

ChimeraLM filters amplification artifacts for
accurate structural variant calling in long-read
single-cell sequencing

Yangyang Li^{1†}, Qingxiang Guo^{1†}, Rendong Yang^{1,2*}

¹Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, 60611, IL, USA.

²Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, 675 N St Clair St, Chicago, 60611, IL, USA.

*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu;

Contributing authors: yangyang.li@northwestern.edu;
qingxiang.guo@northwestern.edu;

†These authors contributed equally to this work.

Abstract

Single-cell genomics provides unprecedented insights into cellular heterogeneity, but Whole Genome Amplification (WGA)—required to obtain sufficient DNA—introduces chimeric artifacts that generate false-positive Structural Variations (SVs) and undermine biological interpretations. Current computational methods cannot reliably distinguish amplification-induced artifacts from genuine rearrangements. Here we present ChimeraLM, a genomic language model that learns sequence-level features to discriminate biological sequences from WGA artifacts. Validated on nanopore long-read data, ChimeraLM achieves 95% recall with 70% precision, reduces chimeric reads by ~90%, and preserves 72–92% of true SVs. This improves SV validation rates 8–11 fold and eliminates artificial inversion (INV) bias, restoring SV type distributions to bulk-like profiles. Attention visualization reveals that ChimeraLM focuses on chimeric junction regions, learning mechanistically interpretable features that generalize across sequencing platforms. By enabling reliable SV detection at single-cell resolution, ChimeraLM addresses a critical data quality barrier in cancer genomics, developmental biology, and somatic mosaicism studies. ChimeraLM is available at <https://github.com/ylab-hi/ChimeraLM>.

047 **Keywords:** Whole Genome Amplification, Single Cell, Genomic Language Model,
048 Structural Variation

049

050

051

052 Main

053

054 Single-cell genomics has revolutionized our resolution of biological heterogeneity,
055 enabling the discovery of rare cell types and the reconstruction of clonal evolution in
056 cancer and development [1–3]. However, a single cell contains only 6–7 picograms of
057 DNA—approximately two genome copies—posing significant technical challenges for
058 comprehensive genomic analysis [4, 5]. Consequently, **WGA** remains an unavoidable
059 prerequisite, amplifying DNA 1,000- to 10,000-fold for high-coverage sequencing [6–
060 8]. While **WGA** provides necessary material for downstream analysis, it introduces
061 systematic errors that severely compromise genomic fidelity, particularly for **SV**
062 detection [9–11].

063 The most pernicious of these errors are chimera artifacts—artificial DNA con-
064 structs formed when highly processive polymerases, such as phi29 in **Multiple Dis-**
065 **placement Amplification (MDA)** [12], switch templates during amplification [9–11, 13].
066 These chimeras join discontinuous genomic loci into single molecules, mimicking the
067 structural signatures of biological **translocations (TRAs)** and **INVs** [10]. In long-read
068 sequencing, which is otherwise ideal for resolving complex **SVs**, chimeric reads can
069 constitute 42–76% of the **WGA** data [9], rendering standard **SV** callers unreliable [14–
070 19]. Because these tools rely on alignment heuristics and coverage deviations [14, 20],
071 they frequently misclassify artificial chimeras as genuine variants [21].

072 Distinguishing biological rearrangements from amplification artifacts remains a
073 major computational bottleneck. Current quality control methods rely on hand-
074 crafted features—such as read-pair orientation or localized coverage drops—that fail
075 to capture the sequence-intrinsic patterns of **WGA** errors [11, 13, 22]. This limitation
076 blocks the application of single-cell long-read sequencing in contexts where precision
077 is paramount, such as tracking somatic mosaicism or validating CRISPR off-target
078 effects.

079 We reasoned that **WGA** artifacts possess latent sequence motifs and structural
080 patterns distinct from genomic sequences, learnable without reliance on reference
081 alignment. Here, we present ChimeraLM, a platform-agnostic **Genomic Language**
082 **Model (GLM)** to identify and filter **WGA** artifacts with single-read resolution.

083 Leveraging advances in DNA foundation models [23–26], ChimeraLM treats arti-
084 fact detection as a sequence modeling task rather than an alignment problem. By
085 attending to long-range dependencies and contextual features within raw reads [23–
086 28], ChimeraLM achieves ~90% reduction in chimeric reads while preserving 72–92%
087 of true **SVs**. We demonstrate that this approach restores the fidelity of single-cell **SV**
088 calling, enabling robust characterization of genomic heterogeneity at the single-cell
089 level.

090

091

092

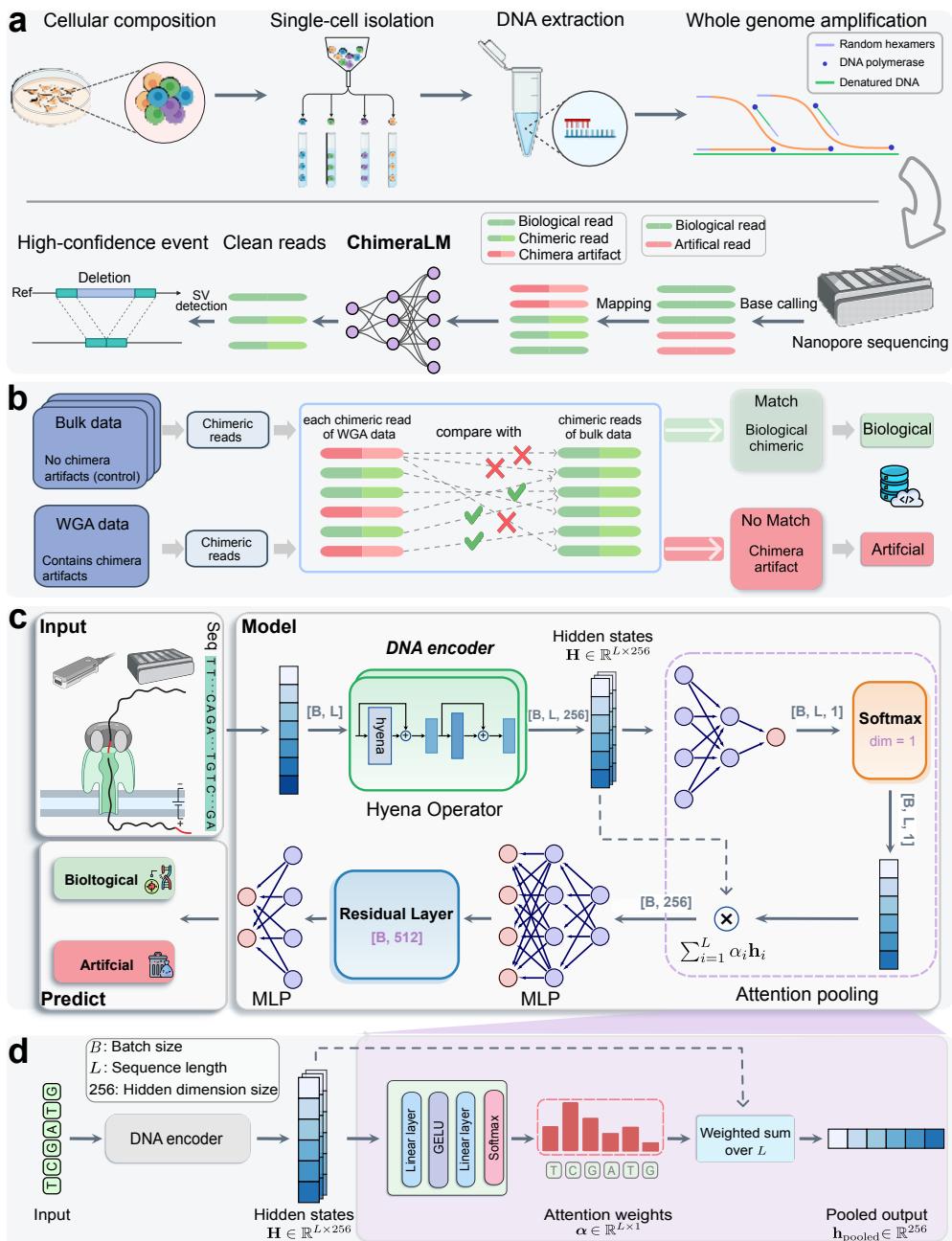


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts. **(a)** Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA. Template switching during amplification generates chimeric artifacts (red) alongside biological reads (green). After base calling and mapping, ChimeraLM classifies chimeric reads as biological or artificial, enabling downstream SV detection on filtered data. **(b)** Ground truth label generation. Chimeric reads from WGA data are compared against chimeric reads from bulk sequencing of the same cell line. Reads matching bulk data are labeled biological (green); non-matching reads are labeled artificial (red). **(c)** ChimeraLM architecture. Input DNA sequences (batch size B , sequence length L) are tokenized at single-nucleotide resolution and encoded into hidden states $H \in \mathbb{R}^{L \times 256}$ through DNA encoder (HyenaDNA [23]). Hyena operators capture long-range dependencies. Attention pooling aggregates position-specific features, and multilayer perceptron (MLP) layers with residual connections process pooled representations for binary classification. **(d)** Attention pooling mechanism. Attention weights $\alpha \in \mathbb{R}^{L \times 1}$ are computed through linear layers with GELU activation and softmax normalization, assigning importance scores to each position. The weighted sum produces a fixed-dimensional representation $h_{\text{pooled}} \in \mathbb{R}^{256}$. Created with BioRender.com.

139 **Results**

140

141 **Overview of ChimeraLM workflow and model architecture**

142 Single-cell genomics relies on **WGA** to obtain sufficient DNA for sequencing (Fig. 1a).
143 The standard workflow includes single-cell isolation, DNA extraction, **WGA**, long-read
144 sequencing (e.g., **Oxford Nanopore Technologies (ONT)**), base calling, and alignment
145 to the reference genome. During amplification, template-switching events introduce
146 artificial chimeric reads, resulting in alignment files containing a mixture of authentic
147 and artificial sequences that can mimic SVs and confound variant detection.
148

ChimeraLM integrates directly into this pipeline as a pre-processing filter, operating
149 after read alignment but before SV detection (Fig. 1a). It evaluates each chimeric
150 read—sequences with multiple alignments to distant genomic locations—and classifies
151 it as either biological (genuine) or artificial (**WGA**-induced). This binary classification
152 enables retention of authentic genomic sequences while removing amplification
153 artifacts prior to variant calling.

Supervised training required a high-confidence labeled dataset (Fig. 1b; Extended
155 Data Fig. 1a). We constructed this dataset using sequencing data from the PC3
156 prostate cancer cell line, which provides both **WGA**-amplified and non-amplified (bulk)
157 genomic data. The key assumption is that bulk sequencing contains only genuine
158 genomic sequences, whereas **WGA** data includes both genuine and artificial chimeras.
159 Chimeric reads from the PC3 **WGA** PromethION dataset were systematically com-
160 pared against three independent bulk datasets (**ONT** PromethION, **ONT** MinION,
161 and **Pacific Biosciences (PacBio)**; see **Methods**). **WGA** reads whose chimeric structures
162 were absent from all three bulk datasets were labeled artificial; reads with structures
163 validated in one or more bulk datasets were labeled biological.

This labeling strategy identified 12,670,396 artificial chimeric reads (zero bulk
165 matches) and 293,180 biological chimeric reads from the **WGA** dataset (Extended
166 Data Fig. 1b). To construct the training dataset, we retained all 293,180 biological
167 reads, subsampled an equal number of artificial reads, and augmented the biological
168 class with 178,748 chimeric reads sampled directly from bulk sequencing data—genuine
169 structural rearrangements unaffected by amplification. This intentional class imbal-
170 ance prioritizes recall of true biological reads, minimizing loss of genuine variants
171 during filtering. The final dataset of 765,108 labeled reads was partitioned into training
172 (70%), validation (20%), and internal test (10%) sets using stratified splitting.
173

The ChimeraLM architecture (Fig. 1c) addresses three technical challenges inher-
174 ent to long-read classification: processing variable-length sequences spanning tens
175 of kilobases, maintaining single-nucleotide resolution to detect abrupt compositional
176 changes at chimeric junctions, and aggregating variable-length representations into
177 fixed-dimensional classification outputs. Input sequences are tokenized at single-
178 nucleotide resolution to preserve complete sequence information. The encoder employs
179 Hyena operators [29], which achieve subquadratic scaling with sequence length,
180 enabling analysis of full-length reads without fragmentation. We initialized the DNA
181 encoder with weights from HyenaDNA [23], a genomic foundation model pre-trained
182 on diverse DNA sequences. An attention pooling mechanism (Fig. 1d) aggregates
183

184

information across the entire read by computing learned, position-specific weights, producing a fixed-dimensional representation. This representation is processed through **MLP** layers with residual connections, and a final softmax layer outputs classification probabilities.

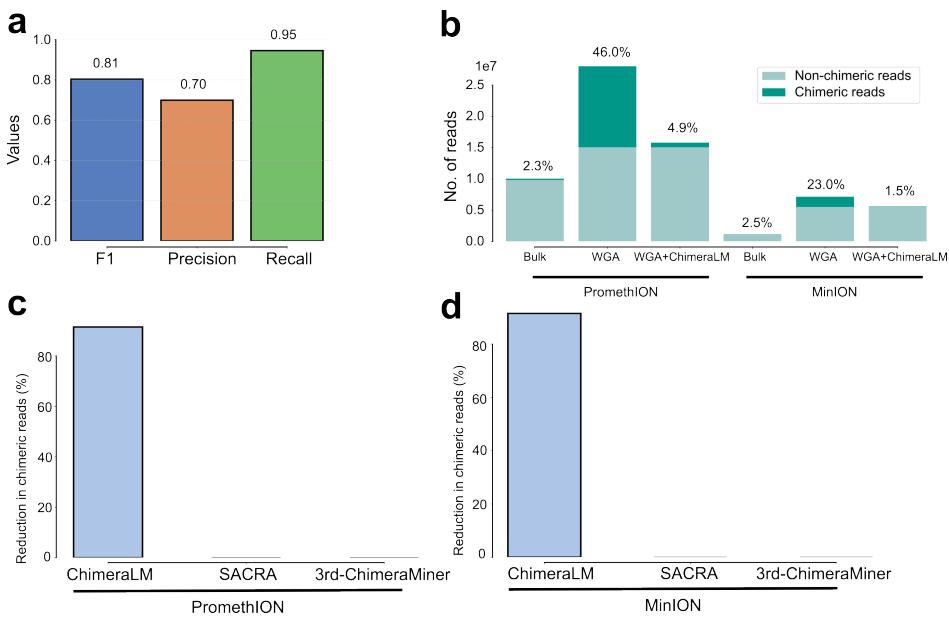


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves recall of 0.95, precision of 0.70, and F1 score of 0.81. (b) Chimeric read reduction across sequencing platforms. Stacked bars show proportions of chimeric (dark teal) and non-chimeric (light teal) reads in bulk, **WGA**, and ChimeraLM-filtered samples. ChimeraLM reduces chimeric read frequencies from 46.0% to 4.9% (PromethION) and from 23.0% to 1.5% (MinION), approaching bulk levels (2.3% and 2.5%, respectively). (c,d) Benchmarking against existing methods on PromethION (c) and MinION (d). ChimeraLM achieves approximately 90% reduction in chimeric reads on both platforms; SACRA and 3rd-ChimeraMiner show no detectable reduction.

ChimeraLM achieves high accuracy and reduces artifacts to near-bulk levels across platforms

We evaluated ChimeraLM’s classification accuracy on the held-out test set, which comprised reads with known biological or artificial status (Fig. 2a). The model achieved an F1 score of 0.81, with recall of 0.95 indicating that 95% of chimeric artifacts were correctly identified—critical for minimizing downstream false-positive **SV** calls—and precision of 0.70 confirming that the majority of flagged reads were true artifacts.

We next assessed practical effectiveness on the full PC3 **WGA** datasets across PromethION and MinION platforms (Fig. 2b). Bulk sequencing established low baseline chimeric read rates (2.3% for PromethION; 2.5% for MinION), whereas **WGA**

231 increased artifact load to 46.0% and 23.0%, respectively. After ChimeraLM filtering,
232 chimeric content dropped to 4.9% (PromethION) and 1.5% (MinION)—10- to 15-
233 fold reductions—while retaining 15.8 million and 5.6 million biological reads. This
234 restoration to near-bulk levels demonstrates effective separation of genuine reads from
235 **WGA**-induced artifacts.

236 We benchmarked ChimeraLM against SACRA [22] and 3rd-ChimeraMiner [13],
237 existing tools for detecting amplification-induced chimeras (Fig. 2c,d). ChimeraLM
238 achieved approximately 90% reduction in chimeric reads on both platforms; neither
239 SACRA nor 3rd-ChimeraMiner showed detectable reduction (0%).

240 The MinION results are particularly notable: this platform served as a com-
241 pletely independent test set, as the model was trained exclusively on PromethION
242 data. Effective generalization to MinION confirms that ChimeraLM learns universal
243 sequence-level features of **WGA**-induced artifacts rather than platform-specific signa-
244 tures. This cross-platform robustness suggests applicability beyond nanopore to other
245 long-read and short-read sequencing technologies.

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

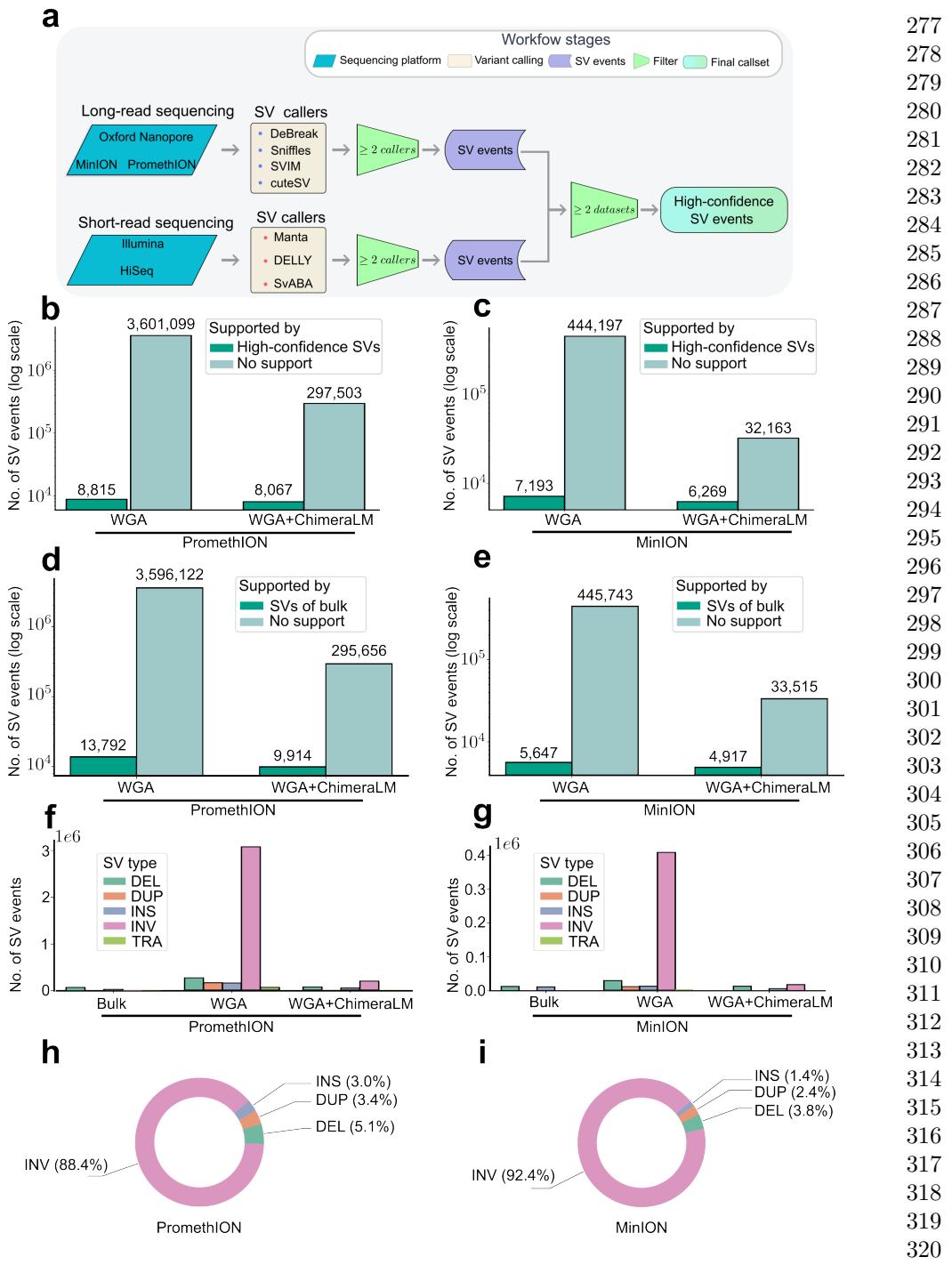


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of high-confidence SV reference dataset. PC3 bulk DNA was sequenced on multiple platforms (ONT PromethION, MinION, and Illumina HiSeq) and analyzed with multiple SV callers. Events detected by ≥2 callers on the same platform and supported by both long-read and short-read data were designated as gold standard SVs. (b,c) SV validation against gold standard for PromethION (b) and MinION (c). Stacked bars show SV calls (log scale) classified as gold standard-supported (dark teal) or unsupported (light teal). ChimeraLM reduces unsupported calls while preserving validated events. (d,e) SV validation against platform-matched long-read bulk sequencing for PromethION (d) and MinION (e). This reference captures true variants that may be missed by short-read data. (f,g) SV type distribution for PromethION (f) and MinION (g). Unfiltered WGA shows elevated INVs, which are reduced to bulk-like levels after ChimeraLM filtering. (h,i) Composition of artifact-supported SVs for PromethION (h) and MinION (i). Pie charts show SV types among events supported exclusively by chimeric reads, representing false positives eliminated by ChimeraLM.

323 **ChimeraLM substantially reduces false-positive structural
324 variant calls**

325 To quantify ChimeraLM’s impact on **SV** calling accuracy, we compared variant calls
326 from unfiltered and ChimeraLM-filtered **WGA** data against two independent reference
327 standards (Fig. 3).

328 We first established a high-confidence gold standard **SV** dataset from bulk PC3
329 DNA sequenced on multiple platforms (**ONT** PromethION, **ONT** MinION, and Illu-
330 mina HiSeq) and analyzed with multiple **SV** callers (Fig. 3a; Extended Data Table 1).
331 **SVs** detected by ≥ 2 callers on the same platform and supported by both long-read
332 and short-read data were retained as gold-standard events.

333 Unfiltered **WGA** data contained extensive false-positive **SVs** (Fig. 3b,c). On Prom-
334 ethION, raw **WGA** produced 3.6 million **SV** calls, of which only 8,815 (0.24%) matched
335 gold standard events—over 99% were artifacts. After ChimeraLM filtering, total calls
336 dropped to 305,570 while retaining 8,067 gold standard events (91.5% of true variants),
337 raising the validation rate to 2.64% (11-fold improvement). MinION data showed simi-
338 lar improvements: calls reduced from 451,390 to 38,432 with validation rate increasing
339 from 1.59% to 16.3% (10-fold improvement) while retaining 87.2% of true variants.

340 We next performed platform-matched validation, comparing **WGA**-derived **SV** calls
341 against long-read bulk sequencing from the same platform (Fig. 3d,e). This reference
342 captures true **SVs** that may be missed by short-read data, providing a more inclusive
343 measure of recall. ChimeraLM increased validation rates from 0.38% to 3.24% on
344 PromethION (8.5-fold improvement) and from 1.25% to 12.79% on MinION (10-fold
345 improvement), while retaining 71.9% and 87.1% of bulk-supported events, respectively.

346 Together, ChimeraLM reduces false-positive **SV** calls by 8–11 fold while preserving
347 72–92% of true variants, substantially enhancing the signal-to-noise ratio in single-cell
348 **SV** discovery.

350
351 **ChimeraLM restores unbiased SV-type distributions**

352 Amplification artifacts can distort the apparent spectrum of **SVs**. We compared **SV**
353 type distributions across bulk, unfiltered **WGA**, and ChimeraLM-filtered datasets
354 (Fig. 3f,g). Bulk sequencing showed balanced proportions of **deletions (DELs)**, **duplica-**
355 **tions (DUPs)**, **insertions (INSs)**, **INVs**, and **TRAs**. Unfiltered **WGA** data exhibited
356 dramatic overrepresentation of **INVs** on both platforms, consistent with amplification
357 artifacts. After ChimeraLM filtering, **SV** distributions were restored toward bulk-like
358 profiles: excessive **INVs** were markedly reduced while other categories remained sta-
359 ble, reflecting selective removal of artifact-supported **INVs** rather than indiscriminate
360 loss of genuine signals.

361 To characterize the artifact composition, we analyzed **SV** calls supported exclu-
362 sively by reads classified as chimera artifacts (Fig. 3h,i). These artifact-supported
363 events were dominated by **INVs**, comprising 88.4% on PromethION and 92.4% on
364 MinION—consistent with template-switching junctions producing inversion-like align-
365 ment signatures. Smaller fractions of **DELs** (5.1% and 3.8%), **DUPs** (3.4% and 2.4%),
366 and **INSs** (3.0% and 1.4%) indicate that **WGA**-induced chimeras can mimic diverse
367 **SV** categories.

Although **INVs** predominate, the presence of other **SV** types among chimeric events indicates that comprehensive filtering—rather than inversion-specific correction—is essential for accurate **SV** detection. By restoring biologically representative **SV** type distributions, ChimeraLM enables robust and interpretable characterization of **SV** in single cells.

369
370
371
372
373
374
375
376
377
378
379
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

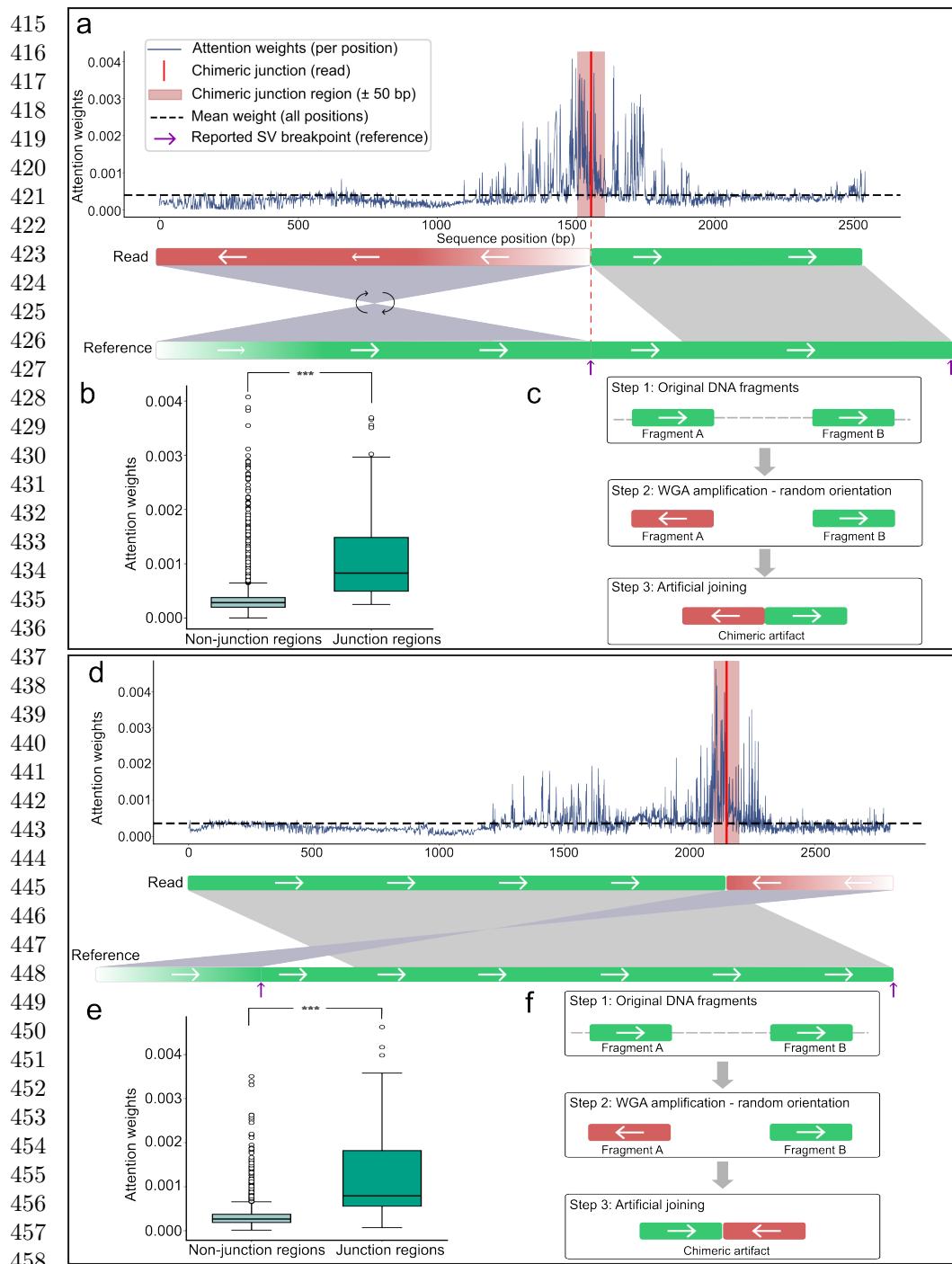


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions.

(a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per position (blue) with mean attention (dashed line). Red vertical lines mark junction positions; pink shading indicates junction region (± 50 bp). Lower panels show read alignments with orientation transitions at junctions (green = forward, red = reverse-complemented). (b,e) Quantitative analysis showing significantly elevated attention weights in junction versus non-junction regions ($P = 5.3 \times 10^{-14}$ and $P = 6.8 \times 10^{-15}$; Wilcoxon rank-sum test). (c,f) Proposed chimera formation mechanisms. During WGA, DNA fragments from distant loci undergo random strand orientation changes before template switching joins them with discordant orientations, producing inversion-like alignment signatures. The two examples illustrate forward-to-reverse and reverse-to-forward transitions.

Attention visualization reveals interpretable classification features	461
	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 investigated whether ChimeraLM’s attention mechanism highlights biologically meaningful regions (Fig. 4). For representative chimeric reads, attention weight profiles showed low baseline values across most positions but pronounced peaks at junction regions where template switching joins DNA fragments from distinct genomic loci (Fig. 4a,d). These peaks coincided with alignment breakpoints characterized by orientation changes between adjacent read segments—the defining signature of WGA-induced artifacts.

Attention weights within junction regions (± 50 bp) were significantly higher than in non-junction regions (Wilcoxon rank-sum test, $P = 5.3 \times 10^{-14}$ and $P = 6.8 \times 10^{-15}$; Fig. 4b,e), indicating that ChimeraLM learns mechanistically relevant features rather than spurious correlations.

Schematic reconstruction of the amplification process supports this interpretation (Fig. 4c,f). During WGA, DNA fragments from distant loci undergo random strand orientation changes before being joined by template switching, producing artificial junctions with discordant orientations that generate inversion-like alignment signatures. The model’s attention peaks effectively capture these orientation discontinuities, localizing chimeric junctions at single-nucleotide resolution.

Discussion

WGA enables genomic analysis from single cells but introduces chimeric artifacts that compromise SV detection. ChimeraLM addresses this challenge through sequence-level classification of biological versus artificial reads, providing an upstream filtering strategy that removes problematic sequences before downstream analysis rather than correcting errors post hoc.

Our results demonstrate several advantages for long-read single-cell sequencing: approximately 90% reduction in chimeric reads while retaining 72–92% of true SVs, 8–11 fold reduction in false-positive SV calls, and consistent performance across PromethION and MinION without platform-specific retraining. The complete failure of existing tools SACRA and 3rd-ChimeraMiner to reduce chimeric content underscores the inadequacy of current heuristic approaches and highlights the advantage of learning sequence-level features directly from data.

ChimeraLM’s effectiveness reflects deep learning’s ability to capture complex sequence patterns difficult to encode in rule-based filters. Traditional quality control methods rely on predefined metrics such as mapping quality or read depth [13, 22], which may not effectively distinguish chimeric artifacts from biological reads. By learning directly from sequence data, ChimeraLM discovers subtle compositional and structural features that differentiate authentic sequences from amplification artifacts. The model also offers interpretability through attention visualization: attention weights concentrate at junctions where template switching joins DNA fragments from distinct loci, matching the known mechanism of chimera formation and building confidence in predictions.

507 The improved reliability of **SV** detection has direct implications for single-cell
508 genomics. Studies of chromosomal instability, clonal evolution, and **SV** burden have
509 been constrained by high false-positive rates in **WGA** data [21, 30]. ChimeraLM
510 enables more confident identification of genuine **SVs**, supporting research in cancer
511 genomics, developmental biology, and aging where single-cell resolution is essential.

512 Several limitations warrant consideration. The current model processes reads
513 independently; integrating contextual features such as coverage or phasing infor-
514 mation may enhance accuracy. **Graphics Processing Unit (GPU)** resources are
515 recommended for large-scale datasets, though **Central Processing Unit (CPU)** infer-
516 ence remains feasible for smaller studies. Future work should prioritize validation
517 across diverse cell types, **WGA** protocols (**Multiple Annealing and Looping-based**
518 **Amplification Cycles (MALBAC)** [31], **Linear Amplification via Transposon Insertion**
519 (**LIANTI**) [5], **Primary Template-directed Amplification (PTA)** [19], and **droplet-based**
520 **MDA (dMDA)** [32]), and sequencing platforms including PacBio HiFi. The sequence-
521 level approach suggests effective transfer to other platforms, though fine-tuning may
522 optimize performance.

523 More broadly, ChimeraLM illustrates the potential of **GLMs** for data quality
524 control. The framework could extend to other amplification-dependent technologies,
525 including cell-free DNA analysis, ancient DNA studies, and metagenomic sequencing
526 from low-biomass samples. Attention-based interpretability also opens opportuni-
527 ties for mechanistic studies of template-switching dynamics, potentially guiding
528 development of improved amplification protocols.

529 In summary, ChimeraLM provides a practical and interpretable framework for
530 improving long-read single-cell genomic data quality, enhancing **SV** detection reliabil-
531 ity while revealing mechanistic features of amplification-induced artifacts.

532

533 **Methods**

534

535 **Cell culture, single-clone preparation, and nanopore sequencing**

536

537 *Cell culture and single-clone establishment*

538 PC3 prostate cancer cells (ATCC® CRL-1435™) were cultured in RPMI-1640 medium
539 supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin at 37 °C
540 with 5% CO₂. To minimize biological heterogeneity, a monoclonal population was
541 established by serial dilution in 96-well plates, ensuring that each culture originated
542 from a single cell. Mycoplasma contamination was routinely tested and confirmed
543 negative prior to DNA extraction.

544

545 *DNA extraction and whole-genome amplification*

546 From the monoclonal population, two types of DNA samples were prepared: a
547 bulk (non-amplified) control and ten single-cell MDA-amplified genomes. Bulk high-
548 molecular-weight DNA was extracted using the Monarch® HMW DNA Extraction
549 Kit for Cells & Blood (New England Biolabs). Individual cells were isolated using
550 1CellDish-60 mm (iBiochips) and amplified using the REPLI-g Advanced DNA Sin-
551 gle Cell Kit (Qiagen) following the manufacturer's protocol. DNA concentration and
552

fragment integrity were assessed with a Qubit 4 fluorometer and Agilent TapeStation (DNA 1000/5000 ScreenTape). Only samples meeting quality standards were used for library construction.	553
	554
	555
	556
	557
Nanopore library preparation and sequencing	558
Sequencing libraries were prepared using the ONT Ligation Sequencing Kit V14 (SQK-LSK114) and sequenced on MinION Mk1C or PromethION P2 Solo devices with R10.4.1 flow cells according to the manufacturer's genomic DNA workflow. Because all single-cell samples originated from the same monoclonal lineage, observed differences between amplified and bulk data primarily reflect MDA-induced artifacts rather than biological variation, providing a controlled experimental setting for downstream analyses.	559
	560
	561
	562
	563
	564
	565
Basecalling and read processing	566
Raw signal files (POD5) were basecalled using Dorado v0.5.0 with the high-accuracy model <code>dna_r10.4.1_e8.2_400bps_hac@v4.3.0</code> [33]. Reads with mean quality < 10 or length < 500 bp were removed. Residual adapters and concatemers were trimmed using Cutadapt v4.0 [34] in two-pass error-tolerant mode. Cleaned reads were aligned to the GRCh38.p13 reference genome using minimap2 v2.26 (<code>map-ont</code> preset) [35]. Resulting BAM files were sorted and indexed with SAMtools v1.16 [36]. Read length and mapping statistics were calculated using NanoPlot v1.46.1 [37]. All samples were processed under identical parameters to ensure consistency across datasets.	567
	568
	569
	570
	571
	572
	573
	574
	575
Chimeric read identification	576
Chimeric reads were identified based on the presence of supplementary alignments in BAM files using the Supplementary Alignment (SA) tag. The SA tag indicates that a read has additional alignments beyond the primary alignment, which is characteristic of chimeric sequences that map to multiple distant genomic locations. To ensure accurate identification, we applied stringent filtering criteria: reads were classified as chimeric only if they (1) were not unmapped, (2) contained the SA tag, (3) were not secondary alignments, and (4) were not supplementary alignments themselves. This filtering approach ensures that only primary alignments with supplementary mapping evidence are considered chimeric, avoiding double-counting of the same chimeric event and excluding low-quality or ambiguous alignments. Reads without the SA tag (single continuous alignments) were classified as non-chimeric. This approach leverages the standard BAM format specification to reliably identify reads with complex alignment patterns.	577
	578
	579
	580
	581
	582
	583
	584
	585
	586
	587
	588
	589
	590
Training data construction	591
Data generation and sources	592
To construct the training dataset, we generated WGA and bulk sequencing data from PC3 cells. The WGA sample was amplified and sequenced on the PromethION P2 platform (ONT), while three independent bulk datasets were produced from non-amplified genomic DNA: bulk PromethION P2, bulk MinION Mk1c (ONT), and bulk PacBio.	593
	594
	595
	596
	597
	598

599 These bulk datasets represent authentic biological sequences free from amplification-
600 induced artifacts. In contrast, **WGA** sequencing includes both genuine genomic reads
601 and artificial chimeras introduced during the amplification process. An additional
602 **WGA** dataset sequenced on the MinION Mk1c platform was reserved exclusively as
603 an independent test set for cross-platform evaluation.

604

605 ***Ground truth annotation and class definition***

606 Ground truth labels were established by systematically comparing chimeric reads from
607 the **WGA** PromethION P2 dataset against those from the three bulk datasets. For
608 each **WGA** chimeric read, all alignment segments—defined by their genomic start
609 and end coordinates—were compared to the corresponding segments of bulk chimeric
610 reads. A **WGA** read was labeled as biological if every segment matched at least one
611 bulk chimeric read within a 1 kb positional tolerance, indicating that the structural
612 configuration is also present in non-amplified DNA. Reads lacking any matching pat-
613 tern across all bulk datasets were labeled as artificial chimeras, presumed to arise
614 from the amplification process. To ensure balanced class representation, additional
615 chimeric reads were randomly sampled from the bulk datasets and labeled as biologi-
616 cal, as these reads originate from genuine genomic rearrangements such as true **SVs**.
617 The final labeled dataset combined the annotated **WGA** PromethION P2 reads with
618 the subsampled bulk chimeric reads and was subsequently partitioned into training,
619 validation, and test sets as described below.

620

621 ***Dataset partitioning and cross-platform validation***

622 The combined labeled dataset, derived from **WGA** PromethION P2 and bulk sequenc-
623 ing data, was divided into training (70%), validation (20%), and internal test (10%)
624 sets using stratified random sampling to maintain class balance. These subsets
625 were used respectively for model training, hyperparameter tuning, and performance
626 evaluation on data from the same sequencing platform.

627

628 To evaluate cross-platform generalization, the complete **WGA** MinION Mk1c
629 dataset was reserved as an independent external test set. This dataset, generated on a
630 different nanopore platform, was never used during model training or internal testing.
631 This two-level evaluation design allowed us to test whether ChimeraLM captures gen-
632 eral sequence features of amplification-induced chimeras rather than platform-specific
633 artifacts.

634

635 **Model architecture**

636 ***DNA encoder***

637 ChimeraLM employs the pre-trained HyenaDNA model [23] as its DNA encoder. This
638 model was pre-trained on large-scale genomic data and provides robust sequence rep-
639 resentations. DNA sequences are tokenized at single-nucleotide resolution, with each
640 base (A, C, G, T, N) mapped to a unique integer token (7, 8, 9, 10, 11, respectively).
641 Special tokens include [CLS]=0, [PAD]=4, and others for sequence processing. Input
642 sequences are truncated at 32,768 bp or padded to enable batch processing.

643

644

For a tokenized input sequence $\mathbf{x} \in \mathbb{Z}^L$, the HyenaDNA generates contextualized hidden representations: 645
646
647

$$\mathbf{H} = \text{HyenaDNA}(\mathbf{x}) \in \mathbb{R}^{L \times 256}$$

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L)$ represents position-wise hidden states with dimension 256. The Hyena operators [29] efficiently capture both local sequence motifs and long-range dependencies essential for distinguishing biological sequences from chimeric artifacts. 650
651
652
653

Attention pooling

To aggregate variable-length sequence representations into fixed-size vectors, ChimeraLM implements attention-based pooling. For hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$, attention weights are computed through a two-layer network: 654
655
656
657
658

$$\begin{aligned}\mathbf{e} &= \text{GELU}(\text{Linear}_{256 \rightarrow 256}(\mathbf{H})) \in \mathbb{R}^{L \times 256} \\ \mathbf{s} &= \text{Linear}_{256 \rightarrow 1}(\mathbf{e}) \in \mathbb{R}^{L \times 1} \\ \boldsymbol{\alpha} &= \text{softmax}(\mathbf{s}) \in \mathbb{R}^{L \times 1}\end{aligned}$$

The pooled representation is the weighted sum of hidden states: 659
660
661
662
663

$$\mathbf{h}_{\text{pooled}} = \sum_{i=1}^L \alpha_i \mathbf{h}_i \in \mathbb{R}^{256}$$

This mechanism assigns learned importance weights to each sequence position, enabling the model to focus on informative regions while accommodating natural variability in read lengths. 664
665
666
667
668

Classification head

The pooled representation is processed through a [MLP](#) with residual connections. The first layer expands dimensionality: 669
670
671
672

$$\mathbf{f}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{Linear}_{256 \rightarrow 512}(\mathbf{h}_{\text{pooled}}))) \in \mathbb{R}^{512}$$

Subsequent residual blocks with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ compute: 673
674
675
676
677
678
679
680

$$\mathbf{f}_{\text{out}} = \text{Dropout}_{0.1}(\text{Linear}_{512 \rightarrow 512}(\text{GELU}(\text{Linear}_{512 \rightarrow 512}(\mathbf{f}_{\text{in}})))) + \mathbf{f}_{\text{in}}$$

where the skip connection enables stable gradient flow during training. The final layer produces binary classification logits: 681
682
683
684
685

$$\mathbf{z} = [z_0, z_1] = \text{Linear}_{512 \rightarrow 2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2$$

where z_0 and z_1 represent logits for biological and artificial chimeric classes, respectively. During inference, the predicted class is $\hat{y} = \text{argmax}_{i \in \{0,1\}} z_i$. 686
687
688
689
690

691 ***Model summary***
 692 The complete ChimeraLM pipeline processes DNA sequences through: (1) single-
 693 nucleotide tokenization, (2) HyenaDNA backbone encoding to generate contextualized
 694 representations, (3) attention pooling to aggregate position-specific features, (4) **MLP**
 695 layers with residual connections to learn classification features, and (5) binary classi-
 696 fication output. The entire model is trained end-to-end using labeled **WGA** and bulk
 697 sequencing data.
 698

699 ***Model training and optimization***
 700

701 ***Training configuration***
 702 ChimeraLM was trained using PyTorch [38] and PyTorch Lightning [39] frameworks.
 703 Input sequences were tokenized using the tokenizer with maximum sequence length of
 704 32,768 bp. Sequences longer than this threshold were truncated; shorter sequences were
 705 padded to enable batch processing. Training employed mixed-precision computation
 706 (bf16) to accelerate training while maintaining numerical stability.
 707

708 ***Optimization procedure***
 709 We used the AdamW optimizer [40] with learning rate $\eta = 1 \times 10^{-4}$ and weight
 710 decay $\lambda = 0.01$. AdamW implements adaptive learning rates with decoupled weight
 711 decay, combining the benefits of Adam optimization with proper L2 regularization.
 712 A ReduceLROnPlateau scheduler dynamically adjusted the learning rate based on
 713 validation loss, reducing it by a factor of 0.1 when no improvement occurred for 10
 714 consecutive epochs. Early stopping with patience of 10 epochs prevented overfitting
 715 by terminating training when validation performance plateaued. A fixed random seed
 716 (12345) ensured reproducibility across training runs.
 717

718 The training objective used cross-entropy loss for binary classification. For a train-
 719 ing example with true class label $y \in \{0, 1\}$ and model logits $\mathbf{z} = [z_0, z_1]$, the loss
 720 is:
 721
$$\mathcal{L}(\mathbf{z}, y) = -\log \left(\frac{\exp(z_y)}{\exp(z_0) + \exp(z_1)} \right) = -z_y + \log(\exp(z_0) + \exp(z_1))$$
 722 where z_0 and z_1 represent logits for biological and artificial chimeric classes, respec-
 723 tively.
 724

725 ***Training implementation***
 726 Training used batch size of 16 sequences with 30 parallel data loading workers. **GPU**
 727 acceleration was employed for efficient processing, with training typically requiring 96-
 728 120 hours depending on dataset size. Model checkpointing saved the best-performing
 729 model based on validation metrics. Configuration management used Hydra [41] to
 730 enable reproducible experimentation.
 731

732
 733
 734
 735
 736

Model evaluation	737
Performance was monitored using accuracy, precision, recall, and F1 score on the validation set after each epoch:	738
Precision = $\frac{TP}{TP + FP}$, Recall = $\frac{TP}{TP + FN}$	739
$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$	740
where TP (true positives) are chimeric reads correctly classified as artificial, TN (true negatives) are biological reads correctly classified as biological, FP (false positives) are biological reads misclassified as artificial, and FN (false negatives) are chimeric reads misclassified as biological. Final model selection was based on best validation performance as determined by early stopping.	741
	742
	743
	744
	745
	746
	747
	748
	749
	750
	751
Model inference and application	752
<i>Inference pipeline</i>	753
To apply ChimeraLM to new WGA sequencing data, the model takes a BAM file as input. Chimeric reads are identified using SA tags and filtered to exclude unmapped, secondary, or supplementary alignments. Each chimeric read sequence is tokenized using the tokenizer (maximum length 32,768 bp, with truncation or padding as needed). The trained model processes sequences in batches, generating two logits $[z_0, z_1]$ for each read corresponding to biological and artificial chimeric classes. Classification is determined by $\hat{y} = \text{argmax}(z_0, z_1)$. ChimeraLM outputs a filtered BAM file containing only reads classified as biological, which can be directly used for downstream analyses including SV calling.	754
	755
	756
	757
	758
	759
	760
	761
	762
	763
	764
Performance evaluation	765
<i>Test set evaluation</i>	766
Final model performance was evaluated on the held-out test set and the independent MinION Mk1c dataset. Metrics (precision, recall, F1 score, accuracy) were computed as described in the training section, where true positives represent chimeric reads correctly classified as artificial and true negatives represent biological reads correctly classified as biological.	767
	768
	769
	770
	771
	772
	773
<i>SV calling</i>	774
SVs were called using multiple tools to ensure comprehensive detection. For long-read data (ONT PromethION P2 and MinION Mk1c), we used Sniffles v2.5 [14, 15], DeBreak v1.2 [16], SVIM v2.0.0 [17], and cuteSV v2.1.1 [18]. For short-read data of the PC3 cell line, we used both the CCLE Illumina whole-genome sequencing dataset and the PRJNA361315 Illumina WGS dataset, processed with Manta v1.6.0 [42], DELLY v1.5.0 [43], and SvABA v1.1.0 [44]. All tools were executed with default recommended parameters.	775
	776
	777
	778
	779
	780
	781
	782

783 ***Gold standard SV dataset construction***

784 A high-confidence gold standard **SV** dataset was generated from bulk PC3 sequencing
785 data to evaluate the impact of ChimeraLM on **SV** detection accuracy (Fig. 3a). All
786 **SV** comparison and breakpoint correction were performed using OctopuSV v0.2.3 [45].
787 We used four datasets: bulk MinION Mk1c, bulk PromethION P2, the CCLE Illumina
788 WGS dataset, and the PRJNA361315 Illumina WGS dataset. Within each dataset, **SV**
789 events supported by at least two independent callers were retained. Variants supported
790 by two or more datasets were designated as gold standard **SVs** for benchmarking.
791

792 ***SV benchmarking analysis***

793 To assess the impact of ChimeraLM on **SV** calling accuracy, we compared **SV** calls from
794 unfiltered **WGA** data and ChimeraLM-filtered **WGA** data against two references: (1)
795 the stringent multi-platform gold standard dataset, and (2) platform-matched long-
796 read bulk sequencing data. Benchmarking was performed using Truvari v4.2.2 [46]
797 with default parameters. **SVs** were considered supported if they matched reference
798 variants within the defined breakpoint tolerance. Validation rates were calculated as
799 the proportion of called **SVs** supported by the reference. This dual benchmarking
800 strategy quantifies both improvements in detecting high-confidence multi-platform
801 **SVs** and the retention of platform-specific true variants.
802

803 ***Benchmarking against existing methods***

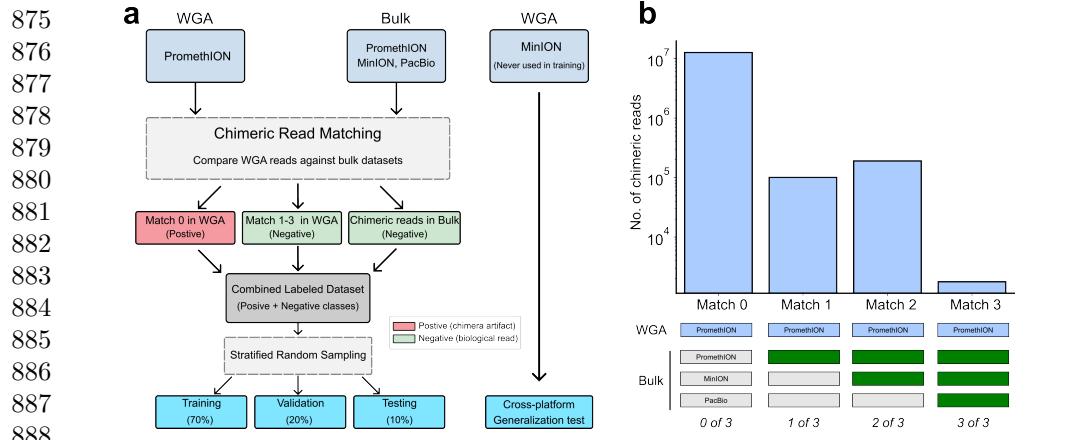
804 ChimeraLM was compared to two existing computational methods for detecting
805 amplification-induced chimeric artifacts: SACRA [22] (GitHub commit 9a2607e) and
806 3rd-ChimeraMiner [13] (GitHub commit 04b5233). Both tools were applied to **WGA**
807 data from PromethION P2 and MinION Mk1c platforms using default parameters as
808 recommended in their documentation. Performance was evaluated by measuring the
809 percentage reduction in chimeric reads relative to unprocessed **WGA** data. Chimeric
810 reads were identified using **WGA** tag-based alignment criteria (reads with **SA** tags
811 indicating split alignments), and reduction rates were calculated as the proportion of
812 chimeric reads removed by each method.
813

814 ***Attention weight analysis***

815 To investigate ChimeraLM’s interpretability, we analyzed attention weights from
816 the pooling mechanism for representative chimeric reads. Attention weights indicate
817 the relative importance assigned to each sequence position during classification. For
818 selected reads, we extracted per-position attention weights and visualized them along-
819 side read alignments to identify whether the model focuses on mechanistically relevant
820 regions.
821

822 Chimeric junction positions were identified from alignment data (defined by break-
823 points in **SA** tags). A window of ± 50 bp surrounding each junction was designated as
824 the junction region. Attention weights within junction region were compared to non-
825 junction regions using the Wilcoxon rank-sum test [47], with statistical significance
826 assessed at $p < 0.001$.
827

Data visualization	829							
Figures were generated using Python with Matplotlib [48] and Seaborn [49].	830							
	831							
Computing resources	832							
Computations were performed on a High Performance Computing (HPC) server with 64-core Intel Xeon Gold 6338 CPU, 256 GB RAM, and two NVIDIA A100 GPUs (80 GB memory each).	833							
	834							
	835							
	836							
	837							
	838							
	839							
Extended Data Table 1 Sequencing and alignment statistics of PC3	840							
	841							
Sample Platform	Reads	Total bases	Total aligned bases	Fraction aligned	Mean length	Mean quality	Average identity	
	($\times 10^6$)	(Gb)	(Gb)		(bp)	(Q)	(%)	
WGA MinION	9.11	14.6	10.4	0.7	1,603	14.3	97.6	842
WGA PromethION	44.69	128.2	69.2	0.5	2,869	14.5	96.1	843
Bulk MinION	0.97	8.1	7.1	0.9	8,310	17.2	97.3	844
Bulk PromethION	8.00	69.9	62.4	0.9	8,732	18.5	97.7	845
								846
								847
								848
								849
								850
Supplementary information.								851
Acknowledgements. We thank Tingyou Wang for guidance on figure preparation. This project was supported in part by NIH grants R35GM142441 and R01CA259388 awarded to RY.								852
Declarations								853
Author Contributions. YL, QG and RY designed the study. YL and QG performed the analysis. QG performed the experiments. YL and QG designed and implemented the model. YL built the command-line tool and documentation. YL, QG and RY wrote the manuscript. RY supervised this work.								854
Data Availability. The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1354861. The dataset includes Oxford Nanopore long-read whole-genome sequencing of PC3 prostate cancer cells and MDA-amplified single-cell derivatives. The individual SRA accessions are as follows: PC3 bulk (MinION Mk1C), SRR35904028; PC3 bulk (PromethION P2), SRR35904029; PC3 10-cell WGA (MinION Mk1C), SRR35904026; PC3 10-cell WGA (PromethION P2), SRR35904027. We can access the data at the following link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA1354861?reviewer=viej6cv6mgbl3n7a9a5k1bsb3							855	
								856
								857
								858
								859
								860
								861
								862
								863
								864
								865
								866
								867
								868
								869
								870
								871
								872
								873
								874



Extended Data Fig. 1 Training dataset construction and ground-truth labeling strategy. (a) Workflow for generating labeled training data. **WGA** PromethION data is compared against three independent bulk sequencing datasets (PromethION, MinION, and PacBio). Reads with no bulk matches (Match 0) are labeled artificial; reads matching one or more bulk datasets (Match 1–3) are labeled biological, along with chimeric reads sampled directly from bulk data. The labeled dataset is split into training (70%), validation (20%), and test (10%) sets. The **WGA** MinION dataset is reserved for independent cross-platform evaluation. (b) Distribution of chimeric read matches. Bar chart shows the number of **WGA** PromethION chimeric reads (log scale) by bulk dataset matches. Match 0 reads ($\sim 10^7$) lacking bulk validation are classified as artificial; Match 1–3 reads with bulk support are classified as biological. The substantial imbalance reflects high prevalence of **WGA**-induced artifacts.

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

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

Acronyms

CPU Central Processing Unit 12

DEL deletion 8

dMDA droplet-based MDA 12

DUP duplication 8

GLM Genomic Language Model	2, 12	921
GPU Graphics Processing Unit	12, 16, 19, 20	922
HPC High Performance Computing	19	923
INS insertion	8	924
INV inversion	1, 2, 7–9	925
LIANTI Linear Amplification via Transposon Insertion	12	926
MALBAC Multiple Annealing and Looping-based Amplification Cycles	12	927
MDA Multiple Displacement Amplification	2	928
MLP multilayer perceptron	3, 5, 15, 16	929
ONT Oxford Nanopore Technologies	4, 7, 8, 13	930
PacBio Pacific Biosciences	4	931
PTA Primary Template-directed Amplification	12	932
SA Supplementary Alignment	13, 17, 18	933
SV Structural Variation	1–3, 5, 7–9, 11, 12, 14, 17, 18	934
TRA translocation	2, 8	935
WGA Whole Genome Amplification	1–8, 10–14, 16–20	936
References		937
[1] Kalf-Ezra, E. <i>et al.</i> Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. <i>Communications Biology</i> 1288 (2024).		938
[2] Navin, N. <i>et al.</i> Tumour evolution inferred by single-cell sequencing. <i>Nature</i> 472 , 90–94 (2011).		939
[3] Sun, C. <i>et al.</i> Mapping recurrent mosaic copy number variation in human neurons. <i>Nature Communications</i> 4220 (2024).		940
[4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. <i>Nature Reviews Genetics</i> 175–188 (2016).		941
[5] Chen, C. <i>et al.</i> Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). <i>Science (new York, N.Y.)</i> 356 , 189–194 (2017).		942
[6] Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives. <i>PLOS Genetics</i> 10 , e1004126 (2014).		943
[7] de Bourcy, C. F. A. <i>et al.</i> A quantitative comparison of single-cell whole genome amplification methods. <i>PLoS ONE</i> e105585 (2014).		944

- 967 [8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification
968 commercial kits using targeted sequencing. *Scientific Reports* 17171 (2021).
- 969
- 970 [9] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement
971 amplification. *Computational and Structural Biotechnology Journal* 1688–1696
972 (2023).
- 973
- 974 [10] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the
975 multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).
- 976
- 977 [11] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displace-
978 ment amplification with whole-genome oxford nanopore sequencing. *Molecular
979 Ecology Resources* e14094 (2025).
- 980
- 981 [12] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple
982 displacement amplification. *Proceedings of the National Academy of Sciences* **99**,
983 5261–5266 (2002).
- 984
- 985 [13] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric
986 sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275
987 (2023).
- 988
- 989 [14] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using
990 single-molecule sequencing. *Nature Methods* 461–468 (2018).
- 991
- 992 [15] Smolka, M. *et al.* Detection of mosaic and population-level structural variants
993 with sniffles2. *Nature Biotechnology* 1571–1580 (2024).
- 994
- 995 [16] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using
996 long sequencing reads with DeBreak. *Nature Communications* 283 (2023).
- 997
- 998 [17] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped
999 long reads. *Bioinformatics* 2907–2915 (2019).
- 1000
- 1001 [18] Jiang, T. *et al.* Long-read-based human genomic structural variation detection
1002 with cuteSV. *Genome Biology* 189 (2020).
- 1003
- 1004 [19] Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with
1005 primary template-directed amplification. *Proceedings of the National Academy of
1006 Sciences* **118**, e2024176118 (2021).
- 1007
- 1008 [20] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and
1009 genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011).
- 1010
- 1011 [21] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection
1012 algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).

- [22] Kiguchi, Y., Nishijima, S., Kumar, N., Hattori, M. & Suda, W. Long-read metagenomics of multiple displacement amplified DNA of low-biomass human gut phageomes by SACRA pre-processing chimeric reads. *DNA Research* **28**, dsab019 (2021). 1013
1014
1015
1016
1017

[23] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution*, Vol. 36, 43177–43201 (Curran Associates, Inc., 2023). 1018
1019

[24] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 287–297 (2025). 1020
1021
1022

[25] Zhou, Z. *et al.* *DNABERT-2: Efficient foundation model and benchmark for multi-species genomes*, 1–24 (OpenReview.net, 2024). 1023
1024
1025

[26] Consens, M. E. *et al.* To transformers and beyond: Large language models for the genome (2023). [arXiv:2311.07621](https://arxiv.org/abs/2311.07621). 1026
1027
1028

[27] Li, Y. *et al.* A genomic language model for chimera artifact detection in nanopore direct rna sequencing. *bioRxiv* (2024). URL <https://www.biorxiv.org/content/early/2024/10/25/2024.10.23.619929>. 1029
1030
1031
1032

[28] Routhier, E. & Mozziconacci, J. Genomics enters the deep learning era. *PeerJ* **10**, e13613 (2022). 1033
1034

[29] Poli, M. *et al.* *Hyena hierarchy: Towards larger convolutional language models*, Vol. 202, 28043–28078 (PMLR, 2023). 1035
1036
1037

[30] Mahmoud, M. *et al.* Structural variant calling: The long and the short of it. *Genome Biology* **20**, 246 (2019). 1038
1039
1040

[31] Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 1622–1626 (2012). 1041
1042
1043
1044

[32] Dippenaar, A. *et al.* Droplet based whole genome amplification for sequencing minute amounts of purified mycobacterium tuberculosis DNA. *Scientific Reports* **14**, 9931 (2024). 1045
1046
1047
1048

[33] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023). 1049
1050

[34] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embojournal* **17**, 10–12 (2011). 1051
1052

[35] Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 3094–3100 (2018). 1053
1054
1055

[36] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* giab008 (2021). 1056
1057
1058

- 1059 [37] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of
1060 long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
- 1061
- 1062 [38] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning*
1063 *library*, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019).
- 1064
- 1065 [39] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub
1066 repository (2019). URL <https://github.com/Lightning-AI/lightning>.
- 1067
- 1068 [40] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).
- 1069
- 1070 [41] Yadan, O. Hydra - a framework for elegantly configuring complex applications.
GitHub repository (2019). URL <https://github.com/facebookresearch/hydra>.
- 1071
- 1072 [42] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for
1073 germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).
- 1074
- 1075 [43] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end
1076 and split-read analysis. *Bioinformatics* i333–i339 (2012).
- 1077
- 1078 [44] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and
1079 indels by local assembly. *Genome Research* 581–591 (2018).
- 1080
- 1081 [45] Guo, Q., Li, Y., Wang, T.-Y., Ramakrishnan, A. & Yang, R. OctopusV and
1082 TentacleSV: A one-stop toolkit for multi-sample, cross-platform structural variant
1083 comparison and analysis. *Bioinformatics* btaf599 (2025).
- 1084
- 1085 [46] English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J.
1086 Truvari: Refined structural variant comparison preserves allelic diversity. *Genome*
1087 *Biology* **23**, 271 (2022).
- 1088
- 1089 [47] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in
1090 python. *Nature Methods* 261–272 (2020).
- 1091
- 1092 [48] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science &*
1093 *Engineering* 90–95 (2007).
- 1094
- 1095 [49] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source*
1096 *Software* 3021 (2021).
- 1097
- 1098
- 1099
- 1100
- 1101
- 1102
- 1103
- 1104