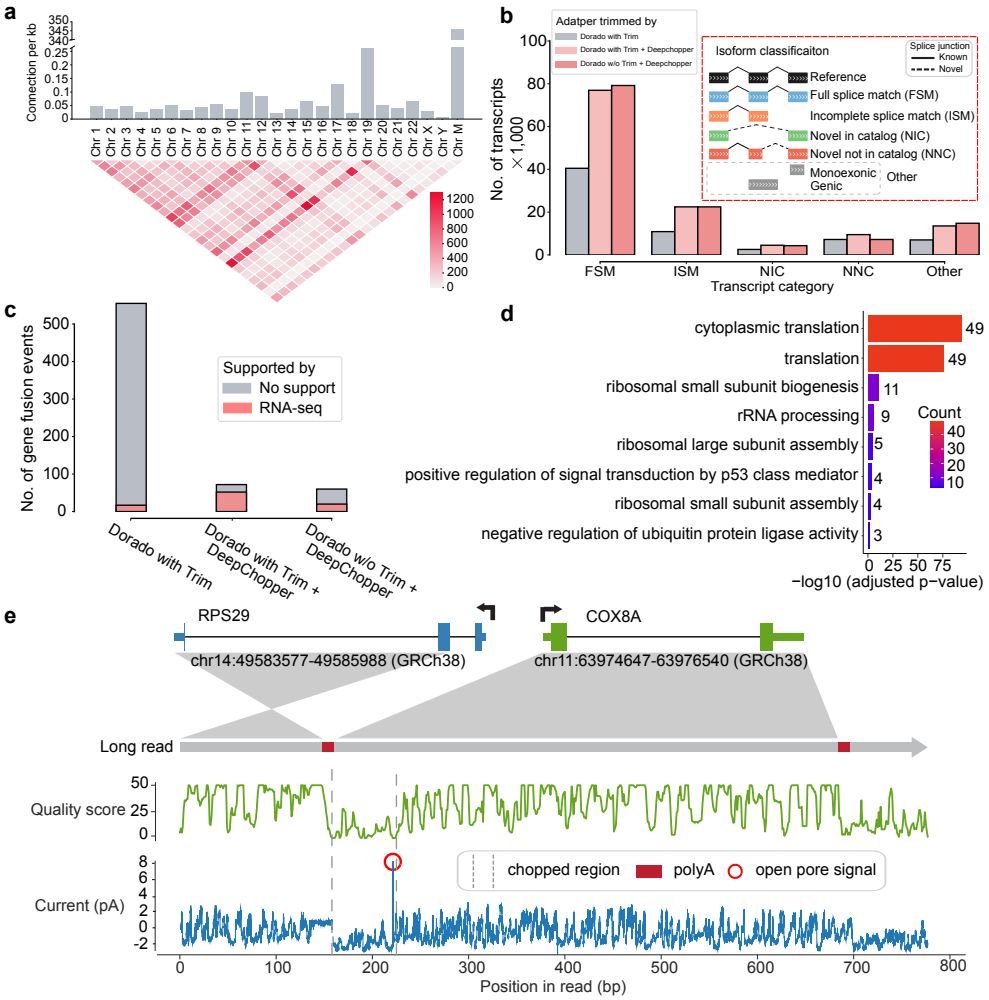


Fig. 3 Characterization of dRNA-seq chimera artifacts and their impact on downstream analysis in VCaP cells (a) The upper bar plot shows the number of chimeric connections per kilobase across chromosomes, highlighting higher chimeric activity in Chr 19 and Chr M. The lower heatmap visualizes interchromosomal connections, with intensity indicating the count of connections between different chromosomes. (b) The bar plot shows the number of transcripts (in thousands) across different isoform classification categories. DeepChopper-processed reads result in a higher number of transcripts compared to Dorado-trimmed reads. The inset details the isoform classification scheme. (c) Detected gene fusions from Dorado adapter-trimmed reads and DeepChopper-processed reads. Gene fusions identified from short-read RNA-seq were used to validate fusion events detected from dRNA-seq. (d) Gene Ontology (GO) enrichment analysis of chimera artifact-affected genes, with color indicating gene count per term. (e) Analysis of a chimeric read artifact detected as an RPS29-COX8A fusion. The schematic shows the fusion between RPS29 (Chr 14) and COX8A (Chr 11). The green plot indicates quality scores along the read, and the blue plot shows raw signal intensity (in pA). The chopped region identified by DeepChopper corresponds to a low-quality segment with low current intensity, polyA, and short open pore signals, suggesting the presence of an ONT adapter.

transcriptome (p-value $< 2.2 \times 10^{-16}$; Extended Data Fig. 7a), while exhibiting a

similar gene length distribution (Extended Data Fig. 7b). Analysis of chimeric junctions across the genome revealed uneven distribution among chromosomes, with the mitochondrial chromosome (Chr M) showing the highest frequency of chimeric connections per base pair—suggesting a potential hotspot for artifact formation (Fig. 3a). This pattern persisted in RNA004 dRNA-seq data (Extended Data Fig. 7c), indicating that chimera artifacts remain a fundamental limitation of dRNA-seq, regardless of chemistry improvements.

We next assessed how DeepChopper correction influences downstream transcriptome analyses. Using IsoQuant [24] to annotate transcripts from VCaP dRNA-seq data, we found that DeepChopper nearly doubled the number of identified transcripts compared to uncorrected data (Fig. 3b). Similar results were observed with RNA004 data (Extended Data Fig. 7d) and when DeepChopper was applied after Dorado's adapter trimming. The largest gains were observed in full-length transcripts (Full splice match (FSM) category), with additional increases in alternatively spliced isoforms (Incomplete splice match (ISM), Novel in catalog (NIC), and Novel not in catalog (NNC) categories). These findings underscore the effectiveness of DeepChopper in mitigating the detrimental effects of chimera artifacts on transcript annotation.

To further assess the implications of artifact removal, we examined gene fusion detection. DeepChopper-corrected reads yielded an 89% reduction in gene fusion calls by FusionSeeker [25] compared to Dorado-trimmed data. Importantly, these reduced fusion calls were not supported by fusions detected in matched short-read RNA-seq data using Arriba [26] (Fig. 3c), suggesting they were false positives. Applying DeepChopper after Dorado trimming yielded consistent results, reinforcing its utility regardless of prior processing steps.

Closer inspection of the filtered gene fusion calls revealed a strong enrichment for ribosomal protein genes (Extended Data Fig. 8a). GO enrichment analyses in VCaP (Fig. 3d) and SG-NEx cell lines (Extended Data Fig. 8b) confirmed this trend, with ribosomal genes frequently appearing in artifact-associated fusions. This enrichment extended to chimera artifacts in RNA004 data as well (Extended Data Fig. 8b). A manual review of a chimeric read identified as an *RPS29–COX8A* fusion revealed that the DeepChopper-processed region—interpreted as an internal adapter sequence—aligned with low-intensity raw current signals, consistent with ONT adapter characteristics (Fig. 3e). The presence of polyA and open pore signals at the boundary of this region further supported an artifact origin rather than a bona fide fusion event.

In conclusion, DeepChopper significantly enhances the integrity of nanopore dRNA-seq data by accurately identifying and removing chimera artifacts that undermine transcript annotation and gene fusion detection. These improvements are robust across different sequencing chemistries and preprocessing pipelines. By addressing a previously underrecognized source of error, DeepChopper exemplifies the power of language model-based approaches in long-read sequencing analysis. Looking ahead, this framework could be extended to detect additional artifact types or applied to new challenges in long-read data interpretation, further advancing the utility of long-read technologies across diverse research domains.

Methods

Cell culture

This is a test. VCaP cell line was obtained from the American Type Culture Collection (ATCC) and cultured under sterile conditions to maintain optimal growth and viability. The cells were grown in Dulbecco's Modified Eagle Medium (DMEM, high glucose; Gibco, Cat# 11-965-092) supplemented with 10% fetal bovine serum (FBS Opti-Gold, Performance Enhanced, US Origin; Gendepot, Cat# F0900-050) to provide essential growth factors. In addition, the culture medium was enriched with 5 mL of 100 mM Sodium Pyruvate (Gendepot, Cat# CA017-010) to support cellular metabolism and 5 mL of Antibiotics-Antimycotics (100×) (Gendepot, Cat# CA002-010) to prevent microbial contamination. Cells were cultured in 100 mm cell culture treated dishes (Thermo Fisher Scientific, Cat# 12-556-002) and incubated at 37 °C in a humidified atmosphere containing 5% CO2, with media changes performed every 72 hours to ensure nutrient availability and waste removal. Cell confluency was regularly monitored and subculturing was performed before reaching 80% confluency to maintain healthy growth conditions and prevent over-confluence stress.

RNA extraction and quantification

Total RNA was extracted using the RNeasy Mini Kit (Qiagen, Cat# 74104) according to the protocol of the manufacturer. The quality and concentration of RNA were assessed using an Agilent 2100 Bioanalyzer. Poly(A)+ RNA was then enriched from total RNA using the Dynabeads^{tm} mRNA Purification Kit (Invitrogen, Cat# 65001), which utilizes oligo (dT) beads for selective mRNA binding. The mRNA was quantified using a Qubit 4 fluorometer and a Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Cat# Q32852). The mRNA preparations were either immediately used to prepare a sequencing library or frozen and stored at −80 °C until further use.

Nanopore sequencing

We performed nanopore dRNA-seq sequencing of the enriched mRNA using two different sets: the RNA002 kits with R9.4.1 flow cells and the RNA004 kits with R10.4.1 flow cells. The decision to incorporate the RNA004 kit, a newly released option, was driven by our intention to test its capabilities in conjunction with our DeepChopper tool to optimize data quality and sequencing efficiency. For the RNA002 library, 1 µg of poly(A)+ RNA was used as input for library preparation using the Direct RNA Sequencing Kit (SQK-RNA002, ONT) following the manufacturer's instructions. Nanopore dRNA-seq employs a reverse transcriptase adapter (RTA) that typically binds to the poly(A) tails of messenger RNA (mRNA); subsequently, a sequencing adapter is ligated to the RTA, which guides the mRNA through the nanopore for sequencing. The prepared library was loaded onto four MinION R9.4 flow cells (FLO-MIN106) and sequenced for 48 hours using the Oxford Nanopore MinION device. For the RNA004 library, 300 ng of poly(A)+ RNA was used as input for library preparation using the Direct RNA Sequencing Kit (SQK-RNA004, ONT) according to the protocol of the manufacturer. The library was then loaded onto a PromethION RNA

Flow Cell (FLO-PRO004RA) and sequenced on the Oxford Nanopore PromethION device for 72 hours.

For Direct cDNA sequencing, we utilized the Direct cDNA Sequencing Kit (SQK-DCS109, ONT) following the manufacturer's protocol. Briefly, 5 µg of total RNA was used as input for first-strand cDNA synthesis using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) with the SSP and VN primers provided in the kit. To eliminate potential RNA contamination, we treated the sample with RNase Cocktail Enzyme Mix (Thermo Fisher Scientific). Second-strand cDNA synthesis was carried out using LongAmp Taq Master Mix (New England Biolabs). The resulting double-stranded cDNA underwent end-repair and dA-tailing using the NEBNext Ultra End Repair/dA-Tailing Module (New England Biolabs). Subsequently, sequencing adapters were ligated to the prepared cDNA using Blunt/TA Ligase Master Mix (New England Biolabs). Between each enzymatic step, the cDNA and libraries were purified using AMPure XP beads (Agencourt, Beckman Coulter). We quantified the libraries using a Qubit Fluorometer 3.0 (Life Technologies) to ensure adequate concentration and quality. The final library was loaded onto a MinION R9.4 flow cell and sequenced on the Oxford Nanopore MinION device for 72 hours.

Training data preparation

We acquired ONT dRNA-seq FAST5 data from the SG-NEx project, which includes six human cell lines: HEK293T, A549, K562, HepG2, MCF7, and HCT116 [17]. The FAST5 files were converted to POD5 format using the POD5 conversion tool (https://pod5-file-format.readthedocs.io). Subsequently, FASTQ files were generated using Dorado (v0.5.2) [18] with adapter trimming disabled (`--no-trim`) and the "rna002_70bps_hac@v3" model. The reads were then aligned to the human reference genome (GRCh38) using minimap2 (v2.24) [20] with ONT direct RNA-specific parameters (-ax splice -uf -k14) for optimized alignment. The resulting SAM files were then converted to BAM format, indexed, and sorted using SAMtools (v1.19.2) [27].

For adapter sequence extraction, we selected primary alignments without supplementary alignments and implemented a refined identification protocol. While 3′ end soft-clipped regions were candidates for adapter sequences, we did not assume all such regions corresponded to adapters. Instead, we incorporated a critical biological refinement step: we first identified polyA tails at the beginning of soft-clipped regions, as these represent reliable biological indicators of transcript termination. Only sequences following these polyA tails were designated as potential adapter sequences, while aligned regions were classified as non-adapter sequences. This approach significantly improved the precision of our training data by distinguishing true adapter sequences from other non-adapter soft-clipped regions that might result from alignment artifacts or sequencing errors. By anchoring our adapter identification to known biological features, we reduced the risk of misclassification and ensured the training data more accurately reflected the natural transcript-adapter boundaries encountered in dRNA-seq.

To create artificial chimeric reads, we randomly combined two non-adapter sequences with one adapter sequence to create FASTQ records. The dataset consists

of positive examples containing adapter sequences (with a 1:1 ratio of 3′ end and internal adapters) and negative examples without any adapter sequences, in a 9:1 ratio. In total, 600,000 data points were generated and divided into training ($N = 480,000$), validation ($N = 60,000$), and test sets ($N = 60,000$) in an 8:1:1 ratio using stratified random sampling.

Language model architecture

DeepChopper approaches adapter sequence identification as a token classification task, utilizing a model with 4.6 million trainable parameters. The system tokenizes biological sequences at single-nucleotide resolution, with each nucleotide (A, C, G, T, and N) serving as a fundamental token. This nucleotide-level granularity enables precise discrimination between artificial adapter sequences and native biological sequences.

At its core, DeepChopper employs HyenaDNA [9] as its primary feature extractor. HyenaDNA processes the input sequence using multiple attention-free linear layers with a receptive field, transforming the nucleotide tokens into rich 256-dimensional feature representations. The model handles variable-length sequences through a padding approach, maintaining consistent performance across different sequence lengths while efficiently capturing long-range dependencies.

These features are then fed through a quality block that incorporates standardized base quality scores. Prior to processing, the quality scores are normalized using z-score standardization ($\mu = 0$, $\sigma = 1$) to ensure numerical stability. The quality block, comprising two MLPs with residual connections (hidden dimensions: 256), processes this normalized quality information while preserving the original sequence features. Each MLP layer is followed by ReLU activation, enhancing the model's ability to learn complex quality-sequence relationships.

The processed sequence features are subsequently fed into a classification head consisting of a two-layer neural network. This architecture transforms the feature representations into classification outputs at the nucleotide level. The classification module employs a softmax activation function to compute probability distributions across two classes: adapter and non-adapter. For a given nucleotide position with output logits z_1 and z_2 (corresponding to adapter and non-adapter classes), the softmax function computes class probabilities as:

$$P(y_i = c) = \frac{e^{z_c}}{\sum_{j=1}^{2} e^{z_j}}$$

where $P(y_i = c)$ represents the probability that nucleotide position i belongs to class c. The final classification decision is based on the class with the higher probability score. In other words, a nucleotide is classified as an adapter if $P(y_i = \text{adapter}) > P(y_i = \text{non-adapter})$.

This nucleotide-level classification strategy allows DeepChopper to identify adapter boundaries with high precision, including both terminal and internal adapter sequences.

Model training

DeepChopper processes sequences up to 32,770 nucleotides in length, excluding any longer sequences from analysis. To ensure efficient batch processing, shorter sequences were padded to this maximum length. The model was trained using a supervised learning approach, utilizing sequences labeled with adapter annotations. Training was performed in a High Performance Computing (HPC) cluster using two A100 Graphics Processing Units (GPUs). The batch size was set to 64, and validation was performed every 20,000 steps. The model with the highest validation F1 score for the base prediction task was selected for subsequent analyses. Training was carried out over 60 epochs, with early stopping applied based on validation performance to mitigate overfitting risks.

The Adam optimizer was used for parameter optimization, with settings of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [28]. A learning rate scheduler was used to reduce the learning rate when validation loss ceased improving, starting with an initial learning rate of 2×10^{-5}. The cross-entropy loss function was used to update the model parameters, defined as follows:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss, N is the total number of tokens in the input sequence, y_i is the ground true label for the i-th token, and $\hat{y}_i$ is the predicted probability for the i-th token.

The average cross-entropy loss across the mini-batch is computed as:

$$\mathcal{L}_{\text{BatchBCE}} = \frac{1}{B}\sum_{j=1}^{B}\mathcal{L}_{\text{BCE}}(\mathbf{y}_j, \hat{\mathbf{y}}_j)$$

where $\mathcal{L}_{\text{BatchBCE}}$ is the average binary cross-entropy loss for the mini-batch, B is the batch size (number of sequences in the mini-batch), and $\mathbf{y}_j$ and $\hat{\mathbf{y}}_j$ are the true labels and predicted probabilities for the j-th sequence in the batch.

The model evaluation metrics included accuracy, precision, recall and the F1 score, calculated using the following equations:

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

The final selection of the model was based on the optimal performance in the validation set. The model is implemented by PyTorch (v2.5.0) [29]. To identify the best hyperparameter configuration, the Hydra (v1.3.2) [30] framework was used.

Ablation Study of Quality Block Component

An ablation study was conducted to compare two model variants: one with the quality block component and one without. Both models were trained using the same dataset of 100,000 samples, following the training procedures described in Training data preparation and Model training. All hyperparameters, including learning rate, batch size, and optimization algorithm, were kept constant across both configurations. The only architectural difference was the inclusion or exclusion of the quality block. Evaluation was performed on a held-out test set using the F1 score as the primary metric.

Sliding window approach for prediction refinement

To improve prediction consistency and reduce local noise, a sliding window approach was implemented for post-processing of nucleotide-level classification outputs. This method extends predicted adapter regions and smooths isolated predictions, better reflecting the typical length distribution of adapter sequences in ONT dRNA-seq data. The approach enhances continuity in adapter-labeled regions and mitigates the occurrence of fragmented or spurious classifications.

Within each window, a voting mechanism was applied to refine the predictions. The majority voting approach uses a straightforward plurality system. For each nucleotide position, predictions from a window contribute "votes" for whether that position belongs to an adapter region or not. The final classification y_i for each nucleotide x_i was determined by the majority vote of all predictions within the window, as defined by the following equation:

$$y_i = \begin{cases} 1 & \text{if } \sum_{j=i-k}^{i+k} p_j > \frac{W}{2} \\ 0 & \text{otherwise} \end{cases}$$

where y_i is the final prediction for the i-th nucleotide, W is the sliding window size, k is half the window size ($k = \frac{W-1}{2}$), and p_j represents the initial predicted label for the j-th nucleotide within the window, where a value of 1 indicates that the nucleotide is part of an adapter sequence, and a value of 0 indicates that it is part of a non-adapter sequence.

The default window size is set to 21 nucleotides, and can be customized using the *–smooth-window* parameter in the DeepChopper implementation to accommodate dataset-specific characteristics.

Post-processing and filtering

After refining the adapter predictions, four filtering steps were applied to enhance the quality of the final results:

1. A predicted adapter sequence must be at least 13 nucleotides long. Sequences shorter than this length threshold are not considered valid adapters.
2. If a read contains more than four adapter sequences, the entire read sequence is retained without any adapter removal.

3. For reads containing four or fewer adapter sequences, the identified adapters are removed and the read is divided into smaller segments.
4. Any segments resulting from this process that are less than 20 nucleotides long are discarded.

Each remaining segment and its corresponding base quality scores are stored as a single read record in the final FASTQ file. This filtering process separates chimeric read artifacts containing internal adapters into multiple segments, while retaining reads with 3′ end adapters as single shortened segments.

All filtering thresholds, including minimum segment length, and maximum adapter count per read, are configurable via command-line parameters in the DeepChopper implementation, allowing users to tailor these settings to dataset-specific requirements or experimental conditions.

BLAT identity calculation

The accuracy of DeepChopper in detecting adapter sequences was evaluated by aligning the identified sequences to the human reference genome using BLAT [21]. A BLAT identity score was defined as the ratio of matched bases to the total sequence length:

$$\text{BLAT Identity} = \frac{\text{Match Length}}{\text{Sequence Length}}$$

In this context, match length refers to the number of bases in the query sequence that align with the reference genome, while sequence length denotes the total length of the query sequence. This score provides a quantitative measure of how closely each identified sequence aligns with the reference genome, serving as an indicator of detection accuracy. The alignments were performed using the PxBLAT (v1.2.1) [31]

Computational benchmarks

All benchmarks were conducted in triplicate using btop (https://github.com/aristocratos/btop, v1.4.0) and nvtop (https://github.com/Syllo/nvtop, v3.1.0) to monitor CPU and GPU memory usage, respectively. Evaluations were performed on high-performance computing infrastructure with 16 CPU cores, 60 GB RAM, and dual NVIDIA A100 GPUs (80 GB memory each). Adapter prediction stage used a batch size of 64.

Validation of chimera artifact reduction

Cross-platform validation of dRNA-seq chimera artifacts identified by DeepChopper was conducted leveraging ONT direct cDNA sequencing and additional cDNA-based sequencing platforms. Direct cDNA sequencing validation was performed using six cancer cell lines, including the VCaP dataset generated in this study and five published datasets (A549, K562, HepG2, MCF7, and HCT116) obtained from the SG-NEx project [17]. The direct cDNA data in FAST5 format were converted to POD5 format using the POD5 conversion tool (https://pod5-file-format.readthedocs.io). Subsequently, FASTQ files were generated using Dorado (v0.5.2) [18] with adapter trimming

enabled (`--trim` adapters) and the "dna_r9.4.1_e8_hac@v3.3" model. The reads were then processed using Pychopper (https://github.com/epi2me-labs/pychopper, v2.7.9) and Cutadapt (v4.2) [32] according to a published protocol [33]. The oriented reads were aligned to the human reference genome (GRCh38) using minimap2 (v2.24) [20] with optimized parameters (-ax splice -uf -k14) for spliced alignment. The resulting SAM files were then converted to BAM format, indexed, and sorted using SAMtools (v1.19.2) [27].

Additional cDNA-based long-read sequencing data from the WTC11 (human) and F121-9 (mouse) cell lines were used for further validation, incorporating five distinct platforms: ONT PCR-cDNA, ONT CapTrap, ONT R2C2, PacBio cDNA, and PacBio CapTrap. The raw FASTQ files (and FASTA files for ONT R2C2) from these datasets were provided by the LRGASP project [22]. For the PCR-cDNA data, the reads were processed using Pychopper (https://github.com/epi2me-labs/pychopper, v2.7.9) and Cutadapt (v4.2) [32], following the protocol described in reference [33]. ONT reads were then aligned to the human reference genome (GRCh38) or mouse reference genome (GRCm39) using minimap2 (v2.24) [20] with the parameters (-ax splice -uf -k14), while PacBio reads were aligned using the parameters (-ax splice:hq -uf). The ONT dRNA-seq data from A549, K562, HepG2, MCF7, HCT116, VCaP, WTC11 and F121-9 cell lines were processed as previously described, except that The F121-9 cell line data was aligned to the mouse reference genome (GRCm39).

To validate the chimeric alignments derived from dRNA-seq, comparisons were made with chimeric alignments identified from cDNA-based data across the specified platforms. Chimeric alignments, defined by a primary alignment and one or more supplementary alignments, each containing the SA tag in the BAM file, were converted into lists of genomic intervals based on their corresponding alignments. The genomic interval lists were then compared between platforms, and overlapping intervals were considered concordant if the distance difference between them was less than 1000 bp. Supporting rates were calculated as the proportion of dRNA-seq chimeric alignments corroborated by cDNA-based platforms, thereby providing cross-validation of chimera artifacts identified by DeepChopper.

Gene expression analysis and transcript classification

Gene expression levels from dRNA-seq were quantified using IsoQuant (v3.1.2) [24], with the parameters (`--data_type nanopore --stranded forward --model_construction_strategy default_ont --sqanti_output`). The "`--sqanti_output`" option enables IsoQuant to generate files containing transcript classification information, analogous to the output provided by SQANTI [34].

Gene fusion identification and visualization

For ONT dRNA-seq data, gene fusions were identified using FusionSeeker (v1.0.1) [25] with default settings. For short-read RNA-seq data, FASTQ files for the VCaP cell line were obtained from the Cancer Cell Line Encyclopedia (CCLE) project [35] under SRA accession SRX5417211. Raw reads were mapped to the hg38 reference genome using STAR (v2.7.11) [36], and gene fusion events were detected with Arriba (v2.4.0) [26].

The gene structure of the RPS29-COX8A fusion was visualized using GSDS (v2.0) [37]. Base quality scores were generated with a custom Python script, and ion current signals were visualized using Squigualiser (v0.6.3) [38]. The circos plot for gene fusion events was visualized using chimeraviz (v1.30.0) [39].

GO enrichment analysis

GO enrichment analysis of biological processes for genes involved in chimera artifacts identified in dRNA-seq data was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) webserver [40].

Computing resource

All computations were performed on a HPC server equipped with a 64-core Intel(R) Xeon(R) Gold 6338 CPU and 256 GB of RAM. The server was also configured with two NVIDIA A100 GPUs, each with 80 GB of memory, enabling efficient processing of both CPU-intensive tasks and GPU-accelerated deep learning workloads.

Data Availability. Raw and processed data generated in this study, including dRNA-seq using the SQK-RNA002 and SQK-RNA004 kits, as well as direct cDNA sequencing of VCaP cells, have been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE277934. A secure token (sdwfckwmbzqdbyx) has been provided for reviewers to access the deposited data.

Code Availability. DeepChopper, implemented in Rust and Python, is open source and available on GitHub (https://github.com/ylab-hi/DeepChopper) under the Apache License, Version 2.0. The package can be installed via PyPI (https://pypi.org/project/deepchopper/) using pip, with wheel distributions provided for Windows, Linux, and macOS to ensure easy cross-platform installation. An interactive demo is available on Hugging Face (https://huggingface.co/spaces/yangliz5/deepchopper), allowing users to test DeepChopper's functionality without local installation. For large-scale analyses, we recommend using DeepChopper on systems with GPU acceleration. Detailed system requirements and optimization guidelines are available in the repository's documentation.

Acknowledgements. This project was supported in part by NIH grants R35GM142441 and R01CA259388 awarded to RY, and NIH grants R01CA256741, R01CA278832, and R01CA285684 awarded to QC.

Author Contributions. YL, TYW and RY designed the study with QC. YL and TYW performed the analysis. QG, YR and XL performed the experiments. YL designed and implemented the model and computational tool. YL, TYW, QG and RY wrote the manuscript. RY supervised this work.

Conflict of interests. RY has served as an advisor/consultant for Tempus AI, Inc. This relationship is unrelated to and did not influence the research presented in this study.

Acronyms

ATCC American Type Culture Collection 11

BLAT BLAST-like alignment tool 5, 7, 8, 16, 26

CCLE Cancer Cell Line Encyclopedia 17

DAVID Database for Annotation, Visualization, and Integrated Discovery 18
dRNA-seq direct RNA sequencing 1, 2, 4–12, 15–18, 23–28

FSM Full splice match 10, 27

GEO Gene Expression Omnibus 18
GLM Genomic Language Model 2
GO Gene Ontology 9, 10, 18, 28
GPU Graphics Processing Unit 14, 18

HPC High Performance Computing 14, 18

ISM Incomplete splice match 10, 27

LCGLM long-context genomic language model 3
LLM Large Language Model 2
LRGASP Long-read RNA-Seq Genome Annotation Assessment Project 8, 17

MLP multilayer perceptron 3, 13
mRNA messenger RNA 11

NIC Novel in catalog 10, 27
NNC Novel not in catalog 10, 27

ONT Oxford Nanopore Technologies 2, 4–12, 15–17, 25

PacBio Pacific Biosciences 5, 8, 17, 25
PCR Polymerase Chain Reaction 2, 8, 17

RTA reverse transcriptase adapter 11

SG-NEx Singapore Nanopore Expression Project 4, 6, 7, 10, 12, 16

References

[1] Garalde, D. R. *et al.* Highly parallel direct rna sequencing on an array of nanopores. *Nature methods* **15**, 201–206 (2018).

[2] Jain, M., Abu-Shumays, R., Olsen, H. E. & Akeson, M. Advances in nanopore direct rna sequencing. *Nature methods* **19**, 1160–1164 (2022).

[3] Smith, M. A. *et al.* Molecular barcoding of native rnas using nanopore sequencing and deep learning. *Genome research* **30**, 1345–1353 (2020).

[4] Liu-Wei, W. *et al.* Sequencing accuracy and systematic errors of nanopore direct rna sequencing. *BMC genomics* **25**, 528 (2024).

[5] epi2me-labs/pychopper: cdna read preprocessing. Github. URL https://github.com/epi2me-labs/pychopper.

[6] Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex minion sequencing. *Microbial Genomics* **3** (2017). URL http://dx.doi.org/10.1099/mgen.0.000132.

[7] Bonenfant, Q., Noé, L. & Touzet, H. Porechop_abi: discovering unknown adapters in oxford nanopore technology sequencing reads for downstream trimming. *Bioinformatics Advances* **3**, vbac085 (2023).

[8] Benegas, G., Ye, C., Albors, C., Li, J. C. & Song, Y. S. Genomic language models: Opportunities and challenges (2024). URL https://arxiv.org/abs/2407.11435. 2407.11435.

[9] Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems* **36** (2024).

[10] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. 1512.03385.

[11] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome **37**, 2112–2120.

[12] Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for multi-species genome. 2306.15006.

[13] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 1–11 (2024).

[14] Lopes, I., Altab, G., Raina, P. & De Magalhães, J. P. Gene size matters: an analysis of gene length in the human genome. *Frontiers in Genetics* **12**, 559998

(2021).

[15] Workman, R. E. *et al.* Nanopore native rna sequencing of a human poly (a) transcriptome. *Nature methods* **16**, 1297–1305 (2019).

[16] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024). URL https://www.science.org/doi/abs/10.1126/science.ado9336.

[17] Chen, Y. *et al.* A systematic benchmark of nanopore long read rna sequencing for transcript level analysis in human cell lines. *BioRxiv* 2021–04 (2021).

[18] PLC., O. N. Dorado. https://github.com/nanoporetech/dorado (2023).

[19] PLC., O. N. Chemistry Technical Document (CHTD_500_v1_revAQ_07Jul2016) (2017). URL https://nanoporetech.com/document/chemistry-technical-document.

[20] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[21] Kent, W. J. Blat—the blast-like alignment tool. *Genome research* **12**, 656–664 (2002).

[22] Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read rna-seq methods for transcript identification and quantification. *Nature methods* 1–15 (2024).

[23] Hewel, C. *et al.* Direct rna sequencing enables improved transcriptome assessment and tracking of rna modifications for medical applications. *bioRxiv* 2024–07 (2024).

[24] Prjibelski, A. D. *et al.* Accurate isoform discovery with isoquant using long reads. *Nature Biotechnology* **41**, 915–918 (2023).

[25] Chen, Y. *et al.* Gene fusion detection and characterization in long-read cancer transcriptome sequencing data with fusionseeker. *Cancer research* **83**, 28–33 (2023).

[26] Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from rna sequencing data. *Genome research* **31**, 448–460 (2021).

[27] Li, H. *et al.* The sequence alignment/map format and samtools. *bioinformatics* **25**, 2078–2079 (2009).

[28] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[29] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[30] Yadan, O. Hydra - a framework for elegantly configuring complex applications. Github (2019). URL https://github.com/facebookresearch/hydra.

[31] Li, Y. & Yang, R. PxBLAT: an efficient python binding library for BLAT. *BMC Bioinf.* **25**, 1–8 (2024).

[32] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).

[33] Grünberger, F., Ferreira-Cerca, S. & Grohmann, D. Nanopore sequencing of rna and cdna molecules in escherichia coli. *Rna* **28**, 400–417 (2022).

[34] Tardaguila, M. *et al.* Sqanti: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome research* **28**, 396–411 (2018).

[35] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

[36] Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

[37] Hu, B. *et al.* Gsds 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**, 1296–1297 (2015).

[38] Samarakoon, H. *et al.* Interactive visualisation of raw nanopore signal data with squigualiser. *Biorxiv* 2024–02 (2024).

[39] Lågstad, S. *et al.* chimeraviz: a tool for visualizing chimeric rna. *Bioinformatics* **33**, 2954–2956 (2017).

[40] Sherman, B. T. *et al.* David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic acids research* **50**, W216–W221 (2022).

Extended data

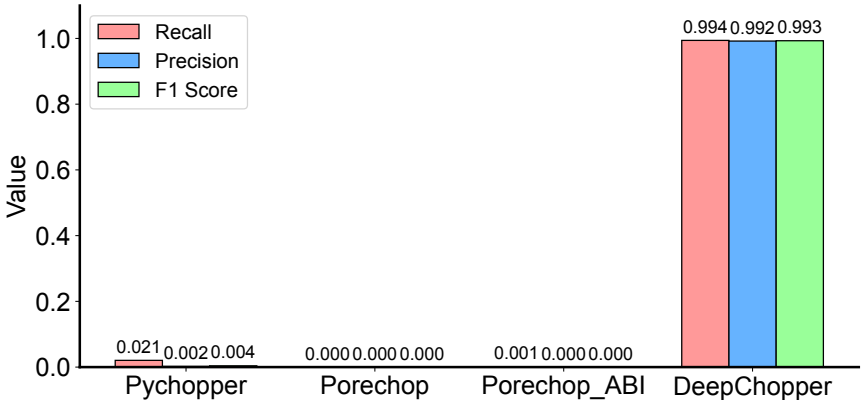
Extended Data Table 1 Summary of Adapter Trimming Tools for analyzing dRNA-seq data

Adapter trimming tool	dRNA-seq terminal adapter trimming	dRNA-seq internal adapter trimming	Trimming existing dRNA-seq datasets (post-basecalling)
Porechop [6]	×	×	×
Porechop_ABI [7]	×	×	×
Pychopper [5]	×	×	×
Dorado [18]	✓	×	×
DeepChopper	✓	✓	✓

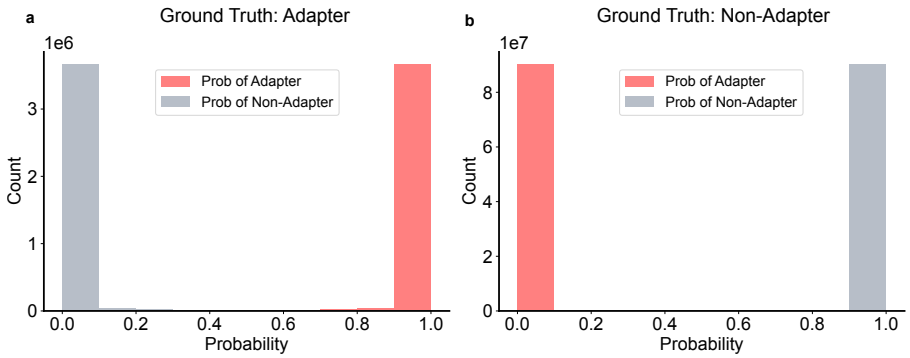
✓ indicates the tool supports this functionality; × indicates the tool does not support this functionality.

Extended Data Table 2 Ablation Study Results for Quality Block

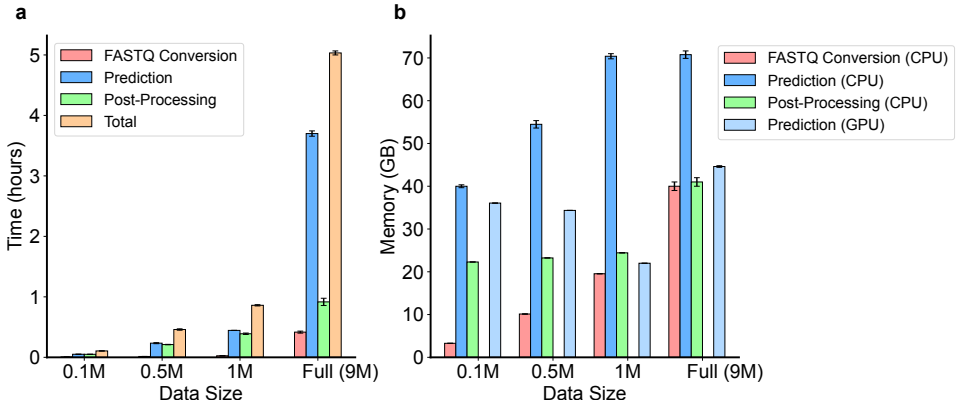
Model Configuration	F1 Score
With Quality Block	**0.99**
Without Quality Block	0.97



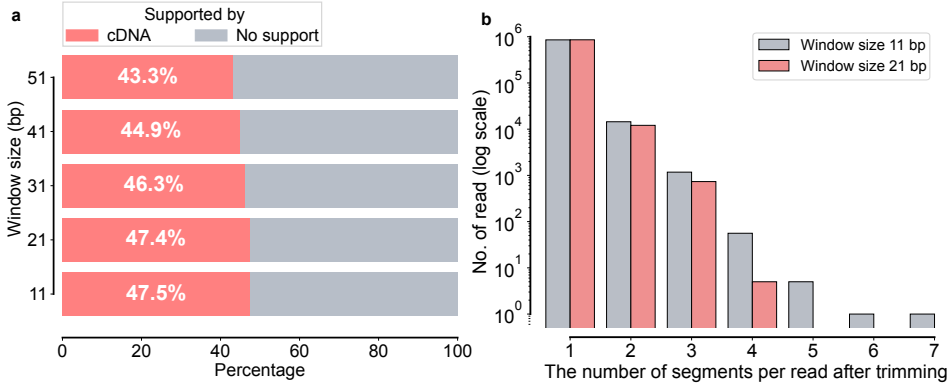
Extended Data Fig. 1 Performance evaluation in a held-out test dataset ($N = 60,000$) showing Recall, Precision, and F1 values for DeepChopper, Pychopper, Porechop, and Porechop_ABI.



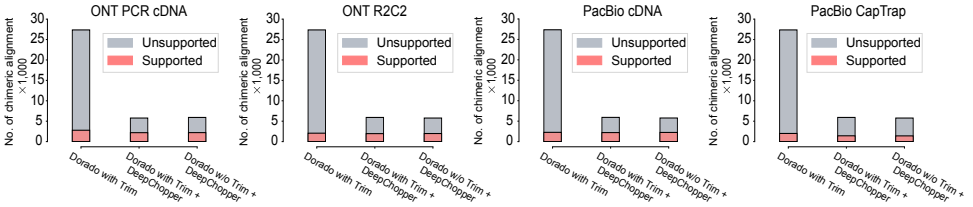
Extended Data Fig. 2 Prediction probability distributions of DeepChopper for the held-out test dataset ($N = 60,000$). (a) Distribution of prediction probabilities for sequences with ground truth adapter classification. Red bars represent the probability of adapter prediction, while gray bars show the probability of non-adapter prediction. The count (y-axis) is shown in millions of sequences (10^6 scale). (b) Distribution of prediction probabilities for sequences with ground truth non-adapter classification. Red bars indicate the probability of adapter prediction, while gray bars show the probability of non-adapter prediction. The count (y-axis) is shown in tens of millions of sequences (10^7 scale). Both distributions demonstrate strong polarization toward correct classification probabilities, indicating the model's high confidence in distinguishing between adapter and non-adapter sequences.



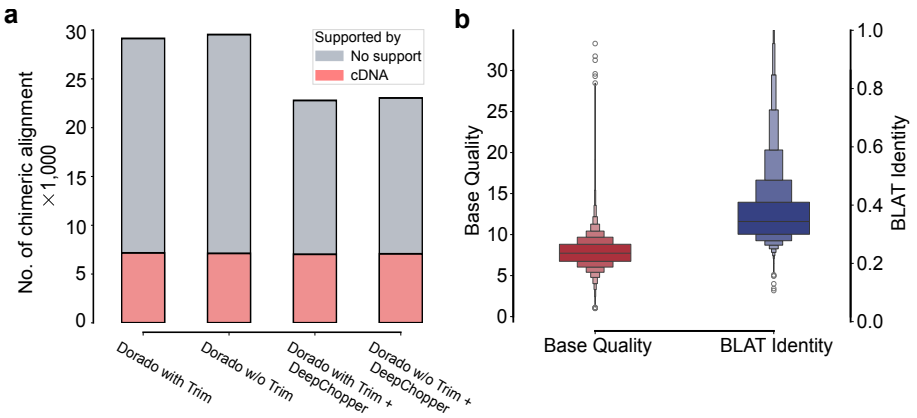
Extended Data Fig. 3 Computational performance metrics across different data sizes from the VCaP cell line dRNA-seq. (a) Runtime analysis showing processing time requirements for different pipeline stages (FASTQ Conversion, Prediction, Post-Processing) and total runtime across four data sizes : subsampled (0.1M, 0.5M, 1M) and full (9M) reads derived from the VCaP dRNA-seq. As data size increases, prediction time becomes the dominant component, with the full dataset requiring approximately 5 hours of total processing time. (b) Memory usage comparison between CPU and GPU implementations across the same data sizes. The prediction stage shows consistently higher memory requirements, with CPU memory usage for prediction reaching approximately 70 GB for the larger datasets. All measurements include error bars representing standard deviation from three runs.



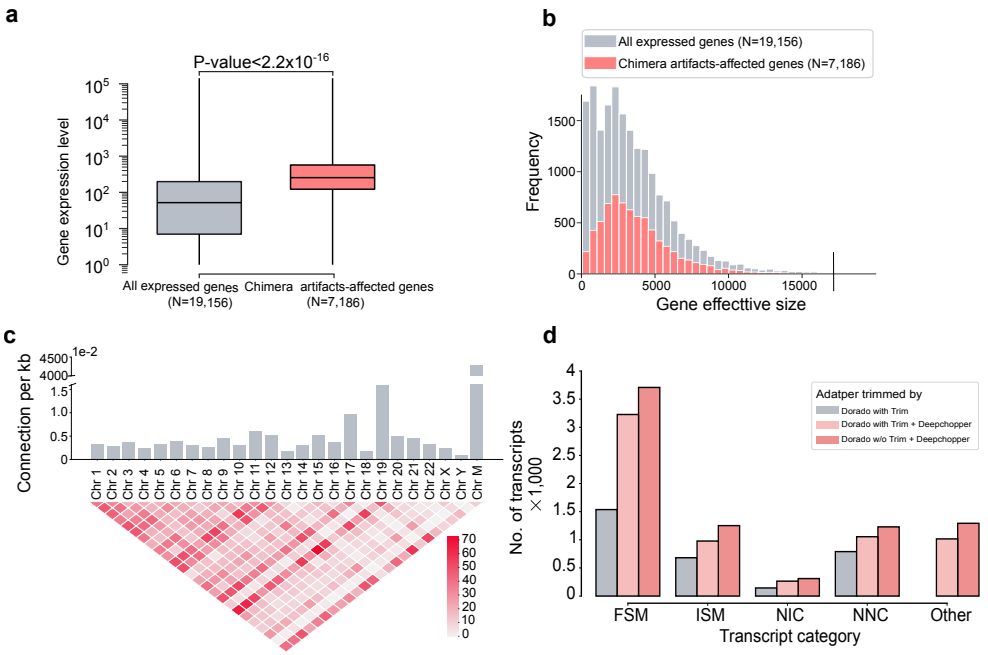
Extended Data Fig. 4 Effect of window size on chimeric alignment detection and read fragmentation. (a) Analysis of different sliding window sizes (11, 21, 31, 41, and 51 nucleotides) showing the percentage of cDNA-supported chimeric alignments (red bars) in VCaP. Higher percentages indicate better support. (b) Distribution of the number of segments per read after trimming (x-axis) for window sizes 11 (gray) and 21 (pink), shown on a logarithmic scale (y-axis). Data represents subsampling of 1M reads from the VCaP dataset. Window size 21 maintains similar detection sensitivity to window size 11 while producing significantly fewer fragmented reads.



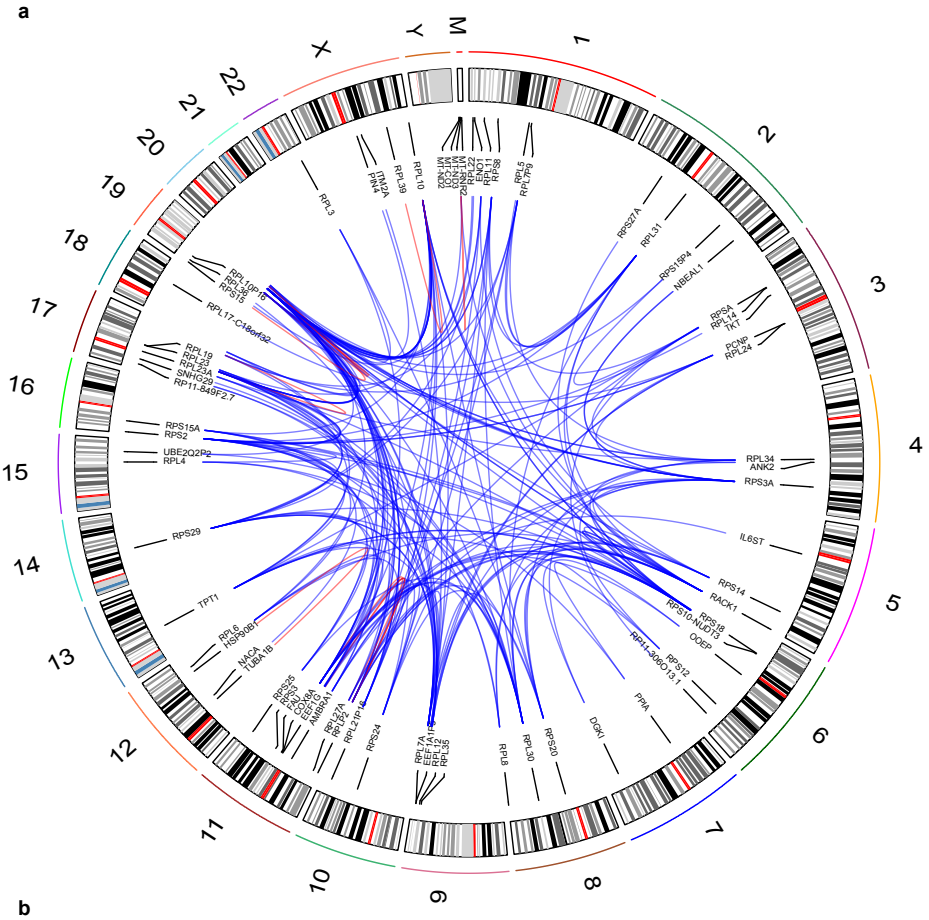
Extended Data Fig. 5 Chimeric alignments from dRNA-seq of the F121-9 cell line (mouse), evaluated for support using additional ONT and PacBio sequencing data with different protocols. DeepChopper-involved methods reduce unsupported chimeric alignments across all methods compared to Dorado with adapter trimming. The bar colors indicate chimeric alignments supported by additional sequencing data (red) and those lacking support (grey).



Extended Data Fig. 6 Evaluation of DeepChopper's predictions on chimeric read artifacts in dRNA-seq data generated using the SQK-RNA004 kit from the VCaP cell line. (a) Number of chimeric alignments (in thousands) identified in VCaP RNA004 dRNA-seq reads processed by Dorado with and without adapter trimming, Dorado with adapter trimming followed by DeepChopper, and DeepChopper. The bar colors indicate chimeric alignments supported by cDNA sequencing (red) and those lacking support (grey). (b) Base quality scores (left) and BLAT alignment identity scores (right) for internal adapter sequences identified by DeepChopper in RNA004 dRNA-seq reads.



Extended Data Fig. 7 Analysis of dRNA-seq chimera artifacts and their genomic and transcriptomic characteristics in VCaP cells. (a) Box plot comparing gene expression levels between all expressed genes (N=19,156) and genes affected by chimera artifacts (N=7,186) in the VCaP dRNA-seq dataset. Chimera artifacts-affected genes exhibit significantly higher expression levels (p-value $< 2.2 \times 10^{-16}$). (b) Distribution of gene effective sizes for all expressed genes and genes affected by chimera artifacts, indicating that the size distributions of genes impacted by chimera artifacts are comparable to those of all expressed genes. (c) Chromosomal distribution and interchromosomal connections from chimeric read artifacts arising from VCaP RNA004 dRNA-seq. The top bar plot shows the number of connections per kilobase for each chromosome, with higher bars indicating more frequent connections. The bottom heatmap visualizes the number of chimeric connections between chromosome pairs, with color intensity representing the connection frequency. (d) Number of detected transcripts across different isoform categories (FSM, ISM, NIC, NNC, and Other) from DeepChopper-identified chimeric read artifacts in VCaP RNA004 dRNA-seq data. DeepChopper-corrected reads resulted in a greater number of transcripts compared to adapter-trimmed reads by Dorado across all categories.



b

Sample	GO Term	Genes	P-value
A549	Cytoplasmic translation	RPL34, RPS21	1.467×10^{-2}
	Translation	RPL34, RPS21	3.040×10^{-2}
HepG2	Translation	EEF1A1, RPS19, RPL13, RPS12	2.172×10^{-4}
	Intracellular iron ion homeostasis	MT-RNR2, FTH1, FTL	7.332×10^{-4}
	Cytoplasmic translation	RPS19, RPL13, RPS12	1.528×10^{-3}
	Intracellular sequestering of iron ion	FTH1, FTL	4.250×10^{-3}
	Translational elongation	EEF1A1, EEF1B2	1.150×10^{-2}
	Iron ion transport	FTH1, FTL	1.630×10^{-2}
	Ribosomal small subunit biogenesis	RPS19, RPS12	4.526×10^{-2}
HCT116	Regulation of translation	RPS3, RPL38, GAPDH	6.151×10^{-3}
	Negative regulation of translation	RPS3, RPL13A, GAPDH	6.525×10^{-3}
	Cytoplasmic translation	RPL4, RPS6, RPL13A, RPSA, RPL7A, RPS29, RPS3, RPL14, RPLP2, RPL13, RPS20, RPL38, RPS2, RPL28	1.008×10^{-24}
	Translation	RPL4, RPS6, RPL13A, RPSA, EEF1A1, RPL7A, RPS29, RPS3, RPL14, RPLP2, RPL13, RPS20, RPL38, RPS2, RPL28	1.566×10^{-22}
VCaP (SQK-RNA004 kit)	Cytoplasmic translation	RPS16, RPLP2, RPL29	7.15×10^{-5}
	Translation	RPS16, EEF1A1P5, RPLP2, RPL29	1.06×10^{-6}

Extended Data Fig. 8 Analysis of gene fusions derived from chimeric read artifacts in dRNA-seq. (a) Circos plot depicting chromosomal connections of gene fusions resulting from chimeric read artifacts in VCaP cells. Blue lines represent inter-chromosomal fusion events, while red lines indicate intra-chromosomal fusions. The outer track displays chromosomal ideograms labeled with respective chromosome numbers. (b) GO enrichment analysis of fusion genes derived from chimeric read artifacts identified by DeepChopper in dRNA-seq data from A549, HepG2, and HCT116 cell lines, and VCaP RNA004 dRNA-seq data. The table lists enriched GO terms of biological processes, associated genes, and the statistical significance (p-values) for each enrichment.