

	001
	002
	003
	004
	005
Genomic Language Model Mitigates Chimera	006
Artifacts in Nanopore Direct RNA Sequencing	007
	008
	009
	010
Yangyang Li ^{1†} , Ting–You Wang ^{1†} , Qingxiang Guo ¹ , Yanan Ren ¹ ,	011
Xiaotong Lu ¹ , Qi Cao ^{1,2} , Rendong Yang ^{1,2*}	012
	013
¹ Department of Urology, Northwestern University Feinberg School of	014
Medicine, 303 E Superior St, Chicago, 60611, IL, USA.	015
² Robert H. Lurie Comprehensive Cancer Center, Northwestern	016
University Feinberg School of Medicine, 675 N St Clair St, Chicago,	017
60611, IL, USA.	018
	019
	020
*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu ;	021
Contributing authors: yangyang.li@northwestern.edu ;	022
tywang@northwestern.edu ; qingxiang.guo@northwestern.edu ;	023
ynren1020@gmail.com ; xiaotong.lu@northwestern.edu ;	024
qi.cao@northwestern.edu ;	025
†These authors contributed equally to this work.	026
	027
	028
	029
	030
Abstract	031
Chimera artifacts in nanopore direct RNA sequencing (dRNA-seq) introduce	032
substantial inaccuracies, complicating downstream applications such as tran-	033
script annotation and gene fusion detection. Current basecalling models are	034
unable to detect or mitigate these artifacts, limiting the reliability and utility	035
of dRNA-seq for transcriptomics research. To address this challenge, we present	036
DeepChopper, a genomic language model specifically designed to identify and	037
remove adapter sequences from base-called dRNA-seq long reads with single-base	038
precision. Operating independently of raw signal or alignment information, Deep-	039
Chopper effectively eliminates chimeric read artifacts, significantly enhancing the	040
accuracy of crucial downstream analyses. This improvement in reliability unlocks	041
the full potential of nanopore dRNA-seq , establishing it as a more robust tool for	042
diverse transcriptomics applications.	043
	044
	045
	046

047 **Introduction**

048

049 Long-read RNA sequencing technologies are revolutionizing transcriptomic research
050 by providing unparalleled resolution for detecting complex splicing and gene fusion
051 events often missed by conventional short-read RNA-seq methods. Among these tech-
052 nologies, **Oxford Nanopore Technologies (ONT) dRNA-seq** stands out by sequencing
053 full-length RNA molecules directly, preserving native RNA modifications and allowing
054 a more accurate and comprehensive analysis of RNA biology. This approach bypasses
055 the inherent limitations of cDNA-based sequencing methods, such as artifacts arising
056 from **Reverse Transcription (RT)**, template switching, and **Polymerase Chain Reaction**
057 (**PCR**) amplification [? ? ?].

058 Despite these advantages, a critical question remains: Does **ONT dRNA-seq** intro-
059 duce technical artifacts? A previous study has suggested that **dRNA-seq** might
060 generate chimera artifacts, leading to multi-mapped reads [?], but systematic char-
061 acterization of these artifacts—their prevalence, formation mechanisms, and persis-
062 tence with modern sequencing chemistries such as **RNA004** and current **Dorado**
063 basecalling [?]—remains limited in peer-reviewed literature. These artifacts may
064 result from ligation during library preparation or chimeric reads produced by soft-
065 ware missing the open pore signal, potentially confounding downstream analyses
066 such as transcriptome assembly, quantification, and detection of alternative splic-
067 ing and gene fusion events. Detecting these chimera artifacts is challenging because
068 long-read aligners often produce chimeric alignments from such artifacts that are indis-
069 tinguishable from those derived from true gene fusion events. Importantly, chimeric
070 read artifacts frequently contain internal adapter sequences [?], suggesting that these
071 adapter-bridged chimeras could theoretically be distinguished from biological chimeras
072 by detecting the presence of internal adapters. However, **ONT dRNA-seq** basecallers,
073 trained in RNA, struggle to properly call these DNA-based adapter sequences under
074 an RNA model [?]. As a result, current adapter detection tools [? ? ?] cannot exploit
075 this feature to eliminate these adapter-bridged chimeras, leaving the issue unresolved
076 (Extended Data Table ??).

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

091

092

Results

DeepChopper Architecture and Training

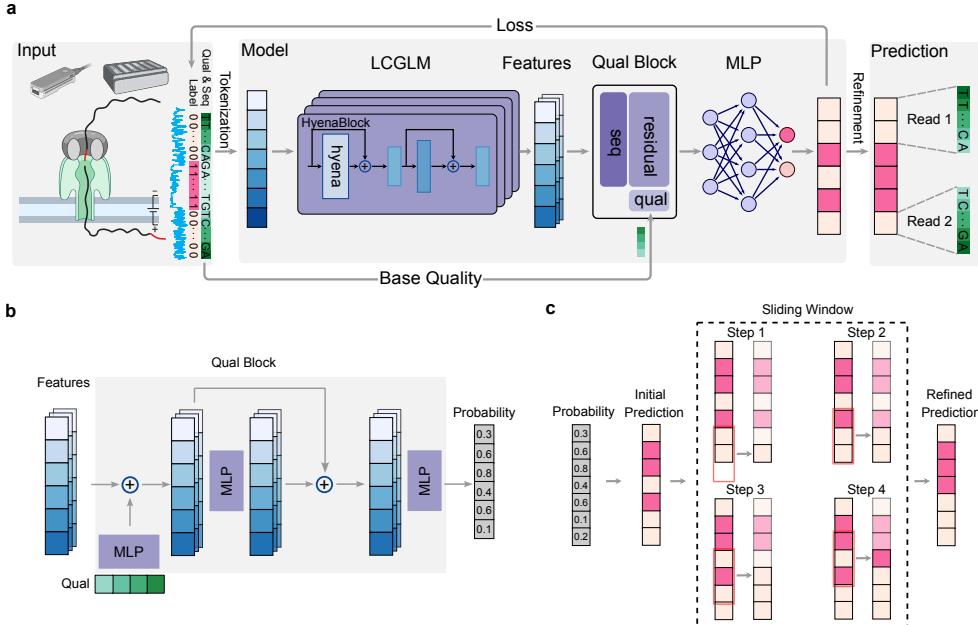


Fig. 1 DeepChopper architecture and methodology. (a) Overview of the DeepChopper model. Raw sequences are first tokenized into vectors and processed by HyenaDNA to generate embedding features. These features are integrated with base quality information in the quality block to produce per-token probability 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.

DeepChopper leverages the [long-context genomic language model \(LCGLM\)](#) HyenaDNA [?], which excels at capturing long-range dependencies (Fig. ??a). To 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 [?] (Fig. ??b, See Methods for details). This addition enables the effective utilization of sequencing base quality, a crucial feature for improving prediction accuracy, particularly for distinguishing genuine adapter sequences from similar motifs that may occur naturally within reads or from low-quality regions. By combining broad contextual understanding with nucleotide-level precision, this hybrid architecture allows DeepChopper to accurately identify and process adapters

093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138

139 sequences. Reads containing internal adapters are split into multiple records, with
140 3' end adapters simultaneously trimmed (Fig. ??a), thereby preserving authentic
141 biological sequences and eliminating chimera artifacts.

142 To further refine the prediction accuracy, DeepChopper implements a post-
143 processing stage using a sliding window and majority vote approach, as illustrated
144 in Fig. ??c. The model applies a sliding window with a stride of 1 across the read,
145 analyzing the distribution of predicted adapter positions within each window (See
146 Methods for details). This refinement process operates on the initial predictions inde-
147 pendently for each position, ensuring that each base's final classification reflects an
148 aggregation of local context without error propagation. By maintaining precise bound-
149 ary detection at single-nucleotide resolution, this strategy ensures that predicted
150 adapter sequences correspond to biologically plausible boundaries, enabling accurate
151 splitting of chimeric reads into their component sub-reads while minimizing spurious
152 fragmentation.

153 Comparing to existing general-purpose GLM, DeepChopper is specifically opti-
154 mized for long-read sequence analysis at single-nucleotide resolution. This fine-grained
155 resolution provides a critical advantage for genomic analysis tasks requiring precise
156 base-level predictions. While DNABERT [?] is limited to input sequences of approxi-
157 mately 512 bp, DNABERT2 [?] to 10,000 bp, and Nucleotide Transformer [?] to 6,000
158 bp, DeepChopper supports input lengths up to 32 kilobases—sufficient to encompass
159 most complete mRNA transcripts. This 32K nucleotide input limit was selected based
160 on both technical and biological considerations. Technically, the constraint reflects
161 the architectural design and context window limitations of the underlying HyenaDNA
162 model. Biologically, human protein-coding transcripts have a median length of approx-
163 imately 2.7 kb (95th percentile: ~ 8.7 kb, 99th percentile: ~ 14 kb), with over 99.97% of
164 transcripts falling below 32 kb based on current annotations (Extended Data Fig. ??).
165 Empirical analysis of our dRNA-seq datasets confirmed that only 0-0.0032% of reads
166 exceeded this threshold across multiple cell lines and chemistries, with mean read
167 lengths ranging from 683 to 1,148 bp (Extended Data Table ??). This extended
168 input capacity, combined with single-nucleotide tokenization, enables DeepChopper to
169 accurately identify non-reference elements such as ONT adapter sequences with base-
170 pair precision—an essential capability for detecting and recovering adapter-bridged
171 chimeras in dRNA-seq data.

172 In addition, DeepChopper's lightweight architecture, consisting of only 4.6 million
173 parameters, makes it computationally efficient and scalable for large-scale dRNA-seq
174 analysis. This is in contrast to models like Evo [?], which require billions of parameters
175 and significantly more computational resources.

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

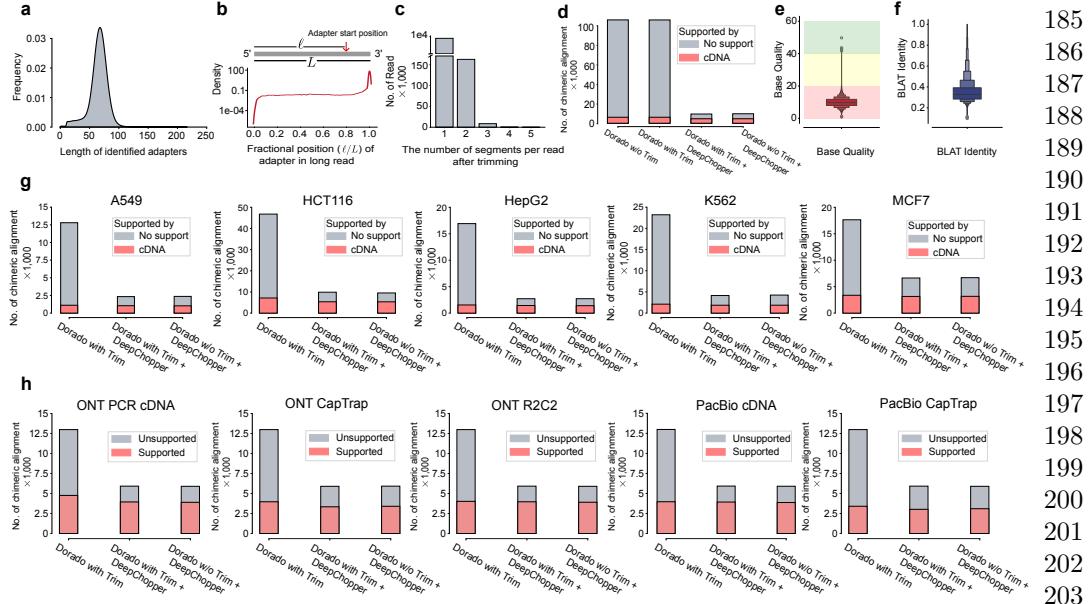


Fig. 2 Detection of chimeric read artifacts in dRNA-seq data using DeepChopper and validation with orthogonal sequencing platforms (a) Length distribution of predicted adapters by DeepChopper in VCaP dRNA-seq data. (b) Distribution of relative adapter position along read length in VCaP dRNA-seq data. 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. Gray bars represent unsupported chimeric alignments (likely artifacts); pink bars represent cDNA-supported chimeric alignments (biological events). 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. Gray bars represent unsupported chimeric alignments (likely artifacts); pink bars represent cDNA-supported chimeric alignments (biological events). 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 Bio-sciences (PacBio) sequencing data with different protocols. DeepChopper-involved methods reduce unsupported chimeric alignments across all methods compared to Dorado with adapter trimming.

DeepChopper Benchmarking and Model Optimization

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

185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230

231 out using the synthetic testing dataset ($n = 60,000$ reads), enabling rigorous assessment
232 of precision, recall, and F1-score metrics. As shown in Extended Data Fig. ??, all
233 existing tools demonstrated negligible performance metrics when processing **dRNA-seq**
234 adapter sequences, indicating fundamental incompatibility with the **dRNA-seq** pro-
235 tocol. In contrast, DeepChopper achieved exceptional accuracy in identifying both
236 terminal and internal adapters, with recall, precision, and F1 scores consistently
237 exceeding 0.99. These results highlight DeepChopper’s unique capability to address
238 the specific complexities inherent to **dRNA-seq** reads and underscore the critical need
239 for purpose-built solutions in this domain.

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

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

262

263 Chimeric Read Artifact Detection in Cancer **dRNA-seq** Data

264

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

274 DeepChopper processes input data through a three-stage pipeline: (1) FASTQ-to-
275 Parquet conversion for efficient input/output, (2) adapter prediction using a neural
276 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. 277
278

We systematically benchmarked DeepChopper’s computational performance to 279
assess scalability for large-scale dRNA-seq studies, testing datasets ranging from 0.1M 280
to 23M reads. Using VCaP read subsamples from 0.1M to 9M reads, we observed near- 281
linear runtime scaling (Extended Data Fig. ??a), with the 9M VCaP dataset requiring 282
approximately 5 hours on two NVIDIA A100 GPUs. Memory usage increased with 283
input size, peaking at approximately 70 GB for CPU-based inference and 56 GB for 284
GPU-based execution (Extended Data Fig. ??b). To evaluate performance at substan- 285
tially larger scale, we benchmarked a merged dataset of 23 million reads combining 286
five cell lines (A549, HCT116, HepG2, K562, and MCF7), which required approxi- 287
mately 10.6 hours total processing time (23 minutes for FASTQ conversion, 8.5 hours 288
for prediction, 1.7 hours for post-processing). Importantly, runtime continued to scale 289
near-linearly and memory usage remained stable (CPU: 70-93 GB, GPU: 34-56 GB) 290
across the entire range, demonstrating that DeepChopper can process datasets sub- 291
stantially larger than 23M reads with no fundamental computational barriers to scaling 292
(See Methods for details). 293

Applying DeepChopper to the full VCaP dataset increased usable read yield by 3%, 294
resulting in 9,357,913 adapter-trimmed reads. It identified 8,218,172 adapter sequences 295
across 7,990,102 reads (87% of total), most measuring 70 bp—consistent with the 296
expected length of the RMX adapter used in ONT’s SQK-RNA002 dRNA-seq kit 297
(Fig. ??a) [?]. Analysis of adapter locations revealed that 7,777,624 reads had adapters 298
at the 3’ end, while 148,452 contained internal adapters (Fig. ??b), indicating that 299
chimeric artifacts are common in VCaP dRNA-seq data. 300

Further examination showed that chimera artifacts could arise from the joining of 301
multiple long reads, with the most frequent pattern involving two reads joined by a sin- 302
gle internal adapter (Fig. ??c). To validate these findings, we analyzed minimap2 [?] 303
chimeric alignments and compared them to a matched VCaP direct cDNA-seq dataset, 304
which we generated as part of this study. Chimeric reads fully supported by cDNA 305
sequencing were considered bona fide events (See Methods for details). Notably, we 306
also evaluated whether ONT’s Dorado basecaller trimming feature could mitigate these 307
artifacts. However, we found that Dorado alone—regardless of trimming—was insuf- 308
ficient to eliminate spurious chimeric alignments. In contrast, DeepChopper reduced 309
unsupported chimeric alignments by around 95% and increased the fraction of cDNA- 310
supported chimeric events from 5.8% to 48.7%, whether applied before or after Dorado 311
trimming (Fig. ??d). These results underscore DeepChopper’s ability to distinguish 312
true biological chimeras from technical artifacts. 313

To further verify the artifactual nature of internal adapters, we analyzed their base 314
quality scores and aligned them to the human reference genome using BLAT [?]. 315
Adapter regions identified within chimera artifacts exhibited significantly lower base 316
quality (Fig. ??e) and poor sequence identity to the reference genome (Fig. ??f), 317
supporting their non-human and non-biological origin. 318

Finally, we evaluated post-processing performance as a function of the sliding win- 319
dow size used for segmentation. Using a 1M-read subsample, we tested window sizes 320
321
322

323 of 11, 21, 31, 41, and 51 bp. Smaller windows yielded slightly higher cDNA sup-
324 port percentages (47.5% for 11 bp vs. 43.3% for 51 bp; Extended Data Fig. ??a),
325 but increased fragmentation of reads into 4+ segments (Extended Data Fig. ??b). A
326 21 bp window provided the optimal balance, maintaining high support while mini-
327 mizing over-segmentation. Based on these results, 21 bp was selected as the default
328 setting, and DeepChopper allows users to adjust this parameter for dataset-specific
329 optimization (See Methods for details).

330

331 Multi-sample Validation Across Platforms and Species

332

333 To further evaluate DeepChopper’s performance beyond the VCaP dataset, we
334 performed multi-sample validation across diverse biological systems and sequenc-
335 ing platforms. We began by analyzing [dRNA-seq](#) data from the [SG-NEx](#) project,
336 comparing chimeric alignments before and after DeepChopper trimming. DeepChop-
337 per detected internal adapters in 0.67-1.25% of reads across these datasets (A549:
338 0.92%, MCF7: 0.67%, HCT116: 1.22%, K562: 0.96%, HepG2: 1.25%), representing
339 15,690-57,122 affected reads per sample (Extended Data Table ??). Critically, inter-
340 nal adapters accounted for 63–85% of all chimeric reads across cell lines, identifying
341 adapter-bridged artifacts as the predominant source of false RNA rearrangement[?].
342 This systematic occurrence of internal adapters across all tested cell lines indicates
343 that adapter-bridged chimeras are not specific to VCaP but represent a general charac-
344 teristic of dRNA-seq data. Across these samples, DeepChopper reduced unsupported
345 chimeric alignments by 62% to 84%, while preserving cDNA-supported chimeric align-
346 ments without noticeable reduction (Fig. ??g). These results reinforce the widespread
347 presence of chimera artifacts in [dRNA-seq](#) and the effectiveness of DeepChopper in
348 selectively removing them without compromising true biological signals.

349

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

358

359 To assess cross-species generalizability, we extended the analysis to the F121-9
360 mouse embryonic stem cell line, also from the [LRGASP](#) dataset. DeepChopper reli-
361 ably removed artifactual chimeric reads not supported by any orthogonal cDNA-based
362 sequencing platform (Extended Data Fig. ??), confirming its applicability to both
363 human and non-human transcriptomes.

364

365 Importantly, across all datasets, DeepChopper consistently outperformed [ONT](#)’s
366 Dorado adapter trimming alone—even when applied as a post-processing step—
367 underscoring its distinct and additive utility in chimera artifact correction.

368

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. We first applied DeepChopper in a zero-shot setting to assess cross-chemistry generalization, as the model was trained exclusively on RNA002 adapter patterns.

In zero-shot application, DeepChopper detected internal adapters in 0.33% of VCaP RNA004 reads (38,878 reads out of 11,714,520 total), lower than the 1.62% observed in VCaP RNA002 (Extended Data Table ??). DeepChopper reduced chimeric alignments by 21% compared to Dorado base-called and adapter-trimmed reads, increasing the proportion of cDNA-supported chimeric alignments from approximately 25% to 30% (Extended Data Fig. ??a). Similar results were observed when DeepChopper was applied after Dorado’s adapter trimming, demonstrating compatibility with standard preprocessing pipelines. 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. ??b).

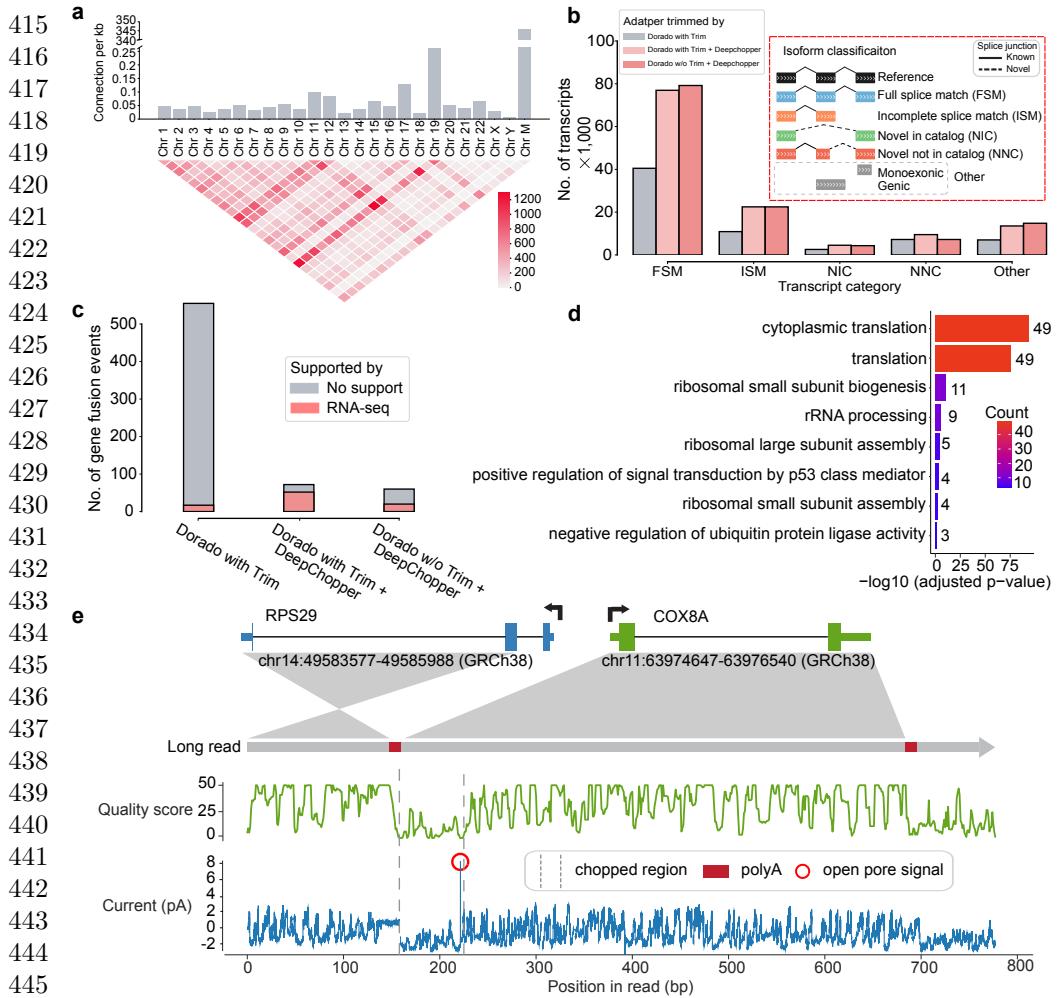
To optimize performance on RNA004 data, we fine-tuned DeepChopper using a dataset with 300,000 reads created from VCaP RNA004 reads (70% training, 20% validation, 10% test) (See Methods for details). The fine-tuned model achieved marginal additional improvement, reducing chimeric alignments by 23-25% compared to Dorado-processed reads—a 3-4% improvement over the zero-shot model (Extended Data Fig. ??). Critically, both the original RNA002-trained model and the RNA004-fine-tuned model preserved all cDNA-supported chimeric alignments, demonstrating that DeepChopper specifically targets adapter-bridged artifacts rather than biological RNA rearrangements[?].

While DeepChopper’s reduction of chimeric alignments in RNA004 (21-22%) is lower than in RNA002 (91%), both the reduced artifact prevalence (0.33% vs 1.62%) and the lower reduction magnitude are expected given chemistry improvements designed to reduce artifact formation [? ?]. Nonetheless, the systematic detection of internal adapters across both RNA002 and RNA004 chemistries confirms that adapter-bridged chimeras remain an inherent characteristic of current dRNA-seq workflows, and DeepChopper’s ability to generalize across chemistries without retraining highlights its robustness for emerging platforms.

Both the original RNA002-trained model and the RNA004-fine-tuned model are available in the DeepChopper repository, providing users with optimized options for different sequencing chemistries.

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 transcriptome (p -value $< 2.2 \times 10^{-16}$; Extended Data Fig. ??a), while exhibiting a similar gene length distribution (Extended Data Fig. ??b). Analysis of chimeric junctions



446 **Fig. 3 Characterization of dRNA-seq chimera artifacts and their impact on downstream**
447 **analysis in VCaP cells** (a) The upper bar plot shows the number of chimeric connections per
448 kilobase across chromosomes, highlighting higher chimeric activity in Chr 19 and Chr M. The lower
449 heatmap visualizes interchromosomal connections, with intensity indicating the count of connections
450 between different chromosomes. (b) The bar plot shows the number of transcripts (in thousands)
451 across different isoform classification categories. DeepChopper-processed reads result in a higher num-
452 ber of transcripts compared to Dorado-trimmed reads. The inset details the isoform classification
453 scheme. (c) Detected gene fusions from Dorado adapter-trimmed reads and DeepChopper-processed
454 reads. Gene fusions identified from short-read RNA-seq were used to validate fusion events detected
455 from dRNA-seq. (d) Gene Ontology (GO) enrichment analysis of chimera artifact-affected genes, with
456 color indicating gene count per term. (e) Analysis of a chimeric read (Read ID: 3b2292e9-43e5-4e40-
457 87d9-ccc23897377c) artifact detected as an RPS29-COX8A fusion. The schematic shows the fusion
458 between RPS29 (Chr 14) and COX8A (Chr 11). The green plot indicates quality scores along the
459 read, and the blue plot shows raw signal intensity (in pA). The chopped region identified by Deep-
460 Chopper corresponds to a low-quality segment with low current intensity, polyA, and short open pore
461 signals, suggesting the presence of an ONT adapter.

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. ??a). This pattern persisted in RNA004 dRNA-seq data (Extended Data Fig. ??c), 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 [?] to annotate transcripts from VCaP dRNA-seq data, we found that DeepChopper nearly doubled the number of identified transcripts compared to uncorrected data (Fig. ??b). Similar results were observed with RNA004 data (Extended Data Fig. ??d) 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 [?] 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 [?] (Fig. ??c), 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. ??a). GO enrichment analyses in VCaP (Fig. ??d) and SG-NEx cell lines (Extended Data Fig. ??b) 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. ??b). 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. ??e). 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 summary, DeepChopper significantly improves the quality of nanopore dRNA-seq data by accurately identifying and splitting adapter-bridged chimera that otherwise confound transcript annotation and gene fusion detection. These improvements are robust across different sequencing chemistries and preprocessing pipelines.

Discussion

DeepChopper addresses a critical gap in ONT dRNA-seq: detecting and recovering adapter-bridged chimeric artifacts through internal adapter identification. Our validation across multiple cell lines, species, and chemistries demonstrates internal adapters occur systematically in 0.33-1.62% of reads—tens to hundreds of thousands per experiment—with each propagating errors to transcript annotation, gene fusion detection,

507 and expression quantification [?]. No prior literature has systematically characterized
508 adapter-bridged chimera formation mechanisms in dRNA-seq; our work addresses this
509 gap while providing the first computational solution. DeepChopper identifies internal
510 adapter sequences marking artificial junctions, then splits chimeras at these bound-
511aries with single-nucleotide precision to recover component sub-reads. This adapter-
512 based approach differs fundamentally from filtering methods that discard problematic
513 reads. The recovery capability depends on detectable internal adapters, distinguishing
514 adapter-bridged artifacts from RT-mediated template-switching in cDNA-seq, which
515 creates chimeras without internal adapters [? ?].

516 DeepChopper leverages GLM capabilities unavailable to conventional approaches:
517 (1) Long-range context [?] enabling full-transcript scanning (median ~2.7 kb, 99th
518 percentile ~14 kb); (2) Transfer learning enabling robust detection despite sequencing
519 errors; (3) Quality-aware predictions integrating per-base confidence (F1: 0.97 → 0.99,
520 Extended Data Table ??); (4) Alignment-free operation eliminating mapping biases.

521 Performance differences between DeepChopper and existing tools
522 (Extended Data Table ??, Extended Data Fig. ??) reflect capability gaps, not qual-
523 ity. Each tool excels at its intended application: Pychopper for cDNA primer
524 detection [? ?] and Dorado for 3' adapter trimming [?]. However, none detect
525 internal adapters in dRNA-seq. Pychopper targets cDNA-specific primers absent
526 in dRNA-seq. Porechop/Porechop_ABI require exact matching or stable k-mer
527 patterns [? ?]—incompatible with unknown, corrupted adapters mis-called by RNA
528 basecallers. Negligible performance on internal detection confirms this is a genuinely
529 unsolved problem.

530 Alignment-based methods employ ruled-based filtering—flagging aberrant align-
531 ment patterns—rather than detecting root causes. Breakinator [?] detects RT-medi-
532 ated foldbacks in cDNA-seq using distance rules; its validation confirms failure in
533 dRNA-seq where RT artifacts don't occur. FLAIR-fusion [?] is a specialized fusion
534 caller, not general quality control. Empirical comparison on VCaP RNA002: Breakina-
535 tor reduced artifacts 62% but decreased cDNA support to 4.8% (below baseline 5.8%),
536 indicating indiscriminate removal (Extended Data Fig. ??). DeepChopper reduced
537 artifacts 91% while increasing cDNA support to 49%. Distance rules cannot distin-
538 guish adapter-bridged artifacts from biological RNA rearrangements [?] and they
539 cannot handle artifacts that do not map as chimeric reads or that fail to map at
540 all (Extended Data Fig. ??). Internal adapter presence definitively indicates artifacts,
541 enabling selective correction while preserving biological complexity.

542 Validation across RNA002 and RNA004 with Dorado basecalling shows adapter-
543 bridged chimeras persist systematically despite chemistry improvements [? ?].
544 RNA004 fine-tuning achieved 3-4% additional improvement beyond zero-shot RNA002
545 model performance (Extended Data Fig. ??), with both models preserving cDNA-
546 supported events. This demonstrates learned probabilistic patterns enable robust
547 generalization without protocol-specific recalibration—critical as technologies evolve.

548 Representative cases illustrate capabilities and limitations. Extended Data Fig. ??
549 shows successful detection with characteristic poly-A patterns in the upstream and low
550 base quality. Challenging scenarios (Extended Data Fig. ??, Extended Data Fig. ??)
551 include incomplete 3' end detection in multi-adapter reads and partial detection of
552

degraded sequences. Solutions include combining Dorado (3' end adapters) with DeepChopper (internal adapters) (Extended Data Fig. ??) and post-processing length filtering (default 20 bp minimum). Despite edge cases, DeepChopper achieves 62-91% artifact reduction while preserving cDNA-supported biological events. 553
554
555
556
DeepChopper demonstrates how GLM address long-read sequencing challenges 557
resisting conventional approaches. Long-range context, error-tolerant learning, single- 558
nucleotide precision, and quality-aware predictions enable detecting corrupted internal 559
adapters that elude exact-matching, k-mer, or alignment methods. Recovery rather 560
than removal represents a paradigm shift from filtering to correction, preserving valuable 561
data while enhancing accuracy. Future directions include extended context for 562
ultra-long transcripts, alternative approaches for RT-mediated cDNA-seq chimeras 563
lacking internal adapters, and expansion to additional platforms. This advance enables 564
confident biological interpretation of dRNA-seq data—from isoform discovery to gene 565
fusion detection—strengthening transcriptomic research in complex biological systems 566
where accurate transcript characterization is essential for understanding gene 567
regulation and cellular function. 568
569

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 confluence was regularly monitored and subculturing was performed before reaching 80% confluence 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™ 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.

599 **Nanopore sequencing**

600 We performed nanopore [dRNA-seq](#) sequencing of the enriched mRNA using two dif-
601 ferent sets: the RNA002 kits with R9.4.1 flow cells and the RNA004 kits with R10.4.1
602 flow cells. The decision to incorporate the RNA004 kit, a newly released option, was
603 driven by our intention to test its capabilities in conjunction with our DeepChop-
604 per tool to optimize data quality and sequencing efficiency. For the RNA002 library,
605 1 µg of poly(A)+ RNA was used as input for library preparation using the Direct
606 RNA Sequencing Kit (SQK-RNA002, [ONT](#)) following the manufacturer's instructions.
607 Nanopore [dRNA-seq](#) employs a [reverse transcriptase adapter \(RTA\)](#) that typically
608 binds to the poly(A) tails of [messenger RNA \(mRNA\)](#); subsequently, a sequencing
609 adapter is ligated to the [RTA](#), which guides the mRNA through the nanopore for
610 sequencing. The prepared library was loaded onto four MinION R9.4 flow cells (FLO-
611 MIN106) and sequenced for 48 hours using the Oxford Nanopore MinION device. For
612 the RNA004 library, 300 ng of poly(A)+ RNA was used as input for library prepara-
613 tion using the Direct RNA Sequencing Kit (SQK-RNA004, [ONT](#)) according to the
614 protocol of the manufacturer. The library was then loaded onto a PromethION RNA
615 Flow Cell (FLO-PRO004RA) and sequenced on the Oxford Nanopore PromethION
616 device for 72 hours.

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

633 **Training data preparation**

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

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**.
645
646
647
648
649
650
651
652
653
654
655
656
657
658

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.
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673

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.
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690

At its core, DeepChopper employs HyenaDNA [?] 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.
680
681
682
683
684
685
686
687
688
689
690

691 across two classes: adapter and non-adapter. For a given nucleotide position with out-
692 put logits z_1 and z_2 (corresponding to adapter and non-adapter classes), the softmax
693 function computes class probabilities as:

694

695
$$P(y_i = c) = \frac{e^{z_c}}{\sum_{j=1}^2 e^{z_j}}$$

697 where $P(y_i = c)$ represents the probability that nucleotide position i belongs to
698 class c . The final classification decision is based on the class with the higher probability
699 score. In other words, a nucleotide is classified as an adapter if $P(y_i = \text{adapter}) >$
700 $P(y_i = \text{non-adapter})$. Hence, a threshold of 0.5 is applied implicitly.

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

704

705 Model training

706

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

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

722
$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

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

726

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

728

729
$$\mathcal{L}_{\text{BatchBCE}} = \frac{1}{B} \sum_{j=1}^B \mathcal{L}_{\text{BCE}}(\mathbf{y}_j, \hat{\mathbf{y}}_j)$$

730

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

734

735

The model evaluation metrics included accuracy, precision, recall and the F1 score, calculated using the following equations: 737
738
739

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

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) [?]. To identify the best hyperparameter configuration, the Hydra (v1.3.2) [?] framework was used. 740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782

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.

RNA004 chemistry fine-tuning

To optimize DeepChopper's performance for the current RNA004 sequencing chemistry, we fine-tuned the model using VCaP RNA004 data. Following the same data preparation methodology described in Training data preparation, we generated synthetic training data from VCaP RNA004 base-called reads. The fine-tuning dataset consisted of 300,000 reads, split into training (70%, 210,000 reads), validation (20%, 60,000 reads), and test (10%, 30,000 reads) sets, with positive examples containing adapter sequences (1:1 ratio of 3' end and internal adapters) and negative examples without adapters in a 9:1 ratio.

The fine-tuning process maintained the same model architecture and hyperparameters as the original RNA002 training but allowed the model to adapt to RNA004-specific error profiles and signal characteristics. Training was performed on two NVIDIA A100 GPUs using the optimization settings described in Model training, with early stopping based on validation F1 score. Fine-tuning required approximately 9 hours of total training time. Performance evaluation compared three conditions: (1) Dorado basecalling with adapter trimming (baseline), (2) Dorado followed by the original RNA002-trained DeepChopper, and (3) Dorado followed by the RNA004-fine-tuned DeepChopper. Chimeric alignments were evaluated against matched cDNA sequencing data to distinguish biological events from artifacts.

783 **Sliding window approach for prediction refinement**

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

790 The refinement operates on the initial (raw) predictions from the model rather than
791 iteratively refining predictions at each step—a design choice driven by the require-
792 ment for precise adapter boundary detection at single-nucleotide resolution. For each
793 base position i , we define a window W_i of size w (default: 21 bp) centered at position
794 i . The final classification y_i is determined by majority vote of all predictions within
795 the window:

797

$$798 \quad y_i = \begin{cases} 1 & \text{if } \sum_{j=i-k}^{i+k} p_j > \frac{w}{2} \\ 0 & \text{otherwise} \end{cases}$$

799

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

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

808

809 **Post-processing and filtering**

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

- 812
- 813 1. A predicted adapter sequence must be at least 13 nucleotides long. Sequences
814 shorter than this length threshold are not considered valid adapters.
 - 815 2. If a read contains more than four adapter sequences, the entire read sequence is
816 retained without any adapter removal.
 - 817 3. For reads containing four or fewer adapter sequences, the identified adapters are
818 removed and the read is divided into smaller segments.
 - 819 4. Any segments resulting from this process that are less than 20 nucleotides long
820 are discarded.

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

825 All filtering thresholds, including minimum segment length, and maximum adapter
826 count per read, are configurable via command-line parameters in the DeepChopper
827

828

implementation, allowing users to tailor these settings to dataset-specific requirements
or experimental conditions. 829
830
831

BLAT identity calculation 832

The accuracy of DeepChopper in detecting adapter sequences was evaluated by aligning
the identified sequences to the human reference genome using [BLAT](#) [?]. A [BLAT](#)
identity score was defined as the ratio of matched bases to the total sequence length: 833
834
835
836

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

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) [?] 837
838
839
840
841
842
843
844
845
846
847

Computational benchmarks 848

All benchmarks were conducted in triplicate using btop
(<https://github.com/aristocratos/btop>, v1.4.0) and nvttop
(<https://github.com/Syllo/nvttop>, 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. 849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874

Transcript length distribution analysis 855

To assess the biological appropriateness of DeepChopper's 32 kb input limit, we analyzed
transcript length distributions from two sources: theoretical annotations and
empirical sequencing data. For theoretical analysis, all protein-coding transcripts were
extracted from Ensembl human genome annotations (GRCh38.115, released July 11,
2025). We calculated median length, 95th percentile, 99th percentile, and the fraction
of transcripts exceeding 32 kb using Python. For empirical analysis, read length
distributions were examined across seven dRNA-seq datasets: A549, MCF7, HCT116,
K562, HepG2, VCaP RNA002, and VCaP RNA004. For each dataset, we calculated
comprehensive statistics including minimum, maximum, mean, standard deviation,
quartiles (Q1, Q2, Q3), and high percentiles (P90, P95, P99). The number and per-
centage of reads exceeding the 32 kb threshold were specifically quantified to assess
the practical impact of this limitation. All read length statistics were computed from
Dorado (v0.5.2) base-called reads with the trim option enabled. 856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874

Validation of chimera artifact reduction 870

Cross-platform validation of [dRNA-seq](#) chimera artifacts identified by DeepChopper
was conducted leveraging [ONT](#) direct cDNA sequencing and additional cDNA-based 871
872
873
874

875 sequencing platforms. Direct cDNA sequencing validation was performed using six
876 cancer cell lines, including the VCaP dataset generated in this study and five pub-
877 lished datasets (A549, K562, HepG2, MCF7, and HCT116) obtained from the **SG-NEx**
878 project [?]. The direct cDNA data in FAST5 format were converted to POD5 format
879 using the POD5 conversion tool (<https://pod5-file-format.readthedocs.io>). Subse-
880 quently, FASTQ files were generated using Dorado (v0.5.2) [?] with adapter trimming
881 enabled (--trim adapters) and the “dna_r9.4.1_e8_hac@v3.3” model. The reads were
882 then processed using Pychopper (<https://github.com/epi2me-labs/pychopper>, v2.7.9)
883 and Cutadapt (v4.2) [?] according to a published protocol [?]. The oriented reads
884 were aligned to the human reference genome (GRCh38) using minimap2 (v2.24) [?]
885 with optimized parameters (-ax splice -uf -k14) for spliced alignment. The resulting
886 SAM files were then converted to BAM format, indexed, and sorted using SAMtools
887 (v1.19.2) [?].

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

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

911 To empirically assess the applicability of alignment-based artifact detection meth-
912 ods to **dRNA-seq**, we compared DeepChopper with Breakinator (v1.0) [?] using VCaP
913 RNA002 data. Three processing pipelines were evaluated: Dorado basecalling with
914 trim option (baseline), Dorado with trim followed by DeepChopper, and Dorado with
915 trim followed by Breakinator. All pipelines started with identical Dorado-trimmed
916 reads to enable direct comparison. Breakinator was applied using default parame-
917 ters as specified in its documentation [?]. Chimeric alignments from each pipeline
918 were classified as cDNA-supported (validated by matched direct cDNA sequencing) or
919

920

unsupported (likely artifacts) using the validation framework described above. Support rates were calculated as the proportion of chimeric alignments corroborated by cDNA-seq data.

Gene expression analysis and transcript classification

Gene expression levels from [dRNA-seq](#) were quantified using IsoQuant (v3.1.2) [?], 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 [?].

Gene fusion identification and visualization

For [ONT dRNA-seq](#) data, gene fusions were identified using FusionSeeker (v1.0.1) [?] 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 [?] under SRA accession SRX5417211. Raw reads were mapped to the hg38 reference genome using STAR (v2.7.11) [?], and gene fusion events were detected with Arriba (v2.4.0) [?]. The gene structure of the RPS29-COX8A fusion was visualized using GSDS (v2.0) [?]. Base quality scores were generated with a custom Python script, and ion current signals were visualized using Squigualiser (v0.6.3) [?]. The circos plot for gene fusion events was visualized using chimeraviz (v1.30.0) [?].

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 [?].

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. Source data are provided with this paper.

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,

967 Linux, and macOS to ensure easy cross-platform installation. An interactive demo
968 is available on Hugging Face (<https://huggingface.co/spaces/yangliz5/deepchopper>),
969 allowing users to test DeepChopper's functionality without local installation. For
970 large-scale analyses, we recommend using DeepChopper on systems with **GPU** accel-
971 eration. Detailed system requirements and optimization guidelines are available in the
972 repository's documentation.

973 **Acknowledgements.** This project was supported in part by NIH grants
974 R35GM142441 and R01CA259388 awarded to RY, and NIH grants R01CA256741,
975 R01CA278832, and R01CA285684 awarded to QC.
976

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

981 **Conflict of interests.** RY has served as an advisor/consultant for Tempus AI, Inc.
982 This relationship is unrelated to and did not influence the research presented in this
983 study.
984

985 **Acronyms**
986

987 **ATCC** American Type Culture Collection [13](#)
988

989 **BLAT** BLAST-like alignment tool [5](#), [7](#), [9](#), [19](#), [31](#)
990

991 **CCLE** Cancer Cell Line Encyclopedia [21](#)
992

993 **DAVID** Database for Annotation, Visualization, and Integrated Discovery [21](#)

994 **dRNA-seq** direct RNA sequencing [1](#), [2](#), [4–15](#), [18](#), [20](#), [21](#), [27](#), [29–34](#)

995

996 **FSM** Full splice match [11](#), [33](#)

997

998 **GEO** Gene Expression Omnibus [21](#)

999 **GLM** Genomic Language Model [2](#), [4](#), [12](#), [13](#)

1000 **GO** Gene Ontology [10](#), [11](#), [21](#), [34](#)

1001 **GPU** Graphics Processing Unit [16](#), [21](#), [22](#)

1002

1003 **HPC** High Performance Computing [16](#), [21](#)

1004 **ISM** Incomplete splice match [11](#), [33](#)
1005

1006 **LCGLM** long-context genomic language model [3](#)

1007 **LLM** Large Language Model [2](#)

1008 **LRGASP** Long-read RNA-Seq Genome Annotation Assessment Project [8](#), [20](#)

1009

1010 **MLP** multilayer perceptron [3](#), [15](#)

1011 **mRNA** messenger RNA [14](#)

1012

NIC Novel in catalog 11, 33	1013
NNC Novel not in catalog 11, 33	1014
ONT Oxford Nanopore Technologies 2, 4–11, 14, 18, 20, 21, 30	1015
PacBio Pacific Biosciences 5, 8, 20, 30	1016
PCR Polymerase Chain Reaction 2, 8, 20	1017
RT Reverse Transcription 2, 12, 13	1018
RTA reverse transcriptase adapter 14	1019
SG-NEx Singapore Nanopore Expression Project 4, 6, 8, 11, 14, 20	1020
References	
[1] Garalde, D. R. <i>et al.</i> Highly parallel direct rna sequencing on an array of nanopores. <i>Nature methods</i> 15 , 201–206 (2018).	1021
[2] Jain, M., Abu-Shumays, R., Olsen, H. E. & Akeson, M. Advances in nanopore direct rna sequencing. <i>Nature methods</i> 19 , 1160–1164 (2022).	1022
[3] Smith, M. A. <i>et al.</i> Molecular barcoding of native rnas using nanopore sequencing and deep learning. <i>Genome research</i> 30 , 1345–1353 (2020).	1023
[4] Liu-Wei, W. <i>et al.</i> Sequencing accuracy and systematic errors of nanopore direct rna sequencing. <i>BMC genomics</i> 25 , 528 (2024).	1024
[5] epi2me-labs/pychopper: cdna read preprocessing. Github. URL https://github.com/epi2me-labs/pychopper .	1025
[6] Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex minion sequencing. <i>Microbial Genomics</i> 3 (2017). URL http://dx.doi.org/10.1099/mgen.0.000132 .	1026
[7] Bonenfant, Q., Noé, L. & Touzet, H. Porechop_abi: discovering unknown adapters in oxford nanopore technology sequencing reads for downstream trimming. <i>Bioinformatics Advances</i> 3 , vbac085 (2023).	1027
[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.	1028
[9] Nguyen, E. <i>et al.</i> Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. <i>Advances in neural information processing systems</i> 36 (2024).	1029
[10] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. 1512.03385.	1030
	1031
	1032
	1033
	1034
	1035
	1036
	1037
	1038
	1039
	1040
	1041
	1042
	1043
	1044
	1045
	1046
	1047
	1048
	1049
	1050
	1051
	1052
	1053
	1054
	1055
	1056
	1057
	1058

- 1059 [11] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained bidirectional
1060 encoder representations from transformers model for DNA-language in genome
1061 **37**, 2112–2120.
- 1062
- 1063 [12] Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for
1064 multi-species genome. [2306.15006](https://doi.org/10.1101/2306.15006).
- 1065
- 1066 [13] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust
1067 foundation models for human genomics. *Nature Methods* 1–11 (2024).
- 1068
- 1069 [14] Lopes, I., Altab, G., Raina, P. & De Magalhães, J. P. Gene size matters: an
1070 analysis of gene length in the human genome. *Frontiers in Genetics* **12**, 559998
1071 (2021).
- 1072
- 1073 [15] Workman, R. E. *et al.* Nanopore native rna sequencing of a human poly (a)
1074 transcriptome. *Nature methods* **16**, 1297–1305 (2019).
- 1075
- 1076 [16] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale
1077 with evo. *Science* **386**, eado9336 (2024). URL <https://www.science.org/doi/abs/10.1126/science.ad09336>.
- 1078
- 1079 [17] Chen, Y. *et al.* A systematic benchmark of nanopore long read rna sequencing
1080 for transcript level analysis in human cell lines. *BioRxiv* 2021–04 (2021).
- 1081
- 1082 [18] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023).
- 1083
- 1084 [19] PLC., O. N. Chemistry Technical Document (CHTD.500.v1.revAQ_07Jul2016)
1085 (2017). URL <https://nanoporetech.com/document/chemistry-technical-document>.
- 1086
- 1087 [20] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
1088 **34**, 3094–3100 (2018).
- 1089
- 1090 [21] Kent, W. J. Blat—the blast-like alignment tool. *Genome research* **12**, 656–664
1091 (2002).
- 1092
- 1093 [22] Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read rna-seq methods
1094 for transcript identification and quantification. *Nature methods* 1–15 (2024).
- 1095
- 1096 [23] Hewel, C. *et al.* Direct rna sequencing enables improved transcriptome assess-
1097 ment and tracking of rna modifications for medical applications. *bioRxiv* 2024–07
1098 (2024).
- 1099
- 1100 [24] Prjibelski, A. D. *et al.* Accurate isoform discovery with isoquant using long reads.
1101 *Nature Biotechnology* **41**, 915–918 (2023).
- 1102
- 1103 [25] Chen, Y. *et al.* Gene fusion detection and characterization in long-read cancer
1104 transcriptome sequencing data with fusionseeker. *Cancer research* **83**, 28–33

- (2023). 1105
- [26] Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from rna sequencing data. *Genome research* **31**, 448–460 (2021). 1106
- [27] Li, H. *et al.* The sequence alignment/map format and samtools. *bioinformatics* **25**, 2078–2079 (2009). 1107
- [28] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 1108
- [29] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019). 1109
- [30] Yadan, O. Hydra - a framework for elegantly configuring complex applications. Github (2019). URL <https://github.com/facebookresearch/hydra>. 1110
- [31] Li, Y. & Yang, R. PxBLAT: an efficient python binding library for BLAT. *BMC Bioinf.* **25**, 1–8 (2024). 1111
- [32] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011). 1112
- [33] Grünberger, F., Ferreira-Cerca, S. & Grohmann, D. Nanopore sequencing of rna and cdna molecules in escherichia coli. *Rna* **28**, 400–417 (2022). 1113
- [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). 1114
- [35] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). 1115
- [36] Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013). 1116
- [37] Hu, B. *et al.* Gsds 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**, 1296–1297 (2015). 1117
- [38] Samarakoon, H. *et al.* Interactive visualisation of raw nanopore signal data with squigualiser. *Biorxiv* 2024–02 (2024). 1118
- [39] Lågstad, S. *et al.* chimeraviz: a tool for visualizing chimeric rna. *Bioinformatics* **33**, 2954–2956 (2017). 1119
- [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). 1120

1151 **Extended data**

1152

1153

1154 **Extended Data Table 1** Summary of Adapter Trimming Tools for analyzing
1155 dRNA-seq data

1156

1157 Adapter trimming tool	1158 dRNA-seq 1159 terminal 1160 adapter 1161 trimming	1162 dRNA-seq 1163 internal 1164 adapter 1165 trimming	1166 Trimming existing 1167 dRNA-seq datasets (post-basecalling)
Porechop [?]	×	×	×
Porechop_ABI [?]	×	×	×
Pychopper [?]	×	×	×
Dorado [?]	✓	×	×
DeepChopper	✓	✓	✓

1166 ✓ indicates the tool supports this functionality; × indicates the tool does not support
1167 this functionality.

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187 **Extended Data Fig. 1 Distribution of transcript length for protein-coding genes.** Analysis
1188 of all protein-coding transcripts from Ensembl GRCh38.115 (released July 2025) shows that >99.99%
1189 of transcripts are below the 32 kb threshold (marked with vertical dashed line). The distribution is
1190 highly skewed toward shorter transcripts, with median length of ~2.7 kb.

1191

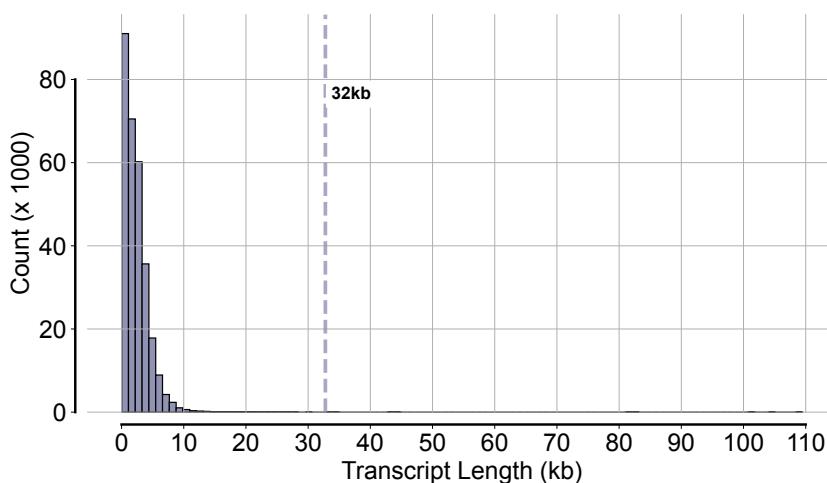
1192

1193

1194

1195

1196



1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242

Extended Data Table 2 Read Length Statistics by Sample

Sample	Reads (M)	Min (bp)	Max (bp)	Mean (bp)	Std Dev (bp)	Q1 (bp)	Q2 (bp)	Q3 (bp)	P90 (bp)	P95 (bp)	P99 (bp)	Reads ≥32kb	% ≥32kb
A549	1.70	5	16,246	907	805	383	700	1,223	1,904	2,440	3,829	0	0
MCF7	3.04	5	28,802	715	623	316	546	911	1,475	1,863	3,052	0	0
HCT116	4.70	5	21,656	889	795	374	669	1,193	1,871	2,431	3,793	0	0
K562	3.06	2	58,395	683	555	319	556	892	1,393	1,736	2,619	2	0
HepG2	1.80	2	46,077	1,148	974	497	864	1,544	2,317	3,025	4,665	1	0
VCaP RNA002	9.18	5	77,474	994	901	462	697	1,279	2,092	2,826	4,399	1	0
VCaP RNA004	11.72	5	225,798	995	971	483	695	1,224	2,025	2,784	4,474	379	0.0032

Q1, Q2, Q3 represent 25th, 50th, and 75th percentiles. P90, P95, P99 represent 90th, 95th, and 99th percentiles. All reads were basecalled using Dorado (v0.5.2) with trim option. VCaP RNA002 and RNA004 represent matched chemistry comparison.

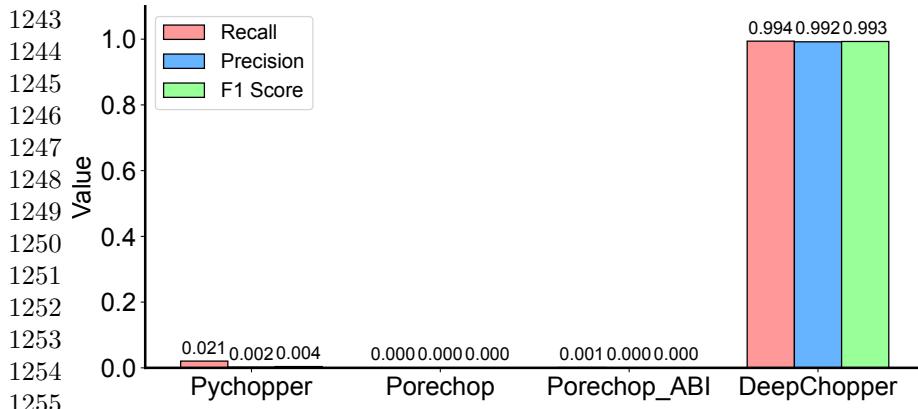
Extended Data Table 3 Ablation Study Results for Quality Block

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

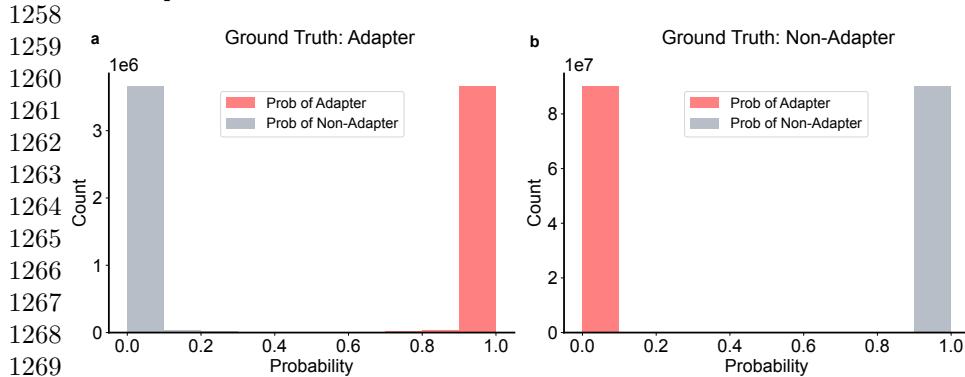
Extended Data Table 4 Internal Adapter Prevalence Across Datasets

Sample	All Reads			Chimeric Reads Only		
	With Internal Adapters (A)	Total (B)	% (A/B)	With Internal Adapters (C)	Total (D)	% (C/D)
A549	15,690	1,703,697	0.92	10,553	12,803	82.43
MCF7	20,340	3,039,468	0.67	11,115	17,646	63.00
HCT116	57,122	4,697,299	1.22	37,823	46,800	80.81
K562	29,436	3,061,722	0.96	19,289	23,214	83.09
HepG2	22,530	1,797,922	1.25	14,331	16,921	84.69
VCaP RNA002	148,452	9,177,422	1.62	98,878	107,265	92.18
VCaP RNA004	38,878	11,714,520	0.33	6,891	29,144	23.65

Total reads from Dorado with trim. Internal adapters detected by DeepChopper after Dorado processing. VCaP RNA002 and RNA004 demonstrate that adapter-bridged chimeras persist across chemistries.



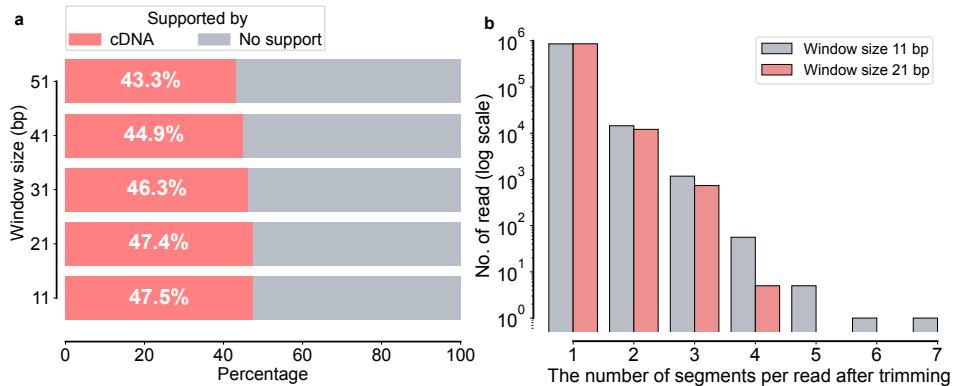
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256 **Extended Data Fig. 2 Performance evaluation in a held-out test dataset ($N = 60,000$)**
1257 showing Recall, Precision, and F1 values for DeepChopper, Pychopper, Porechop, and
1258 Porechop_ABI.



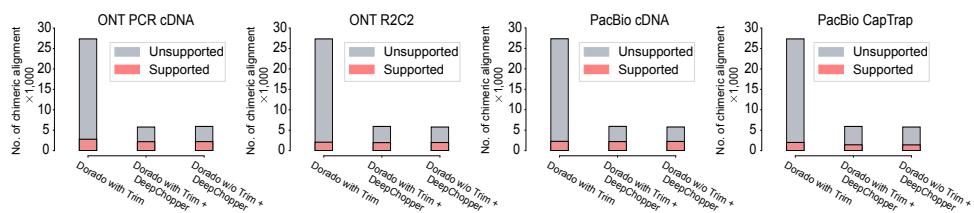
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270 **Extended Data Fig. 3 Prediction probability distributions of DeepChopper for the held-**
1271 **out test dataset ($N = 60,000$).** (a) Distribution of prediction probabilities for sequences with
1272 ground truth adapter classification. Red bars represent the probability of adapter prediction, while
1273 gray bars show the probability of non-adapter prediction. The count (y-axis) is shown in millions of
1274 sequences (10^6 scale). (b) Distribution of prediction probabilities for sequences with ground truth
1275 non-adapter classification. Red bars indicate the probability of adapter prediction, while gray bars
1276 show the probability of non-adapter prediction. The count (y-axis) is shown in tens of millions of
1277 sequences (10^7 scale). Both distributions demonstrate strong polarization toward correct classification
1278 probabilities, indicating the model's high confidence in distinguishing between adapter and non-
1279 adapter sequences.

1278
1279 **Extended Data Fig. 4 Computational performance metrics across different data sizes.**
1280 (a) Runtime analysis showing processing time requirements for different pipeline stages (FASTQ
1281 Conversion, Prediction, Post-Processing) and total runtime across five data sizes: subsampled VCaP
1282 datasets (0.1M, 0.5M, 1M reads), full VCaP dRNA-seq dataset (9M reads), and merged large-scale
1283 dataset (23M reads combining A549, HCT116, HepG2, K562, and MCF7). Runtime scales near-
1284 linearly with data size. As data size increases, prediction time becomes the dominant component,
1285 requiring approximately 5 hours for the 9M dataset and 10.6 hours for the 23M dataset. (b) Memory
1286 usage comparison between CPU and GPU implementations across the same data sizes. The predic-
1287 tion stage shows consistently higher memory requirements. CPU memory usage ranges from 70-93
1288 GB and GPU memory from 34-56 GB across larger datasets, with stable memory footprint indicating
1289 no fundamental barriers to processing substantially larger datasets. All measurements include error
1290 bars representing standard deviation from three technical replicates.

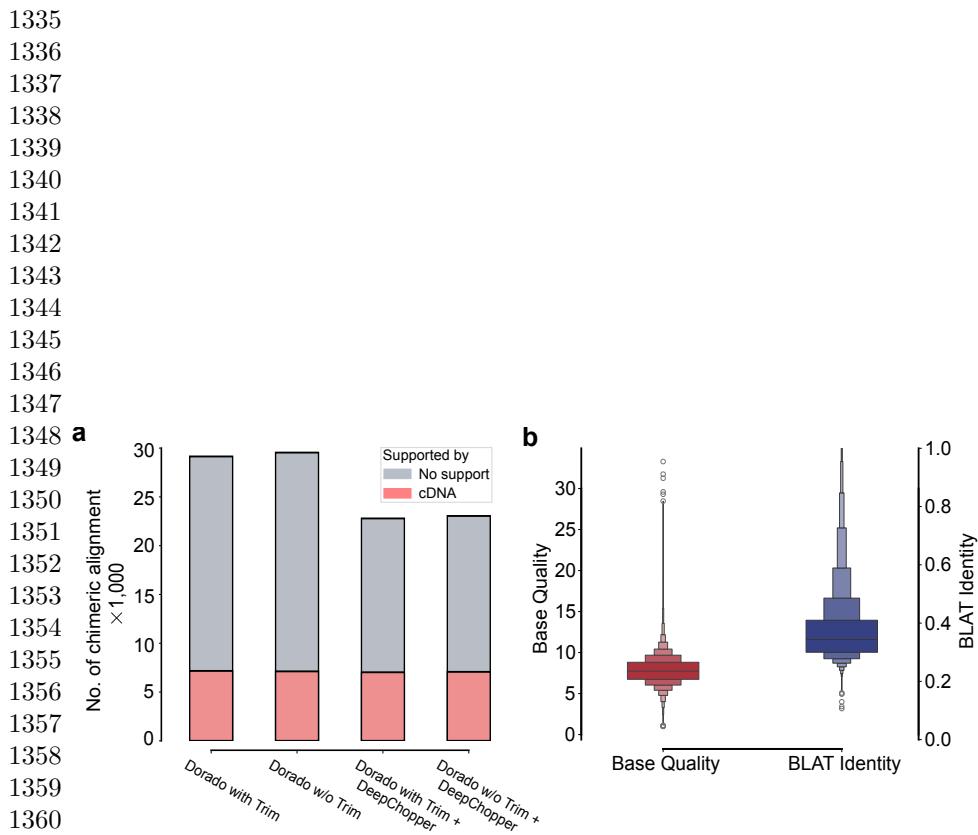
1289
 1290
 1291
 1292
 1293
 1294
 1295
 1296
 1297
 1298
 1299
 1300
 1301
 1302
 1303
 1304
 1305
 1306
 1307
 1308
 1309
 1310
 1311
 1312
 1313
 1314
 1315
 1316
 1317
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334

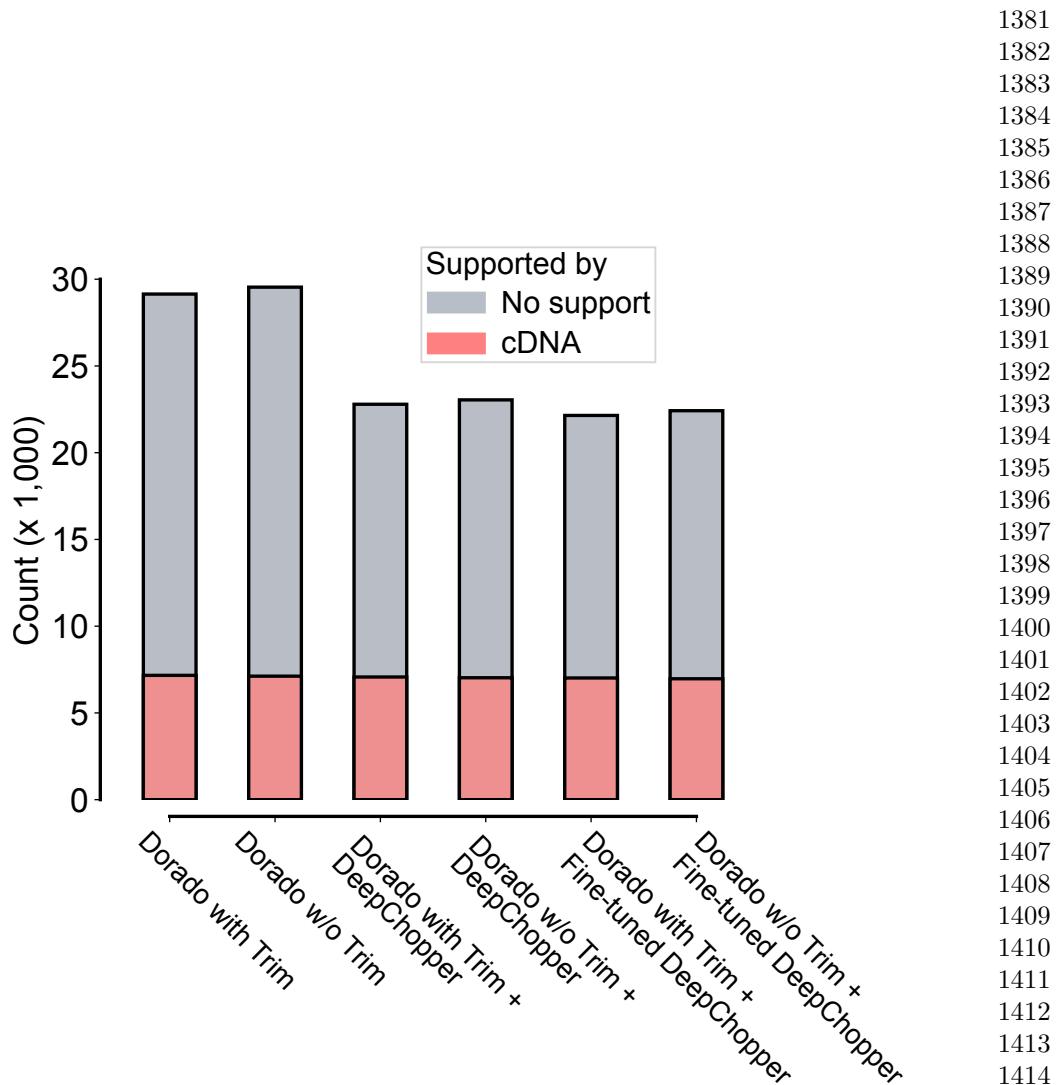


Extended Data Fig. 5 Effect of window size on chimeric alignment detection and read fragmentation. (a) Analysis of different sliding window sizes (11, 21, 31, 41, and 51 nucleotides) showing the percentage of cDNA-supported chimeric alignments (red bars) in VCaP. Higher percentages indicate better support. (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.



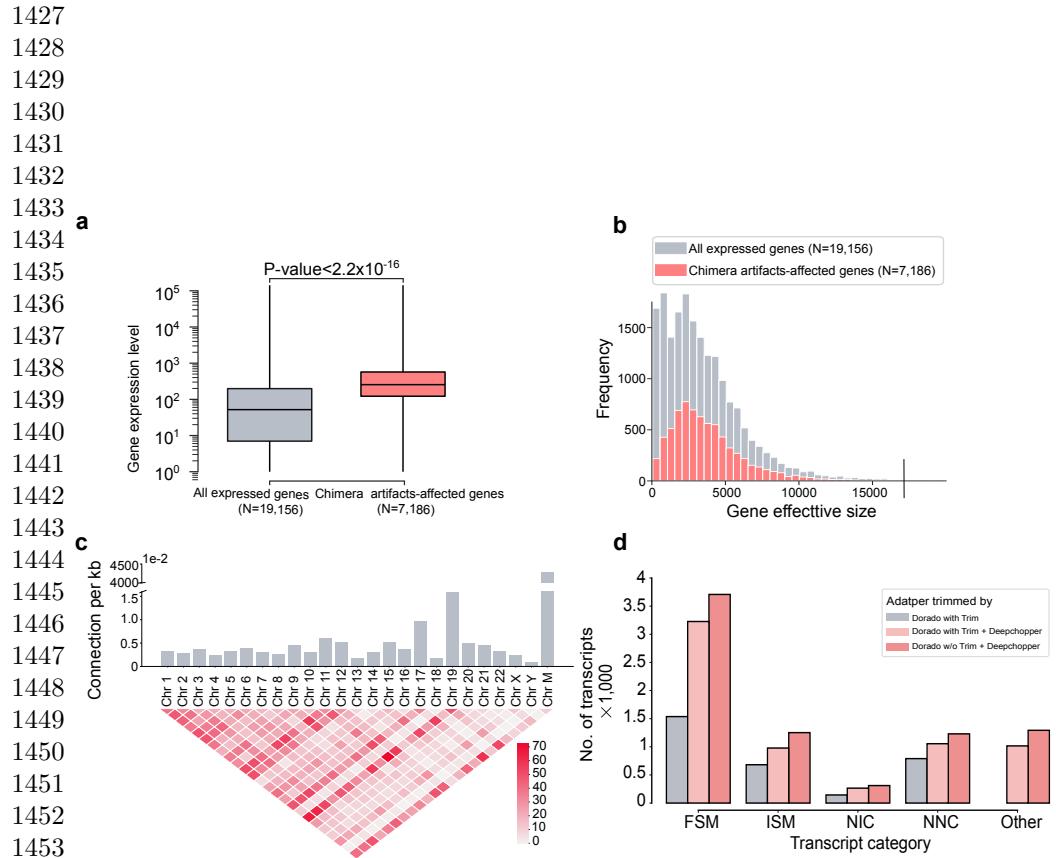
Extended Data Fig. 6 Chimeric alignments from dRNA-seq of the F121-9 cell line (mouse), evaluated for support using additional ONT and PacBio sequencing data with different protocols. DeepChopper-involved methods reduce unsupported chimeric alignments across all methods compared to Dorado with adapter trimming. The bar colors indicate chimeric alignments supported by additional sequencing data (red) and those lacking support (grey).



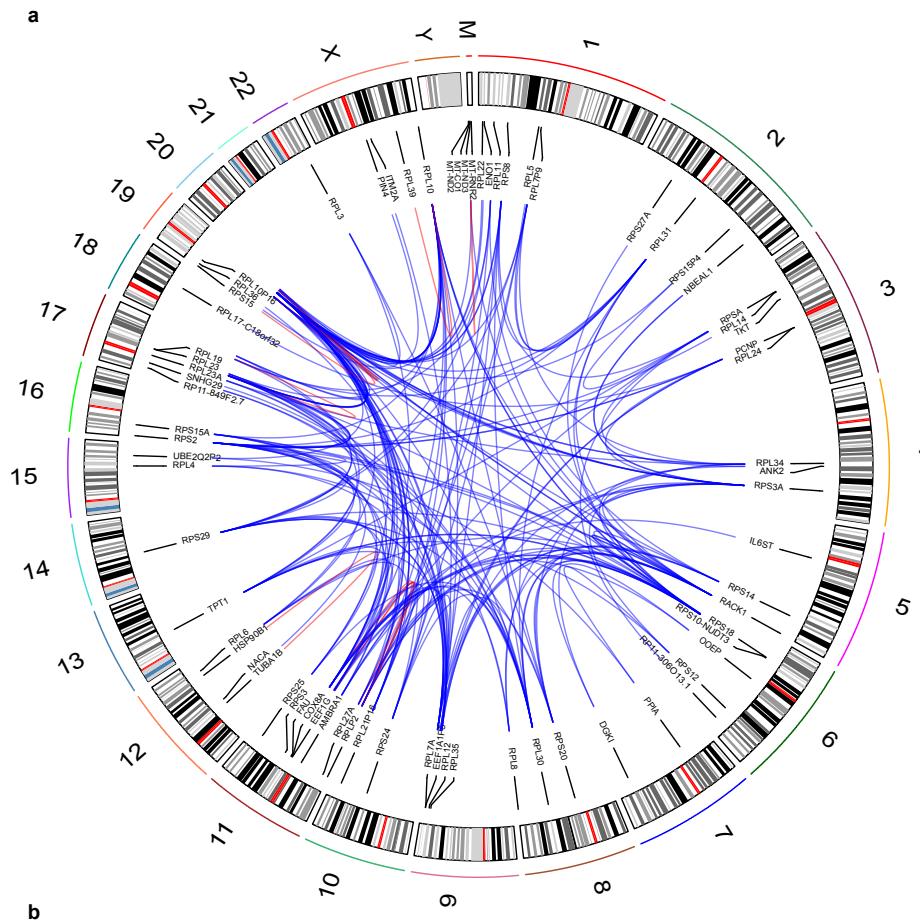


Extended Data Fig. 8 Performance comparison of original and fine-tuned DeepChopper on RNA004 data. Number of chimeric alignments (in thousands) identified in VCaP RNA004 dRNA-seq processed under six conditions: Dorado basecalling with and without adapter trimming, followed by original DeepChopper, and followed by fine-tuned DeepChopper. The bar colors indicate chimeric alignments supported by cDNA sequencing (red) and those lacking support (grey).

1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426



Extended Data Fig. 9 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 < 2.2×10^{-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.

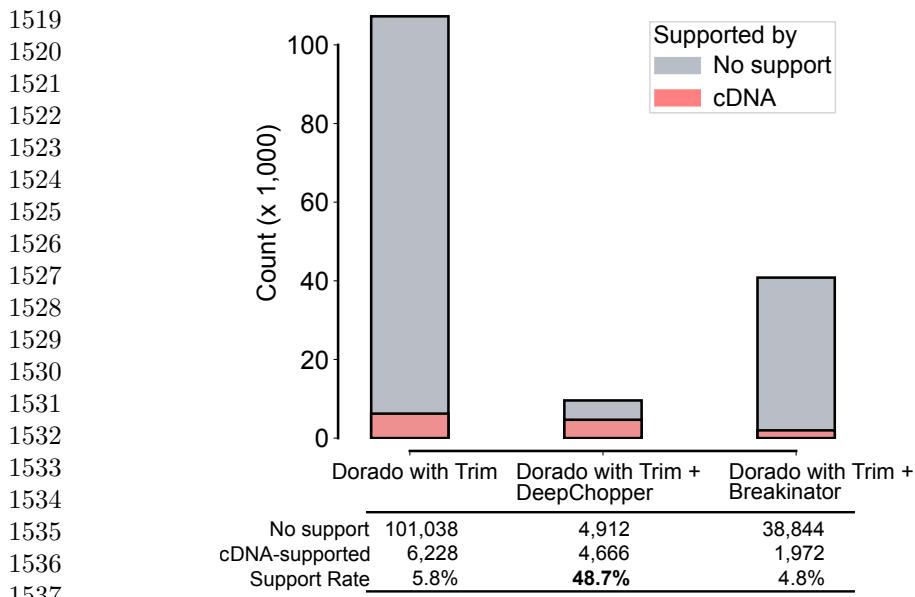


b

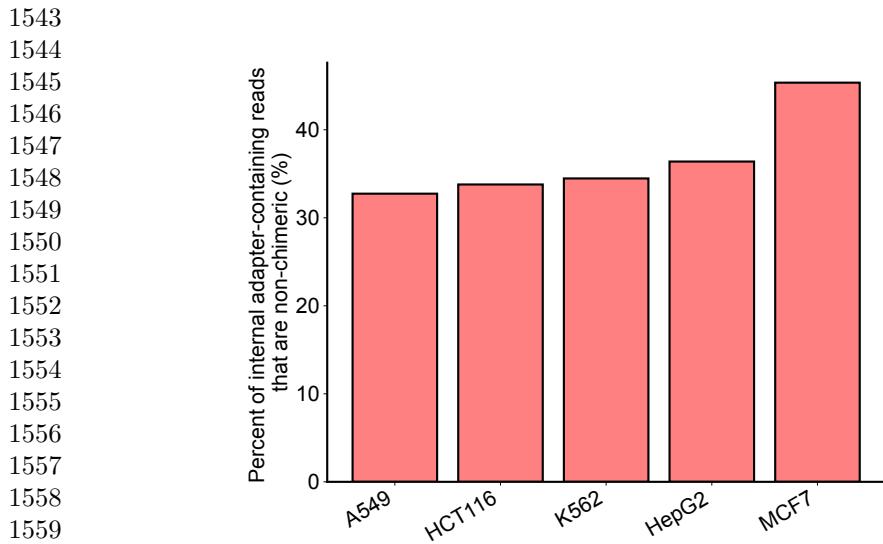
Sample	GO Term	Genes	P-value
A549	Cytoplasmic translation	RPL34, RPS21	1.467×10^{-2}
	Translation	RPL34, RPS21	3.040×10^{-2}
HepG2	Translation	EEF1A1, RPS19, RPL13, RPS12	2.172×10^{-4}
	Intracellular iron ion homeostasis	MT-RNR2, FTH1, FTL	7.332×10^{-4}
	Cytoplasmic translation	RPS19, RPL13, RPS12	1.528×10^{-3}
	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}
HCT116	Ribosomal small subunit biogenesis	RPS19, RPS12	4.526×10^{-2}
	Regulation of translation	RPS3, RPL38, GAPDH	6.151×10^{-3}
VCaP (SQK-RNA004 kit)	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, RPL12, RPS29, RPS3, RPL14, RPLP2, RPL13, RPS20, RPL38, RPS2, RPL28	1.566×10^{-22}
	Cytoplasmic translation	RPS16, RPLP2, RPL29	7.15×10^{-5}
	Translation	RPS16, EEF1A1P5, RPLP2, RPL29	1.06×10^{-6}

Extended Data Fig. 10 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 fusion events. 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.

1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518



1538 Extended Data Fig. 11 **Comparison of chimeric alignment reduction strategies in VCaP**
1539 **RNA002 dRNA-seq data.** Stacked bar plot showing chimeric alignments (in thousands) for three
1540 processing pipelines: Dorado with adapter trimming (baseline), Dorado with adapter trimming fol-
1541 lowed by DeepChopper, and Dorado with adapter trimming followed by Breakinator. Gray bars
1542 represent unsupported chimeric alignments (likely artifacts); pink bars represent cDNA-supported
1543 chimeric alignments (biological events).

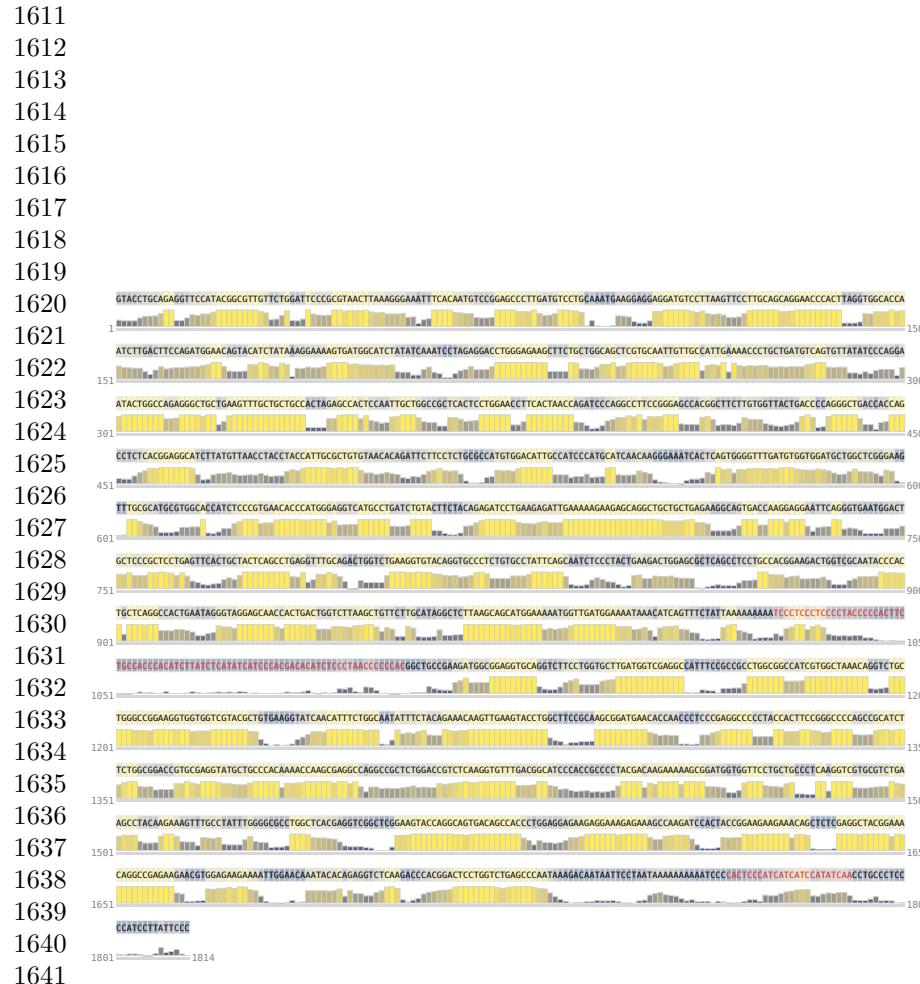


1561 Extended Data Fig. 12 **Percent of internal adapter-containing reads that are non-**
1562 **chimeric** Percentage of internal adapter-containing reads that do not produce chimeric alignments
1563 across five human cell lines (RNA002). Between 33–45% of adapter-containing reads map as single
1564 alignments or fail to map, making them invisible to chimeric alignment-based artifact detection.

figures/finals/figure5e_adapter.pdf

1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610

Extended Data Fig. 13 Representative internal adapter detection with base quality visualization. Representative read (Fig. ??e) from VCap RNA002 processed with Dorado trim followed by DeepChopper, demonstrating internal adapter detection (highlighted in red, positions 155-229) and sub-reads recovery (blue box and purple box). (a) BLAT alignment of upstream sequence (blue box, positions 1-154) to chr14 (RPS29 gene) with 0.98 identity, confirming genuine biological RNA. (b) Full read visualization showing internal adapter and sub-reads. Base quality scores (Q scores) shown as color-coded bars: yellow indicates high quality ($Q > 40$), dark blue indicates low quality ($Q < 10$). The adapter region shows characteristic poly-A sequences upstream and lower base quality compared to flanking biological sequences. The adapter region (75 bp) shows no matches found to the reference genome. Sequences before and after the adapter represent genuine biological RNA from different transcripts artificially joined during library preparation or basecalling. Read ID: 3b2292e9-43e5-4e40-87d9-ccc23897377c. (c) BLAT alignment of downstream sequence (purple box, positions 230-722) to chr11 (COX8A gene) with 0.96 identity, confirming genuine biological RNA from a different transcript.



1642 **Extended Data Fig. 14 Challenge Scenario 1: Incomplete 3' terminal adapter detection**
 1643 **in multi-adapter read.** Representative read from VCaP RNA002 processed with Dorado without
 1644 trim followed by DeepChopper (with refinement applied). The read contains both an internal adapter
 1645 (positions 1026-1105, highlighted in red, correctly detected) and a 3' end adapter that DeepChopper
 1646 failed to completely detect. Base quality scores (Q scores) shown as color-coded bars: yellow indicates
 1647 high quality ($Q > 40$), dark blue indicates low quality ($Q < 10$). The internal adapter shows char-
 1648 acteristic low quality compared to flanking biological sequences. This scenario demonstrates that when
 1649 multiple adapters are present, DeepChopper reliably detects internal adapters (its primary function)
 1650 but may incompletely detect terminal adapters. Read ID: c16c6ade-135b-4073-a1d6-5a9c6900fb2.
 1651
 1652
 1653
 1654
 1655
 1656

