

	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 [1–3].

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 [4], 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 [5]—remains limited in peer-reviewed literature. These artifacts may result
064 from ligation during library preparation or chimeric reads produced by software
065 missing the open pore signal, potentially confounding downstream analyses such
066 as transcriptome assembly, quantification, and detection of alternative splicing and
067 gene fusion events. Detecting these chimera artifacts is challenging because long-read
068 aligners often produce chimeric alignments from such artifacts that are indistin-
069 guishable from those derived from true gene fusion events. Importantly, chimeric
070 read artifacts frequently contain internal adapter sequences [4], 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 [6]. As a result, current adapter detection tools [7–9] cannot exploit
075 this feature to eliminate these adapter-bridged chimeras, leaving the issue unresolved
076 (Extended Data Table 1).

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 [10], 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

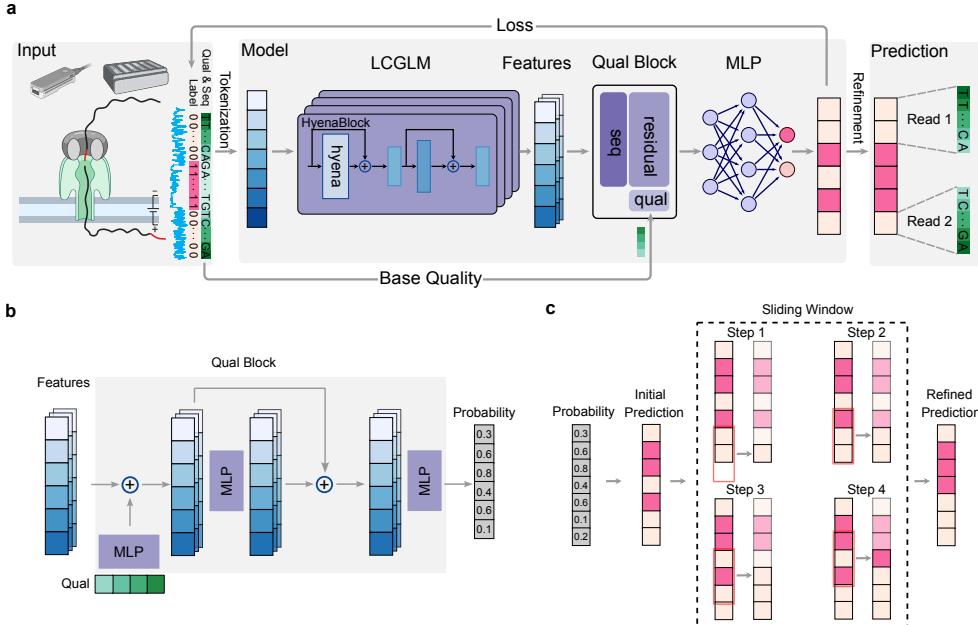


Fig. 1 DeepChopper architecture and methodology. (a) Overview of the DeepChopper model. Raw sequences are first tokenized into vectors and processed by HyenaDNA to generate embedding features. These features are integrated with base quality information in the quality block to produce per-token probability 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 [11], which excels at capturing long-range dependencies (Fig. 1a). 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 [12] (Fig. 1b, 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 3'
140 end adapters simultaneously trimmed (Fig. 1a), thereby preserving authentic biological
141 sequences and eliminating chimera artifacts.

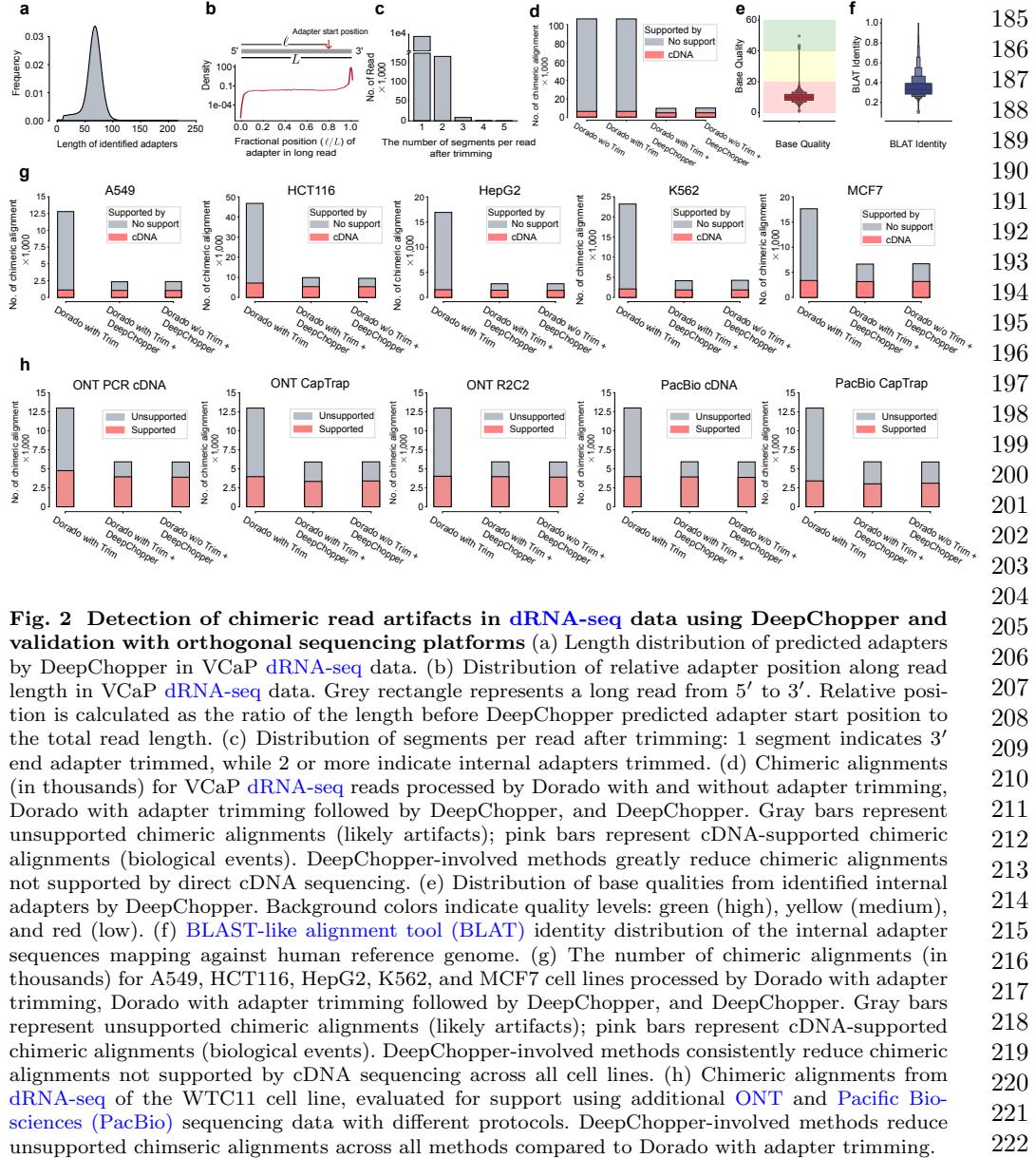
142 To further refine the prediction accuracy, DeepChopper implements a post-
143 processing stage using a sliding window and majority vote approach, as illustrated
144 in Fig. 1c. The model applies a sliding window with a stride of 1 across the read, analyzing
145 the distribution of predicted adapter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

152 Comparing to existing general-purpose GLM, DeepChopper is specifically opti-
153 mized for long-read sequence analysis at single-nucleotide resolution. This fine-grained
154 resolution provides a critical advantage for genomic analysis tasks requiring pre-
155 cise base-level predictions. While DNABERT [13] is limited to input sequences of
156 approximately 512 bp, DNABERT2 [14] to 10,000 bp, and Nucleotide Transformer [15]
157 to 6,000 bp, DeepChopper supports input lengths up to 32 kilobases—sufficient
158 to encompass most complete mRNA transcripts. This 32K nucleotide input limit
159 was selected based on both technical and biological considerations. Technically, the
160 constraint reflects the architectural design and context window limitations of the
161 underlying HyenaDNA model. Biologically, human protein-coding transcripts have
162 a median length of approximately 2.7 kb (95th percentile: ~ 8.7 kb, 99th per-
163 centile: ~ 14 kb), with over 99.97% of transcripts falling below 32 kb based on
164 current annotations (Extended Data Fig. 1). Empirical analysis of our dRNA-seq
165 datasets confirmed that only 0–0.0032% of reads exceeded this threshold across mul-
166 tiple cell lines and chemistries, with mean read lengths ranging from 683 to 1,148
167 bp (Extended Data Table 2). This extended input capacity, combined with single-
168 nucleotide tokenization, enables DeepChopper to accurately identify non-reference
169 elements such as ONT adapter sequences with base-pair precision—an essential
170 capability for detecting and recovering adapter-bridged chimeras in dRNA-seq data.

171 In addition, DeepChopper's lightweight architecture, consisting of only 4.6 million
172 parameters, makes it computationally efficient and scalable for large-scale dRNA-seq
173 analysis. This is in contrast to models like Evo [16], which require billions of parameters
174 and significantly more computational resources.

175 To train DeepChopper for identifying adapter sequences within dRNA-seq long
176 reads, we utilized data from six human cell lines: HEK293T, A549, HCT116, HepG2,
177 K562 and MCF-7 provided by the Singapore Nanopore Expression Project (SG-
178 NEx) [17] (Extended Data Table 2). We curated a training set of 480,000 long reads
179 and a validation set of 60,000 ones initially deemed free of adapters and inserted
180 putative adapter sequences, derived from the raw dRNA-seq data, into these reads
181 to create instances containing internal and 3' end adapters (See Methods for details).
182 An independent test set comprising 60,000 long reads was held out for performance
183 evaluation.

184



DeepChopper Benchmarking and Model Optimization

We conducted comprehensive benchmarking of DeepChopper against existing ONT adapter trimming tools including Pychopper [7], Porechop [8], and Porechop_ABI [9], though it should be noted that none of these existing tools were specifically designed for dRNA-seq data analysis. Performance evaluation was carried out using the synthetic testing dataset ($n = 60,000$ reads), enabling rigorous assessment of precision, recall,

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 and F1-score metrics. As shown in Extended Data Fig. 2, all existing tools demon-
232 strated negligible performance metrics when processing dRNA-seq adapter sequences,
233 indicating fundamental incompatibility with the dRNA-seq protocol. In contrast,
234 DeepChopper achieved exceptional accuracy in identifying both terminal and inter-
235 nal adapters, with recall, precision, and F1 scores consistently exceeding 0.99. These
236 results highlight DeepChopper’s unique capability to address the specific complexi-
237 ties inherent to dRNA-seq reads and underscore the critical need for purpose-built
238 solutions in this domain.

239 In addition to adapter-trimming tools, we also evaluated Breakinator [18], which
240 is designed to detect chimeric artifacts rather than adapter boundaries. Breakinator
241 was benchmarked separately using real VCaP RNA002 dRNA-seq data. As shown in
242 Extended Data Fig. 3, Breakinator reduced detected artifacts by 62% but simultane-
243 ously decreased cDNA support to 4.8% (baseline 5.8%), reflecting substantial loss of
244 biologically supported signal. In contrast, DeepChopper reduced artifacts by 91% while
245 increasing cDNA support to 49%, demonstrating selective removal of true artifacts
246 while preserving biological transcript complexity. A substantial fraction of dRNA-seq
247 artifacts do not appear as chimeric alignments or fail to map (Extended Data Fig. 4),
248 making them inherently undetectable by alignment-rule-based chimera callers.

249 To further evaluate DeepChopper’s classification accuracy, we assessed its perfor-
250 mance at single-nucleotide resolution. As shown in Extended Data Table 3, the model
251 exhibited sharply bimodal probability distributions for base-level classification, strati-
252 fied by ground truth labels. For true adapter bases (Extended Data Fig. 5a), the model
253 consistently assigned high probabilities (> 0.8) to the adapter class while suppress-
254 ing probabilities for the non-adapter class (< 0.2). Conversely, for true non-adapter
255 bases (Extended Data Fig. 5b), the model reliably predicted high probabilities for the
256 non-adapter class and low probabilities for the adapter class. The minimal presence
257 of intermediate probability values suggests that DeepChopper makes highly confident
258 predictions with low ambiguity between classes. This decisive classification behavior
259 reflects the model’s robust feature representation, well-calibrated decision boundaries,
260 and strong generalization to previously unseen data.

261 We further conducted an ablation experiment to assess the contribution of the qual-
262 ity block component in the model architecture. As shown in Extended Data Table 3,
263 inclusion of the quality block led to a marked improvement in performance, with the
264 F1 score increasing from 0.97 to 0.99. This module leverages per-base quality scores to
265 better distinguish genuine adapter sequences from similar motifs that may occur natu-
266 rally within reads. The enhanced performance suggests that incorporating sequencing
267 quality information enables the model to more effectively filter out spurious signal
268 and improve classification robustness. This experiment was performed using the same
269 training methodology as previous evaluations but with a newly generated, focused
270 dataset of 100,000 reads (See Methods for details).

271

272 Chimeric Read Artifact Detection in Cancer dRNA-seq Data

273

274 To assess DeepChopper’s ability to detect chimera artifacts in real data, we generated
275 an independent dRNA-seq dataset using the prostate cancer VCaP cell line, which
276

was excluded from model training. This dataset provides a robust framework for evaluating chimera artifacts in genuine dRNA-seq samples, ensuring that DeepChopper's performance generalizes beyond the training data. We conducted dRNA-seq of VCaP cells using ONT 's SQK-RNA002 chemistry, consistent with that used in the SG-NEx project. Using a MinION sequencer with four R9.4 flow cells, we generated 9,177,639 long reads in FASTQ format, with base-calling performed using SG-NEx 's Dorado software [5].	277 278 279 280 281 282 283
DeepChopper processes input data through a three-stage pipeline: (1) FASTQ-to-Parquet conversion for efficient input/output, (2) adapter prediction using a neural network, and (3) post-processing to trim and segment reads based on predicted adapter positions. To improve runtime, we implemented core functions in Rust, enabled GPU-based inference, and parallelized key components across the pipeline.	284 285 286 287 288
We systematically benchmarked DeepChopper's computational performance to assess scalability for large-scale dRNA-seq studies, testing datasets ranging from 0.1M to 23M reads. Using VCaP read subsamples from 0.1M to 9M reads, we observed near-linear runtime scaling (Extended Data Fig. 6a), with the 9M VCaP dataset requiring approximately 5 hours on two NVIDIA A100 GPUs. Memory usage increased with input size, peaking at approximately 70 GB for CPU-based inference and 56 GB for GPU-based execution (Extended Data Fig. 6b). To evaluate performance at substantially larger scale, we benchmarked a merged dataset of 23 million reads combining five cell lines (A549, HCT116, HepG2, K562, and MCF7), which required approximately 10.6 hours total processing time (23 minutes for FASTQ conversion, 8.5 hours for prediction, 1.7 hours for post-processing). Importantly, runtime continued to scale near-linearly and memory usage remained stable (CPU: 70-93 GB, GPU: 34-56 GB) across the entire range, demonstrating that DeepChopper can process datasets substantially larger than 23M reads with no fundamental computational barriers to scaling (See Methods for details).	289 290 291 292 293 294 295 296 297 298 299 300 301 302 303
Applying DeepChopper to the full VCaP dataset increased usable read yield by 3%, resulting in 9,357,913 adapter-trimmed reads. It identified 8,218,172 adapter sequences across 7,990,102 reads (87% of total), most measuring 70 bp—consistent with the expected length of the RMX adapter used in ONT 's SQK-RNA002 dRNA-seq kit (Fig. 2a) [19]. Analysis of adapter locations revealed that 7,777,624 reads had adapters at the 3' end, while 148,452 contained internal adapters (Fig. 2b), indicating that chimeric artifacts are common in VCaP dRNA-seq data.	304 305 306 307 308 309 310
Further examination showed that chimera artifacts could arise from the joining of multiple long reads, with the most frequent pattern involving two reads joined by a single internal adapter (Fig. 2c). To validate these findings, we analyzed minimap2 [20] chimeric alignments and compared them to a matched VCaP direct cDNA-seq dataset, which we generated as part of this study. Chimeric reads fully supported by cDNA sequencing were considered bona fide events (See Methods for details). Notably, we also evaluated whether ONT 's Dorado basecaller trimming feature could mitigate these artifacts. However, we found that Dorado alone—regardless of trimming—was insufficient to eliminate spurious chimeric alignments. In contrast, DeepChopper reduced unsupported chimeric alignments by around 95% and increased the fraction of cDNA-supported chimeric events from 5.8% to 48.7%, whether applied before or after Dorado	311 312 313 314 315 316 317 318 319 320 321 322

323 trimming (Fig. 2d). These results underscore DeepChopper’s ability to distinguish
324 true biological chimeras from technical artifacts.

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

330 Finally, we evaluated post-processing performance as a function of the sliding
331 window size used for segmentation. Using a 1M-read subsample, we tested window
332 sizes of 11, 21, 31, 41, and 51 bp. Smaller windows yielded slightly higher cDNA
333 support percentages (47.5% for 11 bp vs. 43.3% for 51 bp; Extended Data Fig. 7a),
334 but increased fragmentation of reads into 4+ segments (Extended Data Fig. 7b). A
335 21 bp window provided the optimal balance, maintaining high support while mini-
336 mizing over-segmentation. Based on these results, 21 bp was selected as the default
337 setting, and DeepChopper allows users to adjust this parameter for dataset-specific
338 optimization (See [Methods](#) for details).

339

340 **Multi-sample Validation Across Platforms and Species**

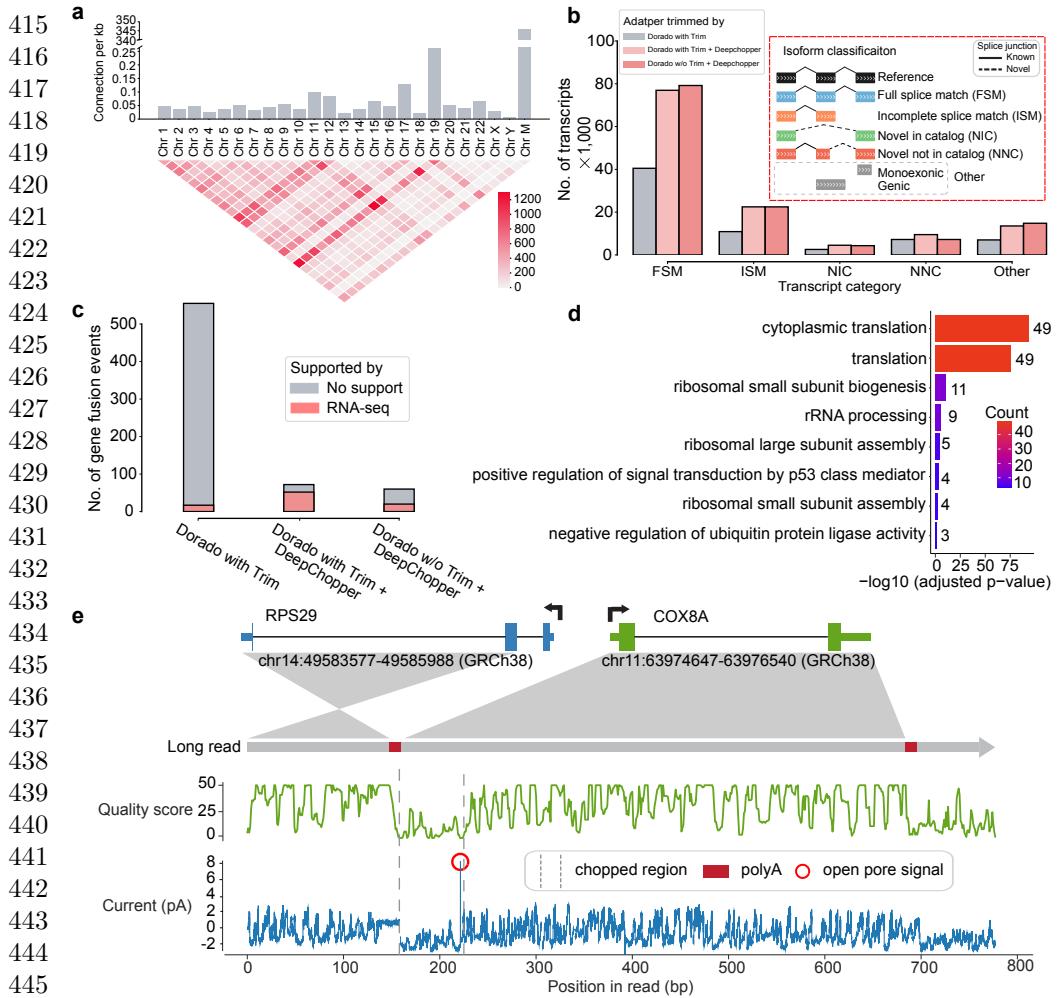
341

342 To further evaluate DeepChopper’s performance beyond the VCaP dataset, we
343 performed multi-sample validation across diverse biological systems and sequenc-
344 ing platforms. We began by analyzing [dRNA-seq](#) data from the [SG-NEx](#) project,
345 comparing chimeric alignments before and after DeepChopper trimming. DeepChop-
346 per detected internal adapters in 0.67–1.25% of reads across these datasets (A549:
347 0.92%, MCF7: 0.67%, HCT116: 1.22%, K562: 0.96%, HepG2: 1.25%), representing
348 15,690–57,122 affected reads per sample (Extended Data Table 4). Critically, inter-
349 nal adapters accounted for 63–85% of all chimeric reads across cell lines, identifying
350 adapter-bridged artifacts as the predominant source of false RNA rearrangement[22].
351 This systematic occurrence of internal adapters across all tested cell lines indicates
352 that adapter-bridged chimeras are not specific to VCaP but represent a general char-
353 acteristic of dRNA-seq data. Across these samples, DeepChopper reduced unsupported
354 chimeric alignments by 62% to 84%, while preserving cDNA-supported chimeric align-
355 ments without noticeable reduction (Fig. 2g). These results reinforce the widespread
356 presence of chimera artifacts in [dRNA-seq](#) and the effectiveness of DeepChopper in
357 selectively removing them without compromising true biological signals.

358 We next applied DeepChopper to the human WTC11 induced pluripotent stem cell
359 line using data from the [Long-read RNA-Seq Genome Annotation Assessment Project](#)
360 ([LRGASP](#)) [23]. This dataset includes cDNA-based long-read sequencing generated
361 with multiple protocols ([PCR](#)-cDNA, CapTrap, R2C2) across [ONT](#) and [PacBio](#) plat-
362 forms, providing a robust benchmarking resource. DeepChopper selectively eliminated
363 only those chimeric alignments not supported by any cDNA-based method (Fig. 2h),
364 further demonstrating its precision in distinguishing genuine chimeras from technical
365 artifacts.

366 To assess cross-species generalizability, we extended the analysis to the F121-9
367 mouse embryonic stem cell line, also from the [LRGASP](#) dataset. DeepChopper reli-
368 ably removed artifactual chimeric reads not supported by any orthogonal cDNA-based

sequencing platform (Extended Data Fig. 8), confirming its applicability to both human and non-human transcriptomes.	369
	370
Importantly, across all datasets, DeepChopper consistently outperformed ONT's Dorado adapter trimming alone—even when applied as a post-processing step—underscoring its distinct and additive utility in chimera artifact correction.	371
	372
	373
	374
Chimera Artifact Analysis in RNA004 Chemistry	375
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.	376
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 4). 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. 9a). 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. 9b).	377
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. 10). 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[22].	378
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 [3, 24]. 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.	379
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.	380
	381
	382
	383
	384
	385
	386
	387
	388
	389
	390
	391
	392
	393
	394
	395
	396
	397
	398
	399
	400
	401
	402
	403
	404
	405
	406
	407
	408
	409
	410
	411
	412
	413
	414



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.

Impact on Downstream Transcriptome Analysis	461
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 < 2.2×10^{-16} ; Extended Data Fig. 11a), while exhibiting a similar gene length distribution (Extended Data Fig. 11b). Analysis of chimeric junctions across the genome revealed uneven distribution among chromosomes, with the mitochondrial chromosome (Chr M) showing the highest frequency of chimeric connections per base pair—suggesting a potential hotspot for artifact formation (Fig. 3a). This pattern persisted in RNA004 dRNA-seq data (Extended Data Fig. 11c), indicating that chimera artifacts remain a fundamental limitation of dRNA-seq, regardless of chemistry improvements.	462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506
We next assessed how DeepChopper correction influences downstream transcriptome analyses. Using IsoQuant [25] to annotate transcripts from VCaP dRNA-seq data, we found that DeepChopper nearly doubled the number of identified transcripts compared to uncorrected data (Fig. 3b). Similar results were observed with RNA004 data (Extended Data Fig. 11d) 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.	490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506
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 [26] 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 [27] (Fig. 3c), suggesting they were false positives. Applying DeepChopper after Dorado trimming yielded consistent results, reinforcing its utility regardless of prior processing steps.	490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506
Closer inspection of the filtered gene fusion calls revealed a strong enrichment for ribosomal protein genes (Extended Data Fig. 12a). GO enrichment analyses in VCaP (Fig. 3d) and SG-NEx cell lines (Extended Data Fig. 12b) 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. 12b). A manual review of a chimeric read identified as an <i>RPS29-COX8A</i> fusion revealed that the DeepChopper-processed region—interpreted as an internal adapter sequence—aligned with low-intensity raw current signals, consistent with ONT adapter characteristics (Fig. 3e). The presence of polyA and open pore signals at the boundary of this region further supported an artifact origin rather than a bona fide fusion event. Extended Data Fig. 13 illustrates the corresponding sequence and base qualities, including the alignment of upstream and downstream segments around the detected adapter, confirming that the apparent <i>RPS29-COX8A</i> fusion arises from an adapter-bridged chimera.	490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506

507 In summary, DeepChopper significantly improves the quality of nanopore **dRNA-**
508 **seq** data by accurately identifying and splitting adapter-bridged chimera that other-
509 wise confound transcript annotation and gene fusion detection. These improvements
510 are robust across different sequencing chemistries and preprocessing pipelines.
511

512 Discussion

513

514 DeepChopper addresses a critical gap in **ONT dRNA-seq**: detecting and recover-
515 ing adapter-bridged chimeric artifacts through internal adapter identification. Our
516 validation across multiple cell lines, species, and chemistries demonstrates internal
517 adapters occur systematically in 0.33-1.62% of reads—tens to hundreds of thousands
518 per experiment—with each propagating errors to transcript annotation, gene fusion
519 detection, and expression quantification [22]. No prior literature has systematically
520 characterized adapter-bridged chimera formation mechanisms in **dRNA-seq**; our work
521 addresses this gap while providing the first computational solution. DeepChopper
522 identifies internal adapter sequences marking artificial junctions, then splits chimeras
523 at these boundaries with single-nucleotide precision to recover component sub-reads.
524 This adapter-based approach differs fundamentally from filtering methods that discard
525 problematic reads. The recovery capability depends on detectable internal adapters,
526 distinguishing adapter-bridged artifacts from **RT**-mediated template-switching in
527 cDNA-seq, which creates chimeras without internal adapters [28, 29].

528 DeepChopper leverages **GLM** capabilities unavailable to conventional approaches:
529 (1) Long-range context [11] enabling full-transcript scanning (median \sim 2.7 kb, 99th
530 percentile \sim 14 kb); (2) Transfer learning enabling robust detection despite sequenc-
531 ing errors; (3) Quality-aware predictions integrating per-base confidence (F1: 0.97 \rightarrow
532 0.99); (4) Alignment-free operation eliminating mapping biases.

533 Performance differences between DeepChopper and existing tools reflect capabil-
534 ity gaps, not quality. Each tool excels at its intended application: Pychopper for
535 cDNA primer detection [7, 30] and Dorado for 3' adapter trimming [5]. However,
536 none detect internal adapters in **dRNA-seq**. Pychopper targets cDNA-specific primers
537 absent in **dRNA-seq**. Porechop/Porechop_ABI require exact matching or stable k-mer
538 patterns [8, 9]—incompatible with unknown, corrupted adapters mis-called by RNA
539 basecallers. Negligible performance on internal detection confirms this is a genuinely
540 unsolved problem.

541 Alignment-based methods employ ruled-based filtering—flagging aberrant align-
542 ment patterns—rather than detecting root causes. Breakinator [18] detects **RT**-
543 mediated foldbacks in cDNA-seq using distance rules; its validation confirms failure in
544 **dRNA-seq** where **RT** artifacts don't occur. FLAIR-fusion [31] is a specialized fusion
545 caller, not general quality control. Distance rules cannot distinguish adapter-bridged
546 artifacts from biological RNA rearrangements [22] and they cannot handle artifacts
547 that do not map as chimeric reads or that fail to map at all. In contrast, internal
548 adapter presence definitively indicates artifacts, enabling selective correction while
549 preserving biological complexity.

550 Validation across RNA002 and RNA004 with Dorado basecalling shows adapter-
551 bridged chimeras persist systematically despite chemistry improvements [3, 24].
552

RNA004 fine-tuning achieved 3-4% additional improvement beyond zero-shot RNA002 model performance, with both models preserving cDNA-supported events. This demonstrates learned probabilistic patterns enable robust generalization without protocol-specific recalibration—critical as technologies evolve. Challenging scenarios (Extended Data Fig. 14, Extended Data Fig. 15) include incomplete 3' end detection in multi-adapter reads and partial detection of degraded sequences. Solutions include combining Dorado (3' end adapters) with DeepChopper (internal adapters) (Extended Data Fig. 16) and post-processing length filtering (default 20 bp minimum). Despite edge cases, DeepChopper achieves 62-91% artifact reduction while preserving cDNA-supported biological events.

DeepChopper demonstrates how [GLM](#) address long-read sequencing challenges resisting conventional approaches. Long-range context, error-tolerant learning, single-nucleotide precision, and quality-aware predictions enable detecting corrupted internal adapters that elude exact-matching, k-mer, or alignment methods. Recovery rather than removal represents a paradigm shift from filtering to correction, preserving valuable data while enhancing accuracy. Future directions include extended context for ultra-long transcripts, alternative approaches for [RT](#)-mediated cDNA-seq chimeras lacking internal adapters, and expansion to additional platforms. This advance enables confident biological interpretation of [dRNA-seq](#) data—from isoform discovery to gene fusion detection—strengthening transcriptomic research in complex biological systems where accurate transcript characterization is essential for understanding gene regulation and cellular function.

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

599 assessed using an Agilent 2100 Bioanalyzer. Poly(A)+ RNA was then enriched from
600 total RNA using the Dynabeadstm mRNA Purification Kit (Invitrogen, Cat# 65001),
601 which utilizes oligo (dT) beads for selective mRNA binding. The mRNA was quantified
602 using a Qubit 4 fluorometer and a Qubit RNA HS Assay Kit (Thermo Fisher Scientific,
603 Cat# Q32852). The mRNA preparations were either immediately used to prepare a
604 sequencing library or frozen and stored at –80 °C until further use.

605

606 Nanopore sequencing

607

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

624

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

640

641 Training data preparation

642

We acquired [ONT dRNA-seq](#) FAST5 data from the [SG-NEx](#) project, which includes six human cell lines: HEK293T, A549, K562, HepG2, MCF7, and HCT116 [17].

644

The FAST5 files were converted to POD5 format using the POD5 conversion tool (https://pod5-file-format.readthedocs.io). Subsequently, FASTQ files were generated using Dorado (v0.5.2) [5] with adapter trimming disabled (`--no-trim`) and the “rna002.70bps_hac@v3” model. The reads were then aligned to the human reference genome (GRCh38) using minimap2 (v2.24) [20] with ONT direct RNA-specific parameters (`-ax splice -uf -k14`) for optimized alignment. The resulting SAM files were then converted to BAM format, indexed, and sorted using SAMtools (v1.19.2) [32].

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

To create artificial chimeric reads, we randomly combined two non-adapter sequences with one adapter sequence to create FASTQ records. The dataset consists of positive examples containing adapter sequences (with a 1:1 ratio of 3' end and internal adapters) and negative examples without any adapter sequences, in a 9:1 ratio. In total, 600,000 data points were generated and divided into training ($N = 480,000$), validation ($N = 60,000$), and test sets ($N = 60,000$) in an 8:1:1 ratio using stratified random sampling.

Language model architecture

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

At its core, DeepChopper employs HyenaDNA [11] 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

691 this normalized quality information while preserving the original sequence features.
692 Each [MLP](#) layer is followed by ReLU activation, enhancing the model's ability to learn
693 complex quality-sequence relationships.

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

701

$$702 P(y_i = c) = \frac{e^{z_c}}{\sum_{j=1}^2 e^{z_j}}$$

703

704 where $P(y_i = c)$ represents the probability that nucleotide position i belongs to
705 class c . The final classification decision is based on the class with the higher probability
706 score. In other words, a nucleotide is classified as an adapter if $P(y_i = \text{adapter}) >$
707 $P(y_i = \text{non-adapter})$. Hence, a threshold of 0.5 is applied implicitly.

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

711

712 Model training

713

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

724 The Adam optimizer was used for parameter optimization, with settings of $\beta_1 = 0.9$
725 and $\beta_2 = 0.999$ [33]. A learning rate scheduler was used to reduce the learning rate
726 when validation loss ceased improving, starting with an initial learning rate of 2×10^{-5} .
727 The cross-entropy loss function was used to update the model parameters, defined as
728 follows:

729

$$730 \mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

731

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

735

736

The average cross-entropy loss across the mini-batch is computed as: 737
738

$$\mathcal{L}_{\text{BatchBCE}} = \frac{1}{B} \sum_{j=1}^B \mathcal{L}_{\text{BCE}}(\mathbf{y}_j, \hat{\mathbf{y}}_j) \quad 739
740
741$$

where $\mathcal{L}_{\text{BatchBCE}}$ is the average binary cross-entropy loss for the mini-batch, B is the batch size (number of sequences in the mini-batch), and \mathbf{y}_j and $\hat{\mathbf{y}}_j$ are the true labels and predicted probabilities for the j -th sequence in the batch. 742
743
744
745
746
747

The model evaluation metrics included accuracy, precision, recall and the F1 score, calculated using the following equations: 748
749
750
751
752
753
754
755
756
757
758

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

The final selection of the model was based on the optimal performance in the validation set. The model is implemented by PyTorch (v2.5.0) [34]. To identify the best hyperparameter configuration, the Hydra (v1.3.2) [35] framework was used. 759
760

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. 761
762
763
764
765
766
767
768
769

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. 770
771
772
773
774
775
776
777
778
779

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 780
781
782

783 two NVIDIA A100 GPUs using the optimization settings described in [Model training](#),
784 with early stopping based on validation F1 score. Fine-tuning required approximately
785 9 hours of total training time. Performance evaluation compared three conditions:
786 (1) Dorado basecalling with adapter trimming (baseline), (2) Dorado followed by the
787 original RNA002-trained DeepChopper, and (3) Dorado followed by the RNA004-
788 fine-tuned DeepChopper. Chimeric alignments were evaluated against matched cDNA
789 sequencing data to distinguish biological events from artifacts.

790

791 Sliding window approach for prediction refinement

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

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

$$804 \\ 805 \\ 806 y_i = \begin{cases} 1 & \text{if } \sum_{j=i-k}^{i+k} p_j > \frac{W}{2} \\ 0 & \text{otherwise} \end{cases} \\ 807$$

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

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

816

817 Post-processing and filtering

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

- 821 1. A predicted adapter sequence must be at least 13 nucleotides long. Sequences
822 shorter than this length threshold are not considered valid adapters.
- 823 2. If a read contains more than four adapter sequences, the entire read sequence is
824 retained without any adapter removal.
- 825 3. For reads containing four or fewer adapter sequences, the identified adapters are
826 removed and the read is divided into smaller segments.
- 827 4. Any segments resulting from this process that are less than 20 nucleotides long
828 are discarded.

Each remaining segment and its corresponding base quality scores are stored as a single read record in the final FASTQ file. This filtering process separates chimeric read artifacts containing internal adapters into multiple segments, while retaining reads with 3' end adapters as single shortened segments. 829
830
831
832

All filtering thresholds, including minimum segment length, and maximum adapter count per read, are configurable via command-line parameters in the DeepChopper implementation, allowing users to tailor these settings to dataset-specific requirements or experimental conditions. 833
834
835
836
837

BLAT identity calculation 838

The accuracy of DeepChopper in detecting adapter sequences was evaluated by aligning the identified sequences to the human reference genome using **BLAT** [21]. A **BLAT** identity score was defined as the ratio of matched bases to the total sequence length: 839
840
841
842

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

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) [36] 843
844
845

Computational benchmarks 846

All benchmarks were conducted in triplicate using btop (https://github.com/aristocratos/btop, v1.4.0) and nvtop (https://github.com/Syllo/nvtop, v3.1.0) to monitor CPU and GPU memory usage, respectively. Evaluations were performed on high-performance computing infrastructure with 16 CPU cores, 60 GB RAM, and dual NVIDIA A100 GPUs (80 GB memory each). Adapter prediction stage used a batch size of 64. 847
848
849
850
851

Transcript length distribution analysis 852

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 RNA-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 percentage of reads exceeding the 32 kb threshold were specifically quantified to assess 853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874

875 the practical impact of this limitation. All read length statistics were computed from
876 Dorado (v0.5.2) base-called reads with the trim option enabled.

877

878 Validation of chimera artifact reduction

879 Cross-platform validation of **dRNA-seq** chimera artifacts identified by DeepChopper
880 was conducted leveraging **ONT** direct cDNA sequencing and additional cDNA-based
881 sequencing platforms. Direct cDNA sequencing validation was performed using six
882 cancer cell lines, including the VCaP dataset generated in this study and five pub-
883 lished datasets (A549, K562, HepG2, MCF7, and HCT116) obtained from the **SG-NEx**
884 project [17]. The direct cDNA data in FAST5 format were converted to POD5 format
885 using the POD5 conversion tool (<https://pod5-file-format.readthedocs.io>). Subse-
886 quently, FASTQ files were generated using Dorado (v0.5.2) [5] with adapter trimming
887 enabled (--trim adapters) and the “dna_r9.4.1_e8_hac@v3.3” model. The reads were
888 then processed using Pychopper (<https://github.com/epi2me-labs/pychopper>, v2.7.9)
889 and Cutadapt (v4.2) [37] according to a published protocol [38]. The oriented reads
890 were aligned to the human reference genome (GRCh38) using minimap2 (v2.24) [20]
891 with optimized parameters (-ax splice -uf -k14) for spliced alignment. The resulting
892 SAM files were then converted to BAM format, indexed, and sorted using SAMtools
893 (v1.19.2) [32].

894 Additional cDNA-based long-read sequencing data from the WTC11 (human) and
895 F121-9 (mouse) cell lines were used for further validation, incorporating five distinct
896 platforms: **ONT PCR-cDNA**, **ONT CapTrap**, **ONT R2C2**, **PacBio** cDNA, and **PacBio**
897 CapTrap. The raw FASTQ files (and FASTA files for **ONT R2C2**) from these datasets
898 were provided by the **LRGASP** project [23]. For the **PCR-cDNA** data, the reads were
899 processed using Pychopper (<https://github.com/epi2me-labs/pychopper>, v2.7.9) and
900 Cutadapt (v4.2) [37], following the protocol described in reference [38]. **ONT** reads
901 were then aligned to the human reference genome (GRCh38) or mouse reference
902 genome (GRCm39) using minimap2 (v2.24) [20] with the parameters (-ax splice -uf
903 -k14), while **PacBio** reads were aligned using the parameters (-ax splice:hq -uf). The
904 **ONT dRNA-seq** data from A549, K562, HepG2, MCF7, HCT116, VCaP, WTC11 and
905 F121-9 cell lines were processed as previously described, except that The F121-9 cell
906 line data was aligned to the mouse reference genome (GRCm39).

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

917 To empirically assess the applicability of alignment-based artifact detection meth-
918 ods to **dRNA-seq**, we compared DeepChopper with Breakinator (v1.0) [18] using VCaP
919 RNA002 data. Three processing pipelines were evaluated: Dorado basecalling with

trim option (baseline), Dorado with trim followed by DeepChopper, and Dorado with trim followed by Breakinator. All pipelines started with identical Dorado-trimmed reads to enable direct comparison. Breakinator was applied using default parameters as specified in its documentation [18]. Chimeric alignments from each pipeline were classified as cDNA-supported (validated by matched direct cDNA sequencing) or unsupported (likely artifacts) using the validation framework described above. Support rates were calculated as the proportion of chimeric alignments corroborated by cDNA-seq data.	921 922 923 924 925 926 927 928 929
Gene expression analysis and transcript classification	930
Gene expression levels from dRNA-seq were quantified using IsoQuant (v3.1.2) [25], 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 [39].	931 932 933 934 935 936
Gene fusion identification and visualization	937 938
For ONT dRNA-seq data, gene fusions were identified using FusionSeeker (v1.0.1) [26] 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 [40] under SRA accession SRX5417211. Raw reads were mapped to the hg38 reference genome using STAR (v2.7.11) [41], and gene fusion events were detected with Arriba (v2.4.0) [27]. The gene structure of the RPS29-COX8A fusion was visualized using GSDS (v2.0) [42]. Base quality scores were generated with a custom Python script, and ion current signals were visualized using Squigualiser (v0.6.3) [43]. The circos plot for gene fusion events was visualized using chimeraviz (v1.30.0) [44].	939 940 941 942 943 944 945 946 947 948
GO enrichment analysis	949 950
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 [45].	951 952 953 954 955
Computing resource	956 957 958 959 960
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.	961 962 963 964 965 966

967 **Code Availability.** DeepChopper, implemented in Rust and Python, is open
968 source and available on GitHub (<https://github.com/ylab-hi/DeepChopper>) under the
969 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,
970 Linux, and macOS to ensure easy cross-platform installation. An interactive demo
971 is available on Hugging Face (<https://huggingface.co/spaces/yangliz5/deepchopper>),
972 allowing users to test DeepChopper's functionality without local installation. For
973 large-scale analyses, we recommend using DeepChopper on systems with GPU
974 acceleration. Detailed system requirements and optimization guidelines are available in the
975 repository's documentation.
976

977 **Acknowledgements.** This project was supported in part by NIH grants
978 R35GM142441 and R01CA259388 awarded to RY, and NIH grants R01CA256741,
979 R01CA278832, and R01CA285684 awarded to QC.
980

981 **Author Contributions.** YL, TYW and RY designed the study with QC. YL
982 and TYW performed the analysis. QG, YR and XL performed the experiments. YL
983 designed and implemented the model and computational tool. YL, TYW, QG and RY
984 wrote the manuscript. RY supervised this work.
985

986 **Conflict of interests.** RY has served as an advisor/consultant for Tempus AI, Inc.
987 This relationship is unrelated to and did not influence the research presented in this
988 study.
989

990 **Acronyms**

991 **ATCC** American Type Culture Collection [13](#)
992

993 **BLAT** BLAST-like alignment tool [5](#), [8](#), [9](#), [19](#), [32](#)
994

995 **CCLE** Cancer Cell Line Encyclopedia [21](#)
996

997 **DAVID** Database for Annotation, Visualization, and Integrated Discovery [21](#)
998 **dRNA-seq** direct RNA sequencing [1](#), [2](#), [4–15](#), [18](#), [20](#), [21](#), [27](#), [31–35](#)
999

1000 **FSM** Full splice match [11](#), [34](#)
1001

1002 **GEO** Gene Expression Omnibus [21](#)
1003

1003 **GLM** Genomic Language Model [2](#), [4](#), [12](#), [13](#)
1004

1004 **GO** Gene Ontology [10](#), [11](#), [21](#), [35](#)
1005

1005 **GPU** Graphics Processing Unit [16](#), [21](#), [22](#)
1006

1006 **HPC** High Performance Computing [16](#), [21](#)
1007

1008 **ISM** Incomplete splice match [11](#), [34](#)
1009

1010 **LCGLM** long-context genomic language model [3](#)
1011

1011 **LLM** Large Language Model [2](#)
1012

1012 **LRGASP** Long-read RNA-Seq Genome Annotation Project [8](#), [20](#)

MLP multilayer perceptron	3, 15, 16	1013
mRNA messenger RNA	14	1014
		1015
NIC Novel in catalog	11, 34	1016
NNC Novel not in catalog	11, 34	1017
		1018
ONT Oxford Nanopore Technologies	2, 4, 5, 7–12, 14, 18, 20, 21, 32	1019
		1020
PacBio Pacific Biosciences	5, 8, 20, 32	1021
PCR Polymerase Chain Reaction	2, 8, 20	1022
		1023
RT Reverse Transcription	2, 12, 13	1024
RTA reverse transcriptase adapter	14	1025
		1026
SG-NEx Singapore Nanopore Expression Project	4, 7, 8, 11, 14, 20	1027
		1028
		1029
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).	1030
		1031
		1032
[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).	1033
		1034
		1035
[3]	Zou, Y. <i>et al.</i> A comparative evaluation of computational models for RNA modification detection using nanopore sequencing with RNA004 chemistry 26 , bbaf404.	1036
		1037
		1038
		1039
[4]	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).	1040
		1041
		1042
[5]	PLC., O. N. Dorado. https://github.com/nanoporetech/dorado (2023).	1043
		1044
		1045
		1046
[6]	Liu-Wei, W. <i>et al.</i> Sequencing accuracy and systematic errors of nanopore direct rna sequencing. <i>BMC genomics</i> 25 , 528 (2024).	1047
		1048
		1049
[7]	epi2me-labs/pychopper: cdna read preprocessing. Github. URL https://github.com/epi2me-labs/pychopper .	1050
		1051
		1052
		1053
[8]	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 .	1054
		1055
		1056
		1057
		1058

- 1059 [10] Benegas, G., Ye, C., Albors, C., Li, J. C. & Song, Y. S. Genomic language
1060 models: Opportunities and challenges (2024). URL <https://arxiv.org/abs/2407.11435>.
1061
- 1062
- 1063 [11] Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at sin-
1064 gle nucleotide resolution. *Advances in neural information processing systems* **36**
1065 (2024).
- 1066
- 1067 [12] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image
1068 Recognition. [1512.03385](https://arxiv.org/abs/1512.03385).
- 1069
- 1070 [13] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained bidirectional
1071 encoder representations from transformers model for DNA-language in genome
1072 **37**, 2112–2120.
- 1073
- 1074 [14] Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for
1075 multi-species genome. [2306.15006](https://arxiv.org/abs/2306.15006).
- 1076
- 1077 [15] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust
1078 foundation models for human genomics. *Nature Methods* 1–11 (2024).
- 1079
- 1080 [16] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale
1081 with evo. *Science* **386**, eado9336 (2024). URL <https://www.science.org/doi/abs/10.1126/science.ado9336>.
- 1082
- 1083 [17] Chen, Y. *et al.* A systematic benchmark of nanopore long read rna sequencing
1084 for transcript level analysis in human cell lines. *BioRxiv* 2021–04 (2021).
- 1085
- 1086 [18] Heinz, J. M., Meyerson, M. & Li, H. Detecting foldback artifacts in long-reads.
- 1087
- 1088 [19] PLC., O. N. Chemistry Technical Document (CHTD_500_v1_revAQ_07Jul2016)
1089 (2017). URL <https://nanoporetech.com/document/chemistry-technical-document>.
- 1090
- 1091 [20] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
1092 **34**, 3094–3100 (2018).
- 1093
- 1094 [21] Kent, W. J. Blat—the blast-like alignment tool. *Genome research* **12**, 656–664
1095 (2002).
- 1096
- 1097 [22] Ma, C., Shao, M. & Kingsford, C. SQUID: Transcriptomic structural variation
1098 detection from RNA-seq **19**, 52.
- 1099
- 1100 [23] Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read rna-seq methods
1101 for transcript identification and quantification. *Nature methods* 1–15 (2024).
- 1102
- 1103 [24] Hewel, C. *et al.* Direct rna sequencing enables improved transcriptome assess-
1104 ment and tracking of rna modifications for medical applications. *bioRxiv* 2024–07

- (2024). 1105
- [25] Prjibelski, A. D. *et al.* Accurate isoform discovery with isoquant using long reads. 1106
Nature Biotechnology **41**, 915–918 (2023). 1107
- [26] Chen, Y. *et al.* Gene fusion detection and characterization in long-read cancer 1108
transcriptome sequencing data with fusionseeker. *Cancer research* **83**, 28–33 1109
(2023). 1110
- [27] Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from rna 1111
sequencing data. *Genome research* **31**, 448–460 (2021). 1112
- [28] Schulz, L. *et al.* Direct long-read RNA sequencing identifies a subset of 1113
questionable exitrons likely arising from reverse transcription artifacts **22**, 190. 1114
- [29] Balázs, Z. *et al.* Template-switching artifacts resemble alternative polyadenylation 1115
20, 824. 1116
- [30] Sessegolo, C. *et al.* Transcriptome profiling of mouse samples using nanopore 1117
sequencing of cDNA and RNA molecules **9**, 14908. 1118
- [31] Felton, C., Tang, A. D., Knisbacher, B. A., Wu, C. J. & Brooks, A. N. Detection of 1119
alternative isoforms of gene fusions from long-read RNA-seq with FLAIR-fusion. 1120
- [32] Li, H. *et al.* The sequence alignment/map format and samtools. *bioinformatics* 1121
25, 2078–2079 (2009). 1122
- [33] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 1123
- [34] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning 1124
library. *Advances in neural information processing systems* **32** (2019). 1125
- [35] Yadan, O. Hydra - a framework for elegantly configuring complex applications. 1126
Github (2019). URL <https://github.com/facebookresearch/hydra>. 1127
- [36] Li, Y. & Yang, R. PxBLAT: an efficient python binding library for BLAT. *BMC Bioinf.* **25**, 1–8 (2024). 1128
- [37] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing 1129
reads. *EMBnet. journal* **17**, 10–12 (2011). 1130
- [38] Grünberger, F., Ferreira-Cerca, S. & Grohmann, D. Nanopore sequencing of rna 1131
and cdna molecules in escherichia coli. *Rna* **28**, 400–417 (2022). 1132
- [39] Tardaguila, M. *et al.* Sqanti: extensive characterization of long-read transcript 1133
sequences for quality control in full-length transcriptome identification and 1134
quantification. *Genome research* **28**, 396–411 (2018). 1135
- 1136
- 1137
- 1138
- 1139
- 1140
- 1141
- 1142
- 1143
- 1144
- 1145
- 1146
- 1147
- 1148
- 1149
- 1150

- 1151 [40] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling
1152 of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- 1153
- 1154 [41] Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**,
1155 15–21 (2013).
- 1156
- 1157 [42] Hu, B. *et al.* Gsds 2.0: an upgraded gene feature visualization server. *Bioinfor-*
1158 *matics* **31**, 1296–1297 (2015).
- 1159
- 1160 [43] Samarakoon, H. *et al.* Interactive visualisation of raw nanopore signal data with
1161 squigualiser. *Biorxiv* 2024–02 (2024).
- 1162
- 1163 [44] Lågstad, S. *et al.* chimeraviz: a tool for visualizing chimeric rna. *Bioinformatics*
1164 **33**, 2954–2956 (2017).
- 1165
- 1166 [45] Sherman, B. T. *et al.* David: a web server for functional enrichment analysis
1167 and functional annotation of gene lists (2021 update). *Nucleic acids research* **50**,
1168 W216–W221 (2022).
- 1169
- 1170
- 1171
- 1172
- 1173
- 1174
- 1175
- 1176
- 1177
- 1178
- 1179
- 1180
- 1181
- 1182
- 1183
- 1184
- 1185
- 1186
- 1187
- 1188
- 1189
- 1190
- 1191
- 1192
- 1193
- 1194
- 1195
- 1196

Extended data

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

Adapter trimming tool	dRNA-seq terminal adapter trimming	dRNA-seq internal adapter trimming	Trimming existing dRNA-seq datasets (post-basecalling)
Porechop [8]	✗	✗	✗
Porechop-ABI [9]	✗	✗	✗
Pychopper [7]	✗	✗	✗
Dorado [5]	✓	✗	✗
DeepChopper	✓	✓	✓

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

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

1243

1244

1245

1246 **Extended Data Table 4** Internal Adapter Prevalence Across Datasets

1247

1248 1249 1250 1251	Sample	All Reads			Chimeric Reads Only		
		With Internal Adapters (A)	Total (B)	% (A/B)	With Internal Adapters (C)	Total (D)	% (C/D)
1252	A549	15,690	1,703,697	0.92	10,553	12,803	82.43
1253	MCF7	20,340	3,039,468	0.67	11,115	17,646	63.00
1254	HCT116	57,122	4,697,299	1.22	37,823	46,800	80.81
1255	K562	29,436	3,061,722	0.96	19,289	23,214	83.09
1256	HepG2	22,530	1,797,922	1.25	14,331	16,921	84.69
1257	VCaP RNA002	148,452	9,177,422	1.62	98,878	107,265	92.18
1258	VCaP RNA004	38,878	11,714,520	0.33	6,891	29,144	23.65

1258 Total reads from Dorado with trim. Internal adapters detected by DeepChopper after Dorado processing.
 1259 VCaP RNA002 and RNA004 demonstrate that adapter-bridged chimeras persist across chemistries.

1260

1261

1262

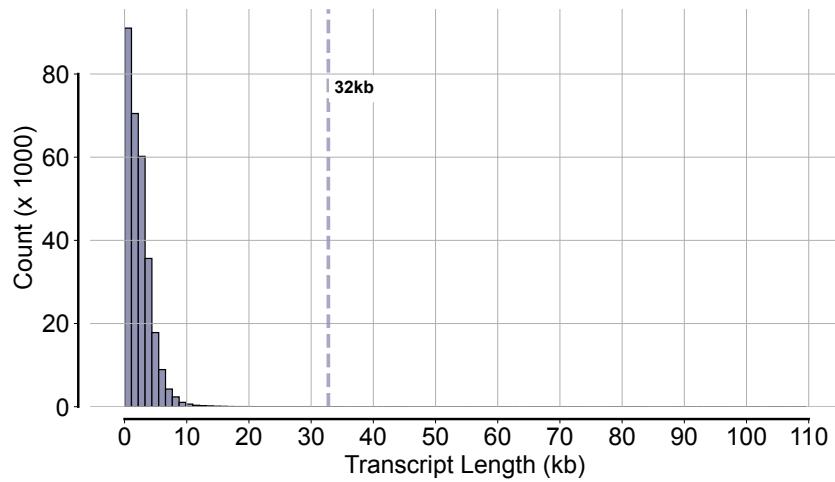
1263

1264

1265

1266

1267



1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

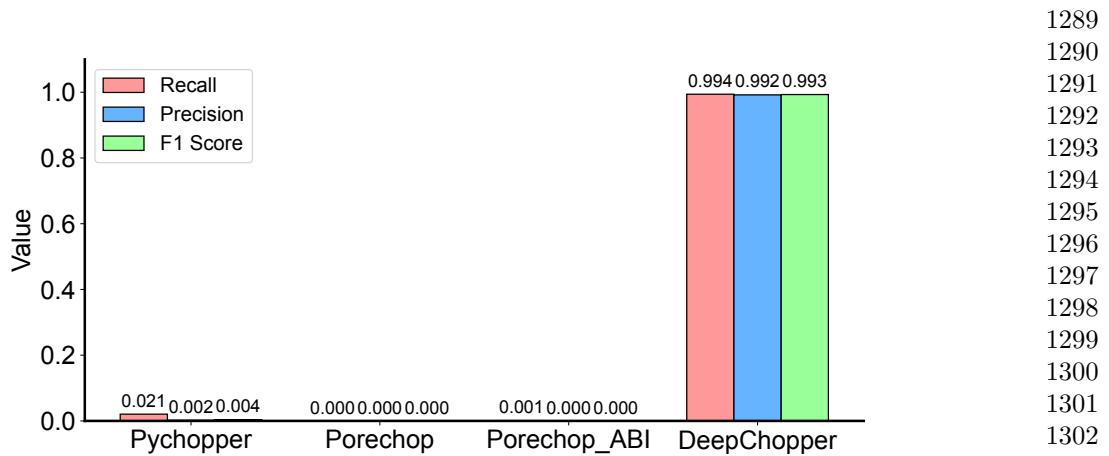
1283

1283 **Extended Data Fig. 1 Distribution of transcript length for protein-coding genes.** Analysis
 1284 of all protein-coding transcripts from Ensembl GRCh38.115 (released July 2025) shows that >99.99%
 1285 of transcripts are below the 32 kb threshold (marked with vertical dashed line). The distribution is
 1286 highly skewed toward shorter transcripts, with median length of ~2.7 kb.

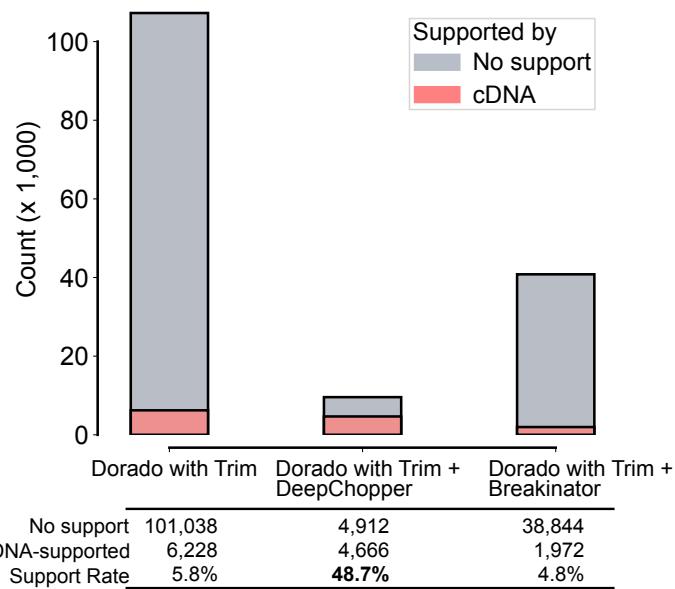
1286

1287

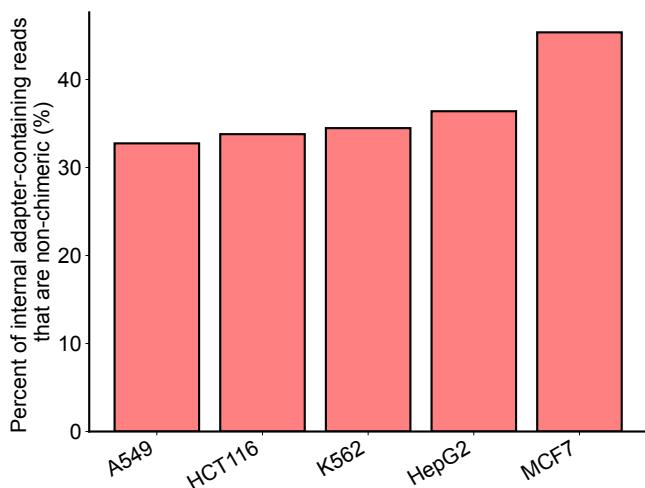
1288



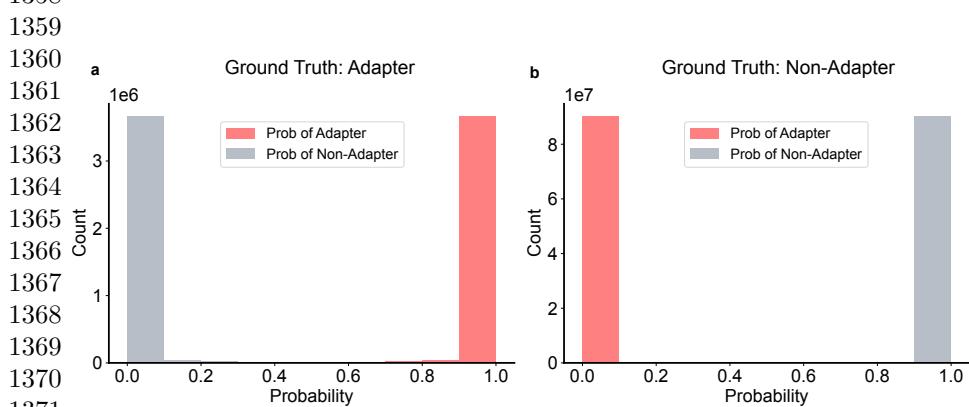
Extended Data Fig. 2 Performance evaluation in a held-out test dataset ($N = 60,000$) showing Recall, Precision, and F1 values for DeepChopper, Pychopper, Porechop, and Porechop_ABI.



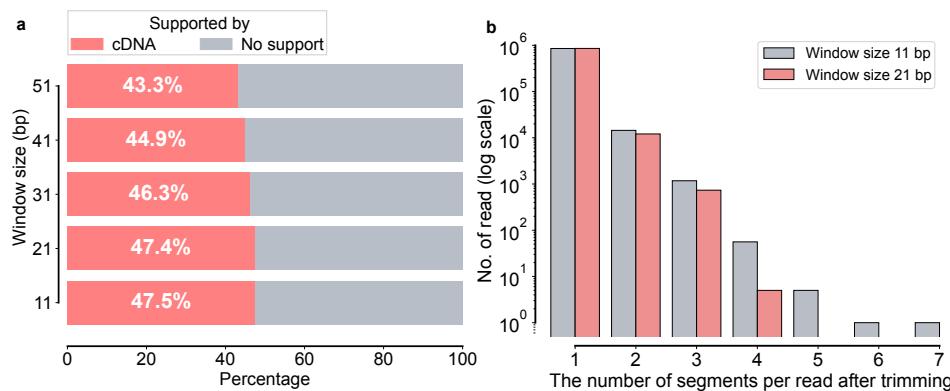
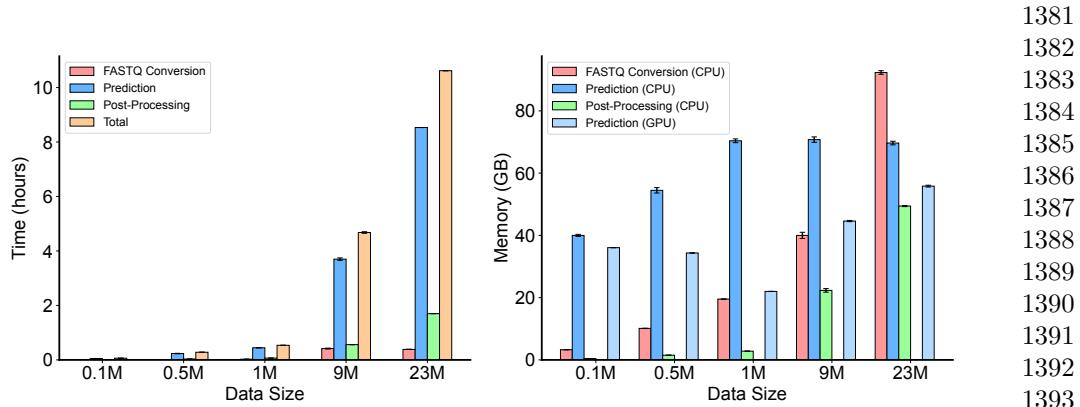
Extended Data Fig. 3 Comparison of chimeric alignment reduction strategies in VCaP RNA002 dRNA-seq data. Stacked bar plot showing chimeric alignments (in thousands) for three processing pipelines: Dorado with adapter trimming (baseline), Dorado with adapter trimming followed by DeepChopper, and Dorado with adapter trimming followed by Breakinator. Gray bars represent unsupported chimeric alignments (likely artifacts); pink bars represent cDNA-supported chimeric alignments (biological events).



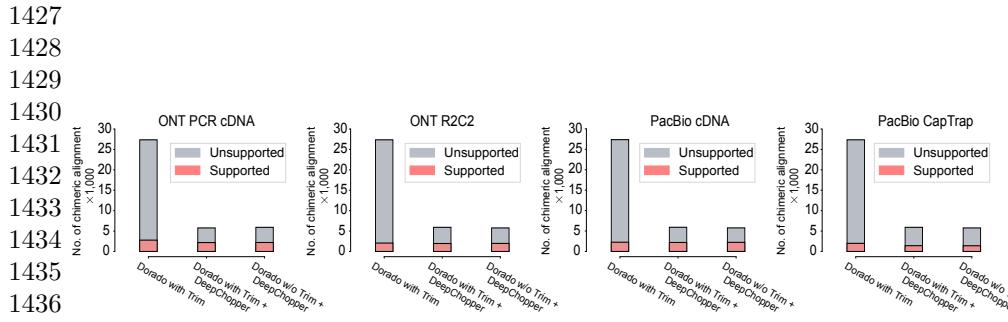
1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349
 1350
 1351
 1352
 1353 **Extended Data Fig. 4 Percent of internal adapter-containing reads that are non-**
 1354 chimeric Percentage of internal adapter-containing reads that do not produce chimeric alignments
 1355 across five human cell lines (RNA002) processed by Dorado with trim followed by DeepChopper.
 1356 Between 33–45% of adapter-containing reads map as single alignments or fail to map, making them
 1357 invisible to chimeric alignment-based artifact detection.



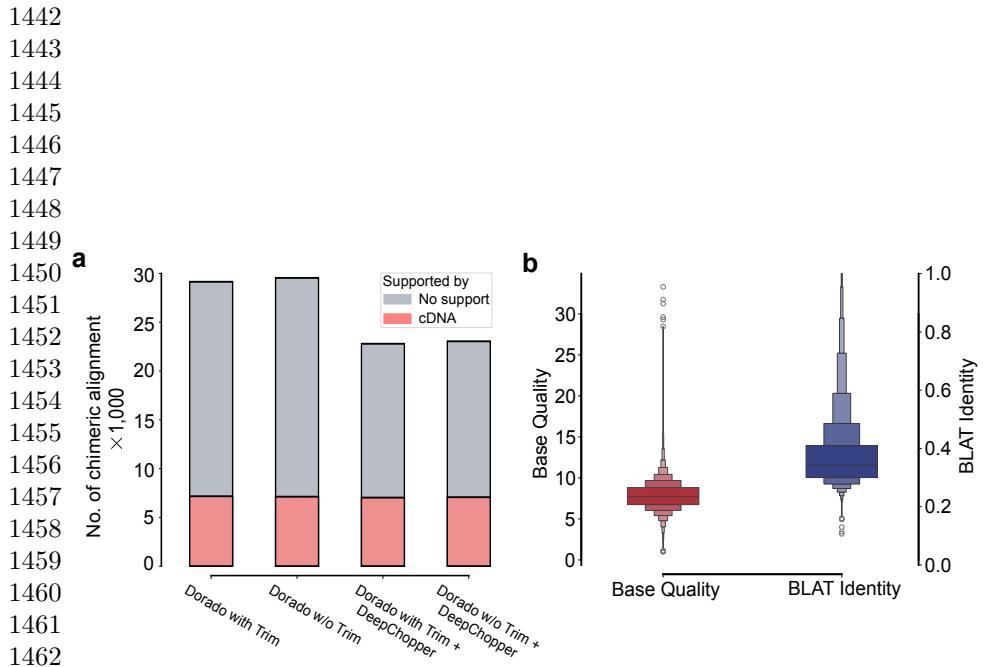
1358
 1359
 1360 **a** Ground Truth: Adapter
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371 **b** Ground Truth: Non-Adapter
 1372 **Extended Data Fig. 5 Prediction probability distributions of DeepChopper for the held-**
 1373 **out test dataset ($N = 60,000$)**. (a) Distribution of prediction probabilities for sequences with
 1374 ground truth adapter classification. Red bars represent the probability of adapter prediction, while
 1375 gray bars show the probability of non-adapter prediction. The count (y-axis) is shown in millions of
 1376 sequences (10^6 scale). (b) Distribution of prediction probabilities for sequences with ground truth
 1377 non-adapter classification. Red bars indicate the probability of adapter prediction, while gray bars
 1378 show the probability of non-adapter prediction. The count (y-axis) is shown in tens of millions of
 1379 sequences (10^7 scale). Both distributions demonstrate strong polarization toward correct classification
 1380 probabilities, indicating the model's high confidence in distinguishing between adapter and non-
 1381 adapter sequences.



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

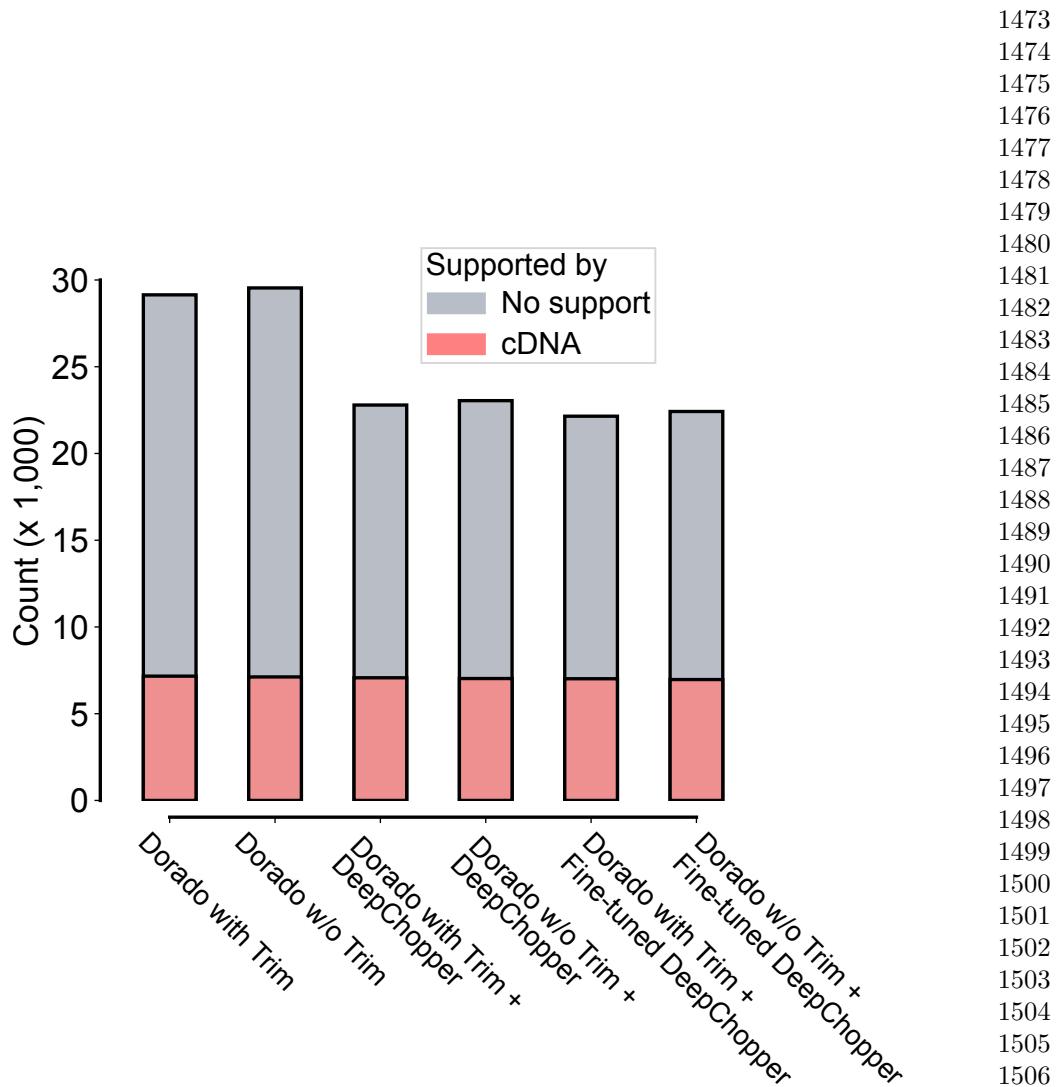


1437 **Extended Data Fig. 8** Chimeric alignments from [dRNA-seq](#) of the F121-9 cell line
1438 (mouse), evaluated for support using additional [ONT](#) and [PacBio](#) sequencing data with
1439 different protocols. DeepChopper-involved methods reduce unsupported chimeric align-
1440 ments across all methods compared to Dorado with adapter trimming. The bar colors
1441 indicate chimeric alignments supported by additional sequencing data (red) and those
1442 lacking support (grey).



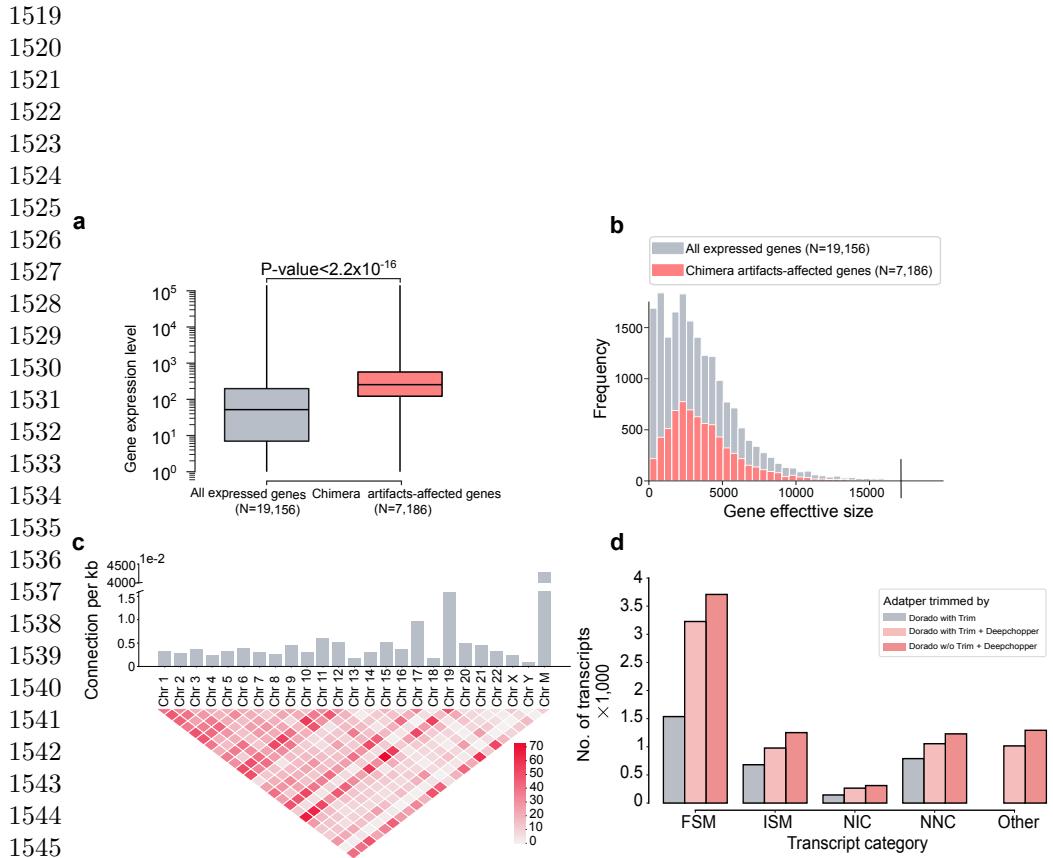
1463 **Extended Data Fig. 9** Evaluation of DeepChopper's predictions on chimeric read arti-
1464 facts in [dRNA-seq](#) data generated using the SQK-RNA004 kit from the VCaP cell line.
1465 (a) Number of chimeric alignments (in thousands) identified in VCaP RNA004 [dRNA-seq](#) reads pro-
1466 cessed by Dorado with and without adapter trimming, Dorado with adapter trimming followed by
1467 DeepChopper, and DeepChopper. The bar colors indicate chimeric alignments supported by cDNA
1468 sequencing (red) and those lacking support (grey). (b) Base quality scores (left) and [BLAT](#) align-
1469 ment identity scores (right) for internal adapter sequences identified by DeepChopper in RNA004
1470 [dRNA-seq](#) reads.

1471
1472



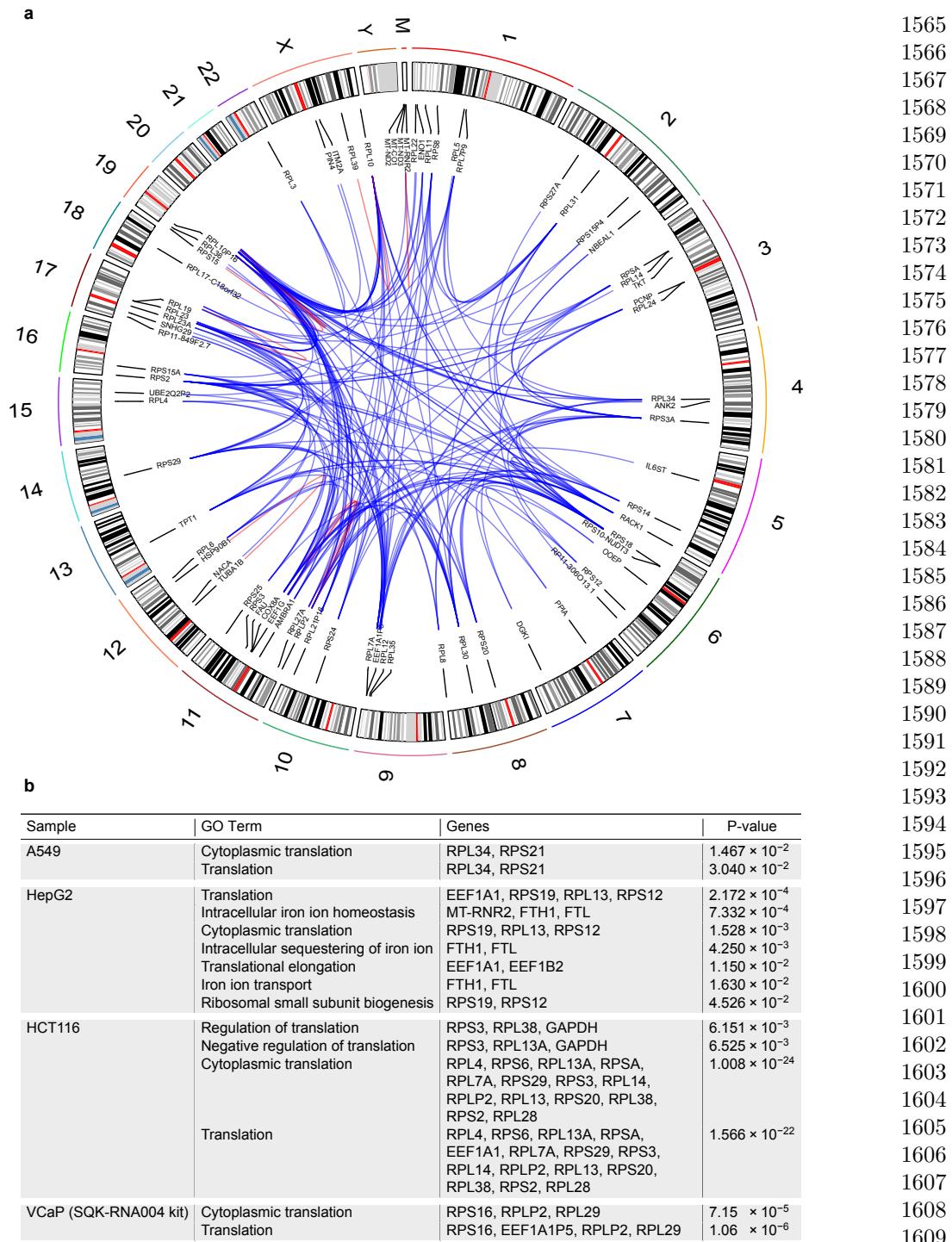
Extended Data Fig. 10 Performance comparison of original and fine-tuned DeepChopper on RNA004 data. Number of chimeric alignments (in thousands) identified in VCaP RNA004 dRNA-seq processed under six conditions: Dorado basecalling with and without adapter trimming, followed by original DeepChopper, and followed by fine-tuned DeepChopper. The bar colors indicate chimeric alignments supported by cDNA sequencing (red) and those lacking support (grey).

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



1546 **Extended Data Fig. 11 Analysis of dRNA-seq chimera artifacts and their genomic and**
1547 **transcriptomic characteristics in VCaP cells.** (a) Box plot comparing gene expression levels
1548 between all expressed genes ($N=19,156$) and genes affected by chimera artifacts ($N=7,186$) in the
1549 VCaP dRNA-seq dataset. Chimera artifacts-affected genes exhibit significantly higher expression lev-
1550 els ($p\text{-value} < 2.2 \times 10^{-16}$). (b) Distribution of gene effective sizes for all expressed genes and genes
1551 affected by chimera artifacts, indicating that the size distributions of genes impacted by chimera
1552 artifacts are comparable to those of all expressed genes. (c) Chromosomal distribution and interchromo-
1553 mosomal connections from chimeric read artifacts arising from VCaP RNA004 dRNA-seq. The top
1554 bar plot shows the number of connections per kilobase for each chromosome, with higher bars indicat-
1555 ing more frequent connections. The bottom heatmap visualizes the number of chimeric connections
1556 between chromosome pairs, with color intensity representing the connection frequency. (d) Number
1557 of detected transcripts across different isoform categories (FSM, ISM, NIC, NNC, and Other) from
1558 DeepChopper-identified chimeric read artifacts in VCaP RNA004 dRNA-seq data. DeepChopper-
1559 corrected reads resulted in a greater number of transcripts compared to adapter-trimmed reads by
1560 Dorado across all categories.

1561
1562
1563
1564



Extended Data Fig. 12 Analysis of gene fusions derived from chimeric read artifacts in dRNA-seq. (a) Circos plot depicting chromosomal connections of gene fusions resulting from chimeric read artifacts in VCaP cells. Blue lines represent inter-chromosomal fusion events, while red lines indicate intra-chromosomal fusions.³⁵ 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.

