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

Yangyang Li[1†], Qingxiang Guo[1†], Rendong Yang[1,2*]

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

*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu;
Contributing authors: yangyang.li@northwestern.edu;
qingxiang.guo@northwestern.edu;
†These authors contributed equally to this work.

Abstract

Single-cell genomics enables unprecedented cellular heterogeneity insights but faces a fundamental challenge: Whole Genome Amplification (WGA) introduces chimeric artifacts that generate false Structural Variations (SVs), undermining biological interpretations. Current computational methods cannot distinguish amplification-induced artifacts from genuine rearrangements. Here we present ChimeraLM, a genomic language model that learns sequence-level features discriminating biological sequences from WGA artifacts. Validated on nanopore data, ChimeraLM achieves 95% recall with 70% precision and reduces chimeric reads by ∼90% while preserving 72–92% of true SVs. This improves SV validation rates 8–11 fold and eliminates false-positive inversion (INV) bias, restoring SV distributions to bulk-like profiles. Attention visualization reveals ChimeraLM focuses on junction regions with single-base precision, learning interpretable features applicable across sequencing technologies. By enabling confident SV detection at single-cell resolution, ChimeraLM addresses a fundamental data quality barrier in cancer genomics, developmental biology, and precision medicine. Available at https://github.com/ylab-hi/ChimeraLM.

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

Main

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

The most pernicious of these errors are chimera artifacts—artificial DNA constructs formed when highly processive polymerases, such as phi29 in Multiple Displacement Amplification (MDA) [?], switch templates during amplification [? ? ? ?]. These chimeras join discontinuous genomic loci into single molecules, mimicking the structural signatures of biological translocations (TRAs) and INVs [?]. In long-read sequencing, which is otherwise ideal for resolving complex SVs, chimeric reads can constitute 42–76% of the WGA data [?], rendering standard SV callers unreliable [? ? ? ? ? ?]. Because these tools rely on alignment heuristics and coverage deviations [? ?], they frequently misclassify artificial chimeras as genuine variants [?].

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

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

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

Results

Overview of ChimeraLM workflow and model architecture

Single-cell genomics relies on WGA to obtain sufficient DNA for sequencing (Fig. 1a). The standard workflow includes single-cell isolation, DNA extraction, WGA, long-read sequencing (e.g., Oxford Nanopore Technologies (ONT)), base calling, and alignment

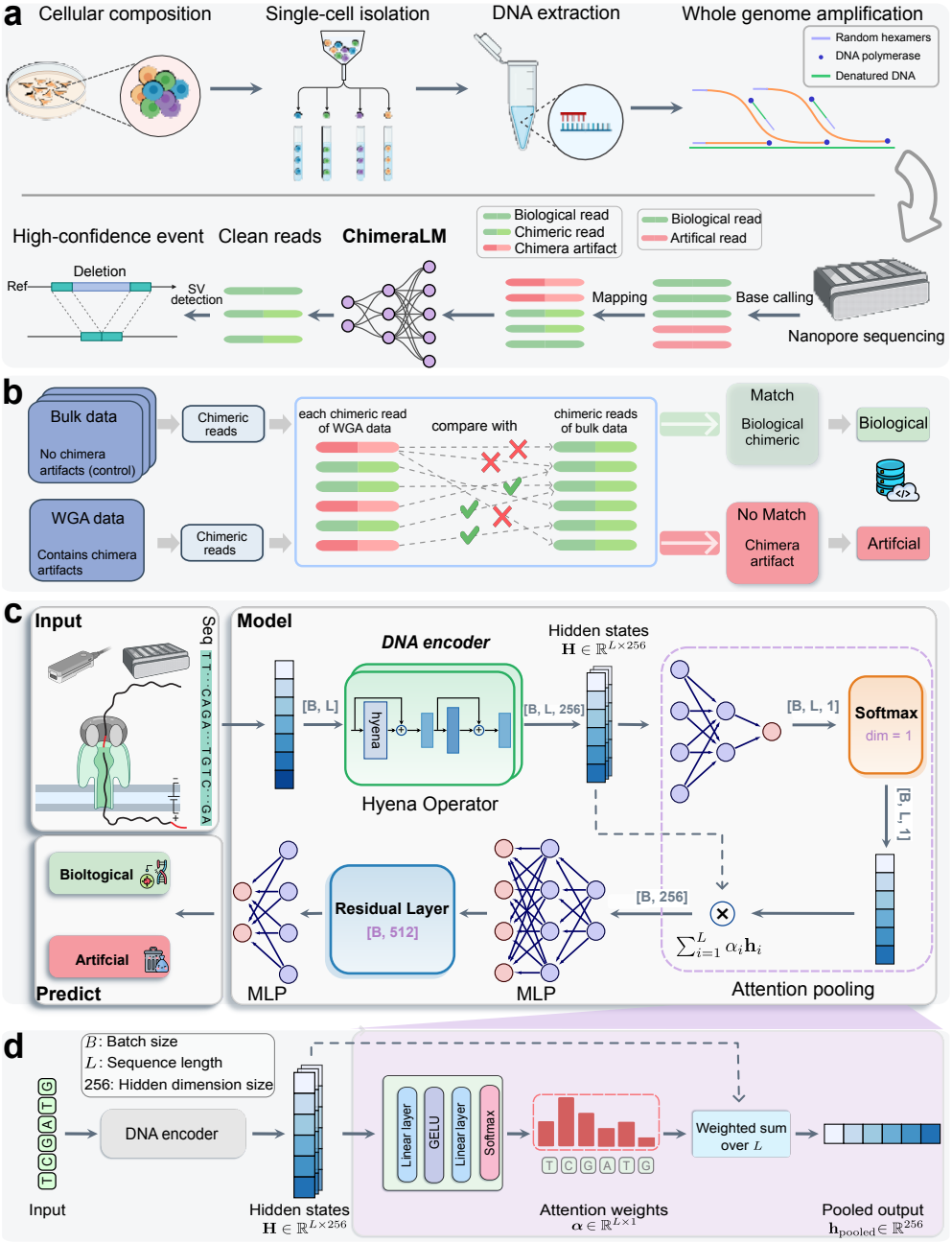


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA for genome amplification. WGA generates chimeric artifacts (red) through template switching during amplification, alongside biological reads (green). After nanopore sequencing, ChimeraLM classifies chimeric reads as biological or artificial, enabling downstream SV detection on clean reads. (b) Ground truth label generation for supervised learning. Chimeric reads from WGA data are compared against all chimeric reads from bulk sequencing data of the same cell line. Reads that match bulk data are labeled as biological (green pathway), while non-matching reads are labeled as chimera artifacts (red pathway). This provides reliable training labels. (c) ChimeraLM architecture. Input DNA sequences (batch size B, sequence length L) are tokenized and encoded into hidden states $\mathbf{H} \in \mathbb{R}^{L\times 256}$ through a DNA encoder (HyenaDNA [**?**]). Hyena operators capture long-range dependencies in genomic sequences. Attention pooling aggregates position-specific features using learned weights. Residual and multilayer perceptron (MLP) layers process pooled features, and a softmax layer outputs binary classification probabilities for biological versus artificial reads. (d) Attention pooling mechanism detail. The DNA encoder (HyenaDNA) transforms input sequences into hidden state $\mathbf{H} \in \mathbb{R}^{L\times 256}$. Attention weights $\boldsymbol{\alpha} \in \mathbb{R}^{L\times 1}$ are computed through linear layers, GELU activation, and softmax normalization, assigning importance scores to each nucleotide position. The weighted sum $\mathbf{h}_{\text{pooled}} = \sum_{i=1}^{L} \alpha_i \mathbf{h}_i$ produces the pooled output $\mathbf{h}_{\text{pooled}} \in \mathbb{R}^{256}$, compressing variable-length sequences into fixed-dimensional representations. Created with BioRender.com.

to the reference genome. During amplification, template-switching events introduce artificial chimeric reads, resulting in alignment files containing a mixture of authentic and artificial sequences that can mimic SVs and confound variant detection.

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

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

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

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

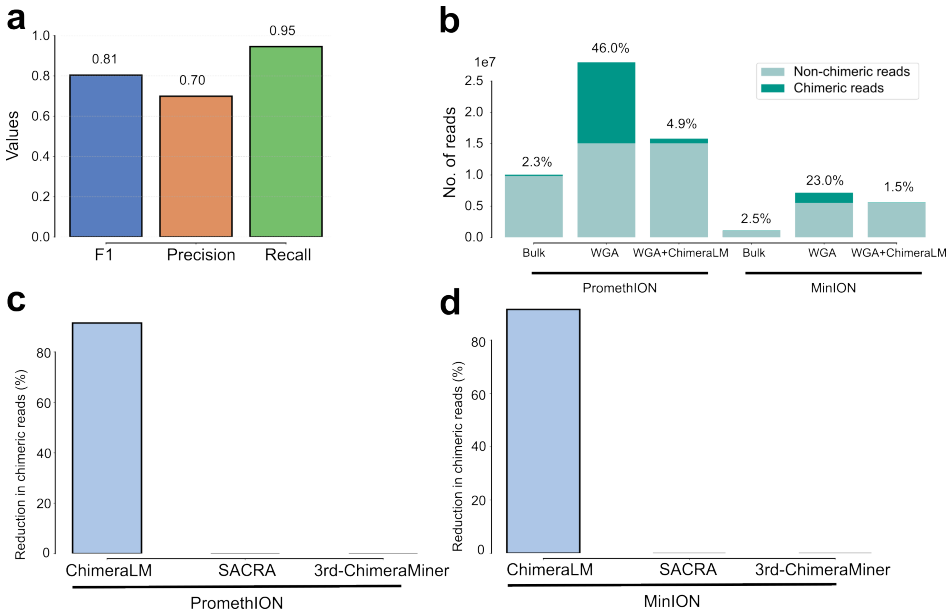


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves high recall (0.95) in identifying chimera artifacts while maintaining acceptable precision (0.70), yielding an F1 score of 0.81 for binary classification of biological versus artificial sequences. (b) Chimeric read reduction across sequencing platforms. Stacked bars show the proportion of chimeric (dark teal) and non-chimeric (light teal) reads in bulk sequencing, WGA-amplified samples, and ChimeraLM-filtered WGA samples. Data from PC3 cell line sequenced on PromethION (left) and MinION (right) platforms demonstrate that 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. ChimeraLM achieves approximately 90% reduction in chimeric reads on both PromethION (c) and MinION (d) platforms, whereas existing computational tools SACRA and 3rd-ChimeraMiner show no detectable reduction in chimeric content.

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 increased artifact load to 46.0% and 23.0%, respectively. After ChimeraLM filtering, chimeric content dropped to 4.9% (PromethION) and 1.5% (MinION)—10- to 15-fold reductions—while retaining 15.8 million and 5.6 million biological reads. This

restoration to near-bulk levels demonstrates effective separation of genuine reads from WGA-induced artifacts.

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

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

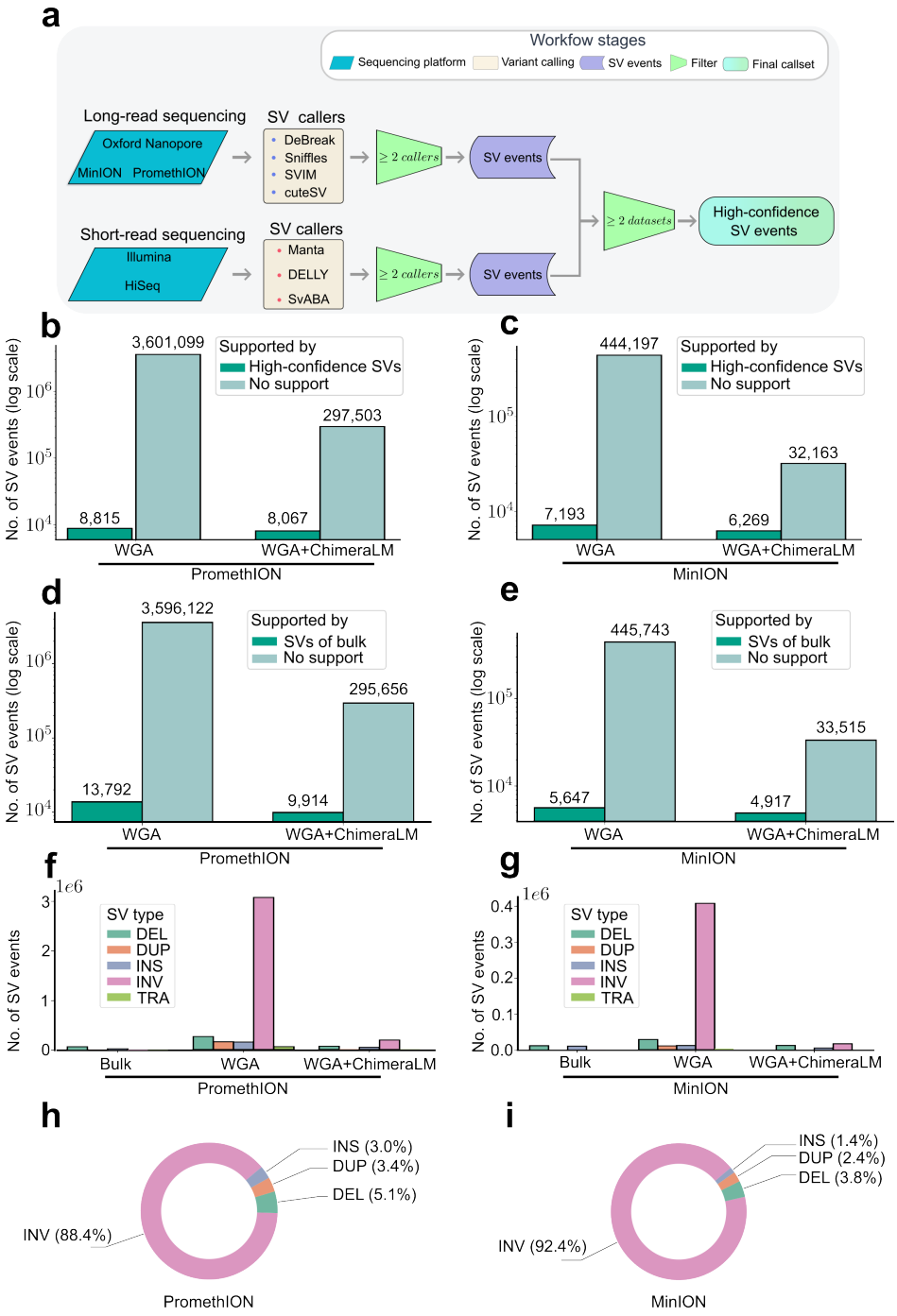


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of high-confidence SV reference dataset. PC3 bulk DNA was sequenced on multiple platforms (ONT PromethION and MinION, Illumina HiSeq) and analyzed with multiple SV calling algorithms. SV events detected by ≥2 callers on the same platform were retained. Events supported by both long-read and short-read platforms were designated as high-confidence gold standard SVs. (b,c) SV validation against multi-platform gold standard. Stacked bars show total SV calls (log scale, numbers above bars) classified as gold standard-supported (dark teal) or unsupported (light teal) for PromethION (b) and MinION (c). ChimeraLM substantially reduces unsupported SV calls while preserving gold standard events. (d,e) SV validation against long-read bulk sequencing (ONT PromethION and MinION). Stacked bars show SV calls classified as bulk-supported (dark teal) or unsupported (light teal) for PromethION (d) and MinION (e). Long-read bulk data from the same platform provides platform-matched validation, capturing true variants that may be specific to long-read detection. (f,g) SV type distribution across processing methods. Bar charts show the number of detected SVs by type: deletion (DEL) (green), duplication (DUP) (orange), insertion (INS) (blue), INV (pink), and TRA (light green) for PromethION (f) and MinION (g). Unfiltered WGA data shows elevated counts across all types, particularly INVs and TRAs, which are reduced to bulk-like levels after ChimeraLM filtering. (h,i) Composition of chimeric artifact-supported SVs. Pie charts show the proportion of SV types among events supported exclusively by reads classified as chimeric artifacts in unfiltered WGA data for PromethION (h) and MinION (i). These represent false-positive SV calls that would be eliminated by ChimeraLM.

ChimeraLM substantially reduces false-positive structural variant calls

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

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

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

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

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

ChimeraLM restores unbiased SV-type distributions

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

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

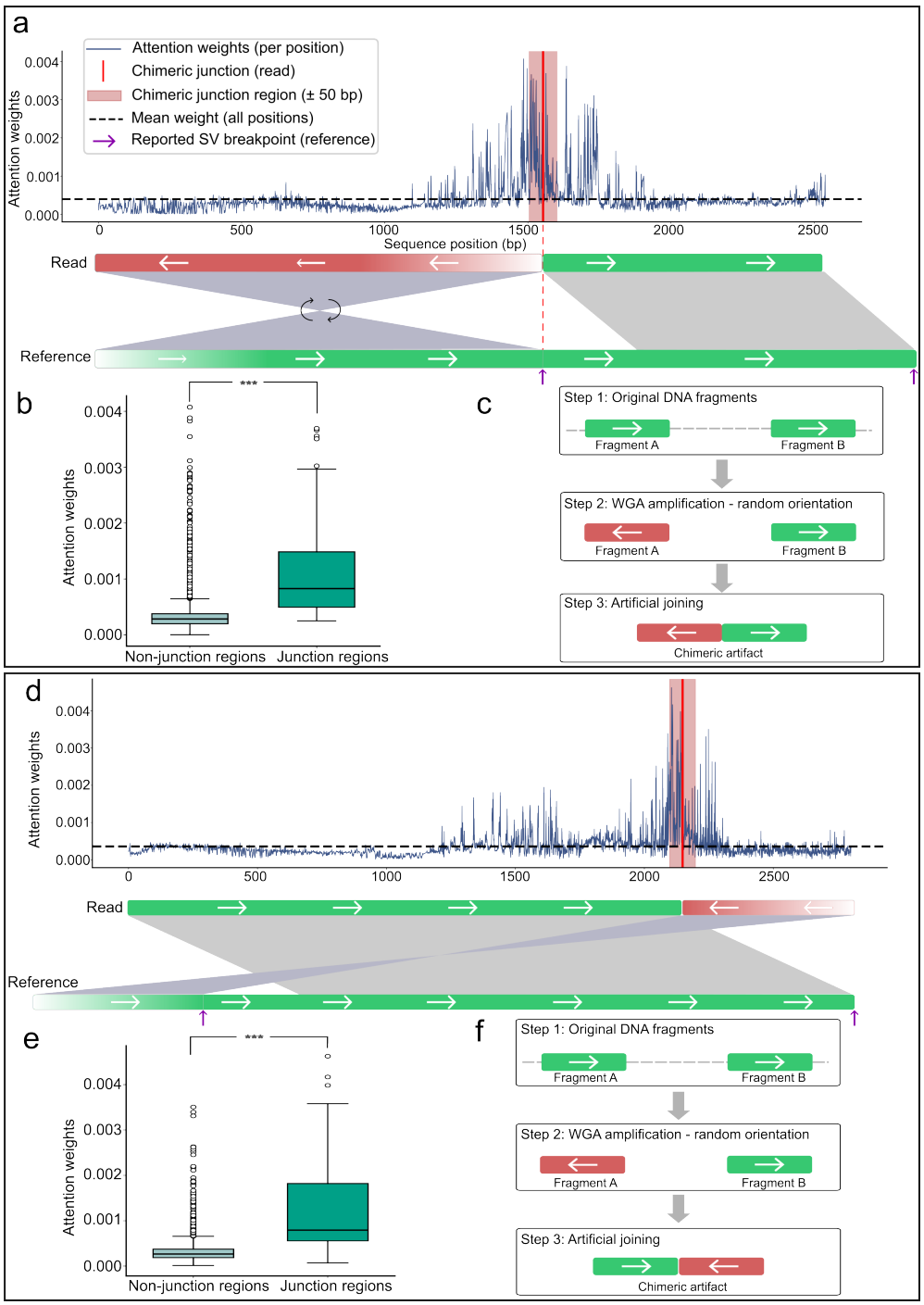


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions. (a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per sequence position (blue line) and mean attention (dashed line). Red vertical lines mark chimeric junction positions, with pink shading indicating junction region (±50 bp). Purple arrows show reported SV breakpoints. Lower panels illustrate read alignments: reads (top bars) show orientation transitions at junctions (green = forward, red = reverse-complemented, arrows indicate strand), while reference genome (bottom bars) maintains continuous forward orientation. Gray regions connect aligned segments. (b,e) Quantitative attention analysis. Box plots show significantly elevated attention weights in junction region versus non-junction regions for both examples ($p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$, respectively; Wilcoxon rank-sum test). (c,f) Proposed chimera formation mechanisms. Step 1: Original DNA fragments from distant genomic loci exist in forward orientation. Step 2: During WGA, one or both fragments may undergo random reverse-complementation. Step 3: Template switching joins the fragments with discordant orientations, creating chimeric artifacts. The two examples illustrate different orientation patterns (forward-to-reverse vs reverse-to-forward transitions) arising from random strand selection during amplification.

Attention visualization reveals interpretable classification features

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, $P = 5.3\times10^{-14}$ and $P = 6.8\times10^{-15}$; 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 has enabled 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, substantially improving SV calling accuracy before downstream analysis. This upstream filtering strategy—removing problematic sequences at the read level rather than correcting errors post hoc—provides a practical solution for single-cell genomics laboratories.

Our results demonstrate several key advantages of ChimeraLM for long-read single-cell sequencing. The method achieves approximately 90% reduction in chimeric reads across nanopore platforms while retaining 72–92% of true SVs. It reduces false-positive SV calls by 8–11 fold, enabling researchers to focus on biologically relevant variants without manually filtering thousands of artifacts. Moreover, ChimeraLM performs consistently across PromethION and MinION without platform-specific retraining, indicating that it captures generalizable sequence features of WGA-induced chimeras. These results underscore the model's robustness across diverse datasets and sequencing conditions.

ChimeraLM's effectiveness reflects the ability of deep learning models to capture complex sequence patterns that are difficult to encode in rule-based filters. Traditional quality control methods rely on predefined metrics such as mapping quality or read depth [**? ?**], 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 genomic sequences from amplification artifacts. Furthermore, the model offers interpretability through attention visualization, allowing researchers to examine which sequence regions drive

classification. Attention weights can concentrate sharply at junctions where template switching joins DNA fragments from distinct loci, matching the known mechanism of chimera formation. Some reads show more diffuse attention distributions, suggesting that ChimeraLM integrates multiple complementary cues—such as junction orientation, compositional biases, and local sequence context—to classify diverse artifact types. This interpretability builds confidence in the model's predictions and provides a lens for probing the molecular processes underlying amplification-induced artifacts.

The improved reliability of SV detection has direct implications for single-cell genomics. Studies of chromosomal instability, clonal evolution, and SV burden in individual cells have long been constrained by high false-positive rates in WGA data [**?** **?**]. ChimeraLM enables more confident identification of genuine SVs, supporting research in cancer genomics, developmental biology, and aging where single-cell resolution is essential for understanding cellular heterogeneity. Although the current model processes reads independently, integrating additional contextual features—such as coverage, mate-pair, or phasing information—could further enhance accuracy. Graphics Processing Unit (GPU) resources are recommended for large-scale datasets, while Central Processing Unit (CPU) inference remains feasible for smaller studies; runtime optimization and model compression may improve accessibility for broader use.

Future work should prioritize validation across diverse biological and technical contexts. First, testing on multiple cell types (primary, stem, or immune cells) and WGA protocols (Multiple Annealing and Looping-based Amplification Cycles (MALBAC), Linear Amplification via Transposon Insertion (LIANTI), Primary Template-directed Amplification (PTA)) will establish biological generalizability. Second, validation on additional sequencing platforms—including PacBio HiFi, Illumina linked-reads, and emerging long-read technologies—will confirm the platform-agnostic design principle. The sequence-level approach suggests ChimeraLM should transfer effectively to any platform, though platform-specific fine-tuning may optimize performance. Third, the interpretability of attention-based models could be leveraged to investigate mechanisms of chimera formation: large-scale analysis of attention patterns may reveal recurrent sequence motifs or genomic contexts associated with template switching, guiding the development of improved amplification protocols. More broadly, ChimeraLM illustrates the potential of GLMs for data quality control applications [**?**]. Architectural innovations such as the Hyena operator for efficient long-range modeling [**?**] may have utility beyond chimera detection, addressing challenges such as contamination, adapter artifacts, and systematic sequencing errors across multiple platforms.

Looking ahead, ChimeraLM's framework could extend beyond single-cell genomics to address quality control challenges in other amplification-dependent technologies, including cell-free DNA analysis, ancient DNA studies, and metagenomic sequencing from low-biomass samples. The model's interpretability through attention visualization also opens opportunities for mechanistic studies of polymerase fidelity and template-switching dynamics across different amplification protocols. Furthermore, integration with emerging single-cell multi-omics platforms could enable simultaneous quality control across genomic, transcriptomic, and epigenomic data layers, providing a unified framework for artifact detection in complex single-cell experiments.

ChimeraLM thus provides a practical and interpretable framework for improving long-read single-cell genomic data quality. By removing WGA-induced chimeric artifacts at the read level and revealing the mechanistic features that drive them, the method not only enhances SV detection reliability but also deepens understanding of amplification-induced bias in single-cell genomics.

Methods

Cell culture, single-clone preparation, and nanopore sequencing

Cell culture and single-clone establishment

PC3 prostate cancer cells (ATCC® CRL-1435™) were cultured in RPMI-1640 medium supplemented with 10% fetal bovine serum and 1% penicillin–streptomycin at 37 °C with 5% CO_2. To minimize biological heterogeneity, a monoclonal population was established by serial dilution in 96-well plates, ensuring that each culture originated from a single cell. Mycoplasma contamination was routinely tested and confirmed negative prior to DNA extraction.

DNA extraction and whole-genome amplification

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

Nanopore library preparation and sequencing

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.

Basecalling and read processing

Raw signal files (POD5) were basecalled using Dorado v0.5.0 with the high-accuracy model `dna_r10.4.1_e8.2_400bps_hac@v4.3.0` [**?**]. Reads with mean quality < 10 or length < 500 bp were removed. Residual adapters and concatemers were trimmed using Cutadapt v4.0 [**?**] in two-pass error-tolerant mode. Cleaned reads were aligned to the GRCh38.p13 reference genome using minimap2 v2.26 (`map-ont` preset) [**?**]. Resulting BAM files were sorted and indexed with SAMtools v1.16 [**?**]. Read length

and mapping statistics were calculated using NanoPlot v1.46.1 [**?**]. All samples were processed under identical parameters to ensure consistency across datasets.

Chimeric read identification

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.

Training data construction

Data generation and sources

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

Ground truth annotation and class definition

Ground truth labels were established by systematically comparing chimeric reads from the WGA PromethION P2 dataset against those from the three bulk datasets. For each WGA chimeric read, all alignment segments—defined by their genomic start and end coordinates—were compared to the corresponding segments of bulk chimeric reads. A WGA read was labeled as biological if every segment matched at least one bulk chimeric read within a 1 kb positional tolerance, indicating that the structural configuration is also present in non-amplified DNA. Reads lacking any matching pattern across all bulk datasets were labeled as artificial chimeras, presumed to arise from the amplification process. To ensure balanced class representation, additional chimeric reads were randomly sampled from the bulk datasets and labeled as biological, as these reads originate from genuine genomic rearrangements such as true SVs. The final labeled dataset combined the annotated WGA PromethION P2 reads with

the subsampled bulk chimeric reads and was subsequently partitioned into training, validation, and test sets as described below.

Dataset partitioning and cross-platform validation

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

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

Model architecture

DNA encoder

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

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

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

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

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:

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

The pooled representation is the weighted sum of hidden states:

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

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.

Classification head

The pooled representation is processed through a MLP with residual connections. The first layer expands dimensionality:

$$\mathbf{f}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{Linear}_{256\to512}(\mathbf{h}_{\text{pooled}}))) \in \mathbb{R}^{512}$$

Subsequent residual blocks with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ compute:

$$\mathbf{f}_{\text{out}} = \text{Dropout}_{0.1}(\text{Linear}_{512\to512}(\text{GELU}(\text{Linear}_{512\to512}(\mathbf{f}_{\text{in}})))) + \mathbf{f}_{\text{in}}$$

where the skip connection enables stable gradient flow during training. The final layer produces binary classification logits:

$$\mathbf{z} = [z_0, z_1] = \text{Linear}_{512\to2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2$$

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

Model summary

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

Model training and optimization

Training configuration

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

Optimization procedure

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

The training objective used cross-entropy loss for binary classification. For a training example with true class label $y \in \{0, 1\}$ and model logits $\mathbf{z} = [z_0, z_1]$, the loss is:

$$\mathcal{L}(\mathbf{z}, y) = -\log\left(\frac{\exp(z_y)}{\exp(z_0) + \exp(z_1)}\right) = -z_y + \log(\exp(z_0) + \exp(z_1))$$

where z_0 and z_1 represent logits for biological and artificial chimeric classes, respectively.

Training implementation

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

Model evaluation

Performance was monitored using accuracy, precision, recall, and F1 score on the validation set after each epoch:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

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

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.

Model inference and application

Inference pipeline

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.

Performance evaluation

Test set evaluation

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.

SV calling

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 [**?** **?**], DeBreak v1.2 [**?**], SVIM v2.0.0 [**?**], and cuteSV v2.1.1 [**?**]. 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 [**?**], DELLY v1.5.0 [**?**], and SvABA v1.1.0 [**?**]. All tools were executed with default recommended parameters.

Gold standard SV dataset construction

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

SV benchmarking analysis

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

Benchmarking against existing methods

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

Attention weight analysis

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

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

Data visualization

Figures were generated using Python with Matplotlib [**?**] and Seaborn [**?**].

Computing resources

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).

Supplementary information.

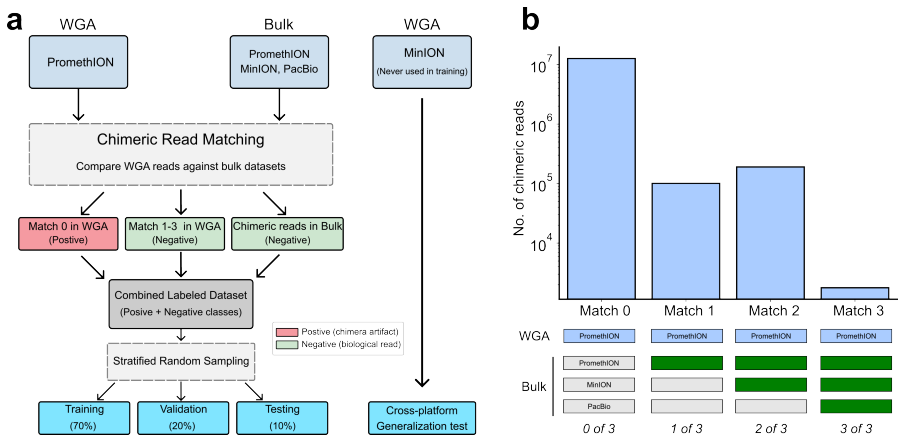
Acknowledgements. We thank Tingyou Wang for guidance on figure preparation. This project was supported in part by NIH grants R35GM142441 and R01CA259388 awarded to RY.

Declarations

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.

Extended Data Table 1 Sequencing and alignment statistics of PC3

Sample	Platform	Reads ($\times 10^6$)	Total bases (Gb)	Total bases aligned (Gb)	Fraction aligned	Mean length (bp)	Mean quality (Q)	Average identity (%)
WGA	MinION	9.11	14.6	10.4	0.7	1,603	14.3	97.6
WGA	PromethION	44.69	128.2	69.2	0.5	2,869	14.5	96.1
Bulk	MinION	0.97	8.1	7.1	0.9	8,310	17.2	97.3
Bulk	PromethION	8.00	69.9	62.4	0.9	8,732	18.5	97.7



Extended Data Fig. 1 Training dataset construction and ground-truth labeling strategy for PC3 cell line. (a) Schematic workflow for generating labeled training data. WGA PromethION data containing both biological and artificial chimeric reads is compared against three independent bulk sequencing datasets from the same cell line (PromethION, MinION, and PacBio platforms). Chimeric reads are classified through systematic matching: reads with no matches across all bulk datasets (Match 0) are labeled as artificial chimeras (positive class, red); reads matching one or more bulk datasets (Match 1–3) are labeled as biological reads (negative class, green), along with chimeric reads sampled directly from bulk data. The combined labeled dataset undergoes stratified random sampling to generate training (70%), validation (20%), and testing (10%) sets for model development. The WGA MinION dataset is reserved as an independent cross-platform generalization test set. (b) Distribution of chimeric read matches between WGA and bulk sequencing datasets. Bar chart showing the number of chimeric reads (y-axis, log scale) grouped by how many bulk datasets (x-axis) contained matching chimeric structures when comparing WGA PromethION reads against bulk sequencing data. "Match 0" indicates reads with no matches in any bulk dataset (classified as artificial chimeras, $\sim 10^7$ reads), whereas "Match 1–3" indicate reads with matches in one, two, or all three bulk datasets (classified as biological reads, $\sim 10^5$ reads each). Color-coded boxes below bars indicate which bulk platforms validated each read category: PromethION (light blue), MinION (white), and PacBio (white); green boxes indicate platform-specific validation. The substantial imbalance between Match 0 ($\sim 10^7$) and Match 1–3 categories ($\sim 10^5$ each) reflects the high prevalence of WGA-induced artifacts, necessitating balanced subsampling for supervised learning.

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=viej6cv6mgbli3n7a9a5k1bsb3

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 7, 8
DUP duplication 7, 8

GLM Genomic Language Model 2, 12
GPU Graphics Processing Unit 12, 17, 19, 21

HPC High Performance Computing 19

INS insertion 7, 8
INV inversion 1, 2, 7–9

LIANTI Linear Amplification via Transposon Insertion 12

MALBAC Multiple Annealing and Looping-based Amplification Cycles 12
MDA Multiple Displacement Amplification 2
MLP multilayer perceptron 3, 4, 16

ONT Oxford Nanopore Technologies 2, 4, 7, 8, 13, 14

PacBio Pacific Biosciences 4
PTA Primary Template-directed Amplification 12

SA Supplementary Alignment 14, 18, 19
SV Structural Variation 1–3, 5, 7–13, 15, 18, 19

TRA translocation 2, 7, 8

WGA Whole Genome Amplification 1–8, 10–16, 18–20

References

[1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. *Communications Biology* 1288 (2024).

[2] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

[3] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons. *Nature Communications* 4220 (2024).

[4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* 175–188 (2016).

[5] Chen, C. *et al.* Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science (new York, N.Y.)* **356**, 189–194 (2017).

[6] Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives. *PLOS Genetics* **10**, e1004126 (2014).

[7] de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* e105585 (2014).

[8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. *Scientific Reports* 17171 (2021).

[9] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal* 1688–1696 (2023).

[10] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).

[11] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displacement amplification with whole-genome oxford nanopore sequencing. *Molecular Ecology Resources* e14094 (2025).

[12] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* **99**, 5261–5266 (2002).

[13] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275 (2023).

[14] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 461–468 (2018).

[15] Smolka, M. *et al.* Detection of mosaic and population-level structural variants with sniffles2. *Nature Biotechnology* 1571–1580 (2024).

[16] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nature Communications* 283 (2023).

[17] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* 2907–2915 (2019).

[18] Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* 189 (2020).

[19] Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences* **118**, e2024176118 (2021).

[20] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011).

[21] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection 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).

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

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

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

[26] Consens, M. E. *et al.* To transformers and beyond: Large language models for the genome (2023). arXiv:2311.07621.

[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.

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

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

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

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

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

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

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

[35] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).

[36] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning library*, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019).

[37] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub repository (2019). URL https://github.com/Lightning-AI/lightning.

[38] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).

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

[40] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).

[41] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* i333–i339 (2012).

[42] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Research* 581–591 (2018).

[43] Guo, Q., Li, Y., Wang, T.-Y., Ramakrishnan, A. & Yang, R. OctopuSV and TentacleSV: A one-stop toolkit for multi-sample, cross-platform structural variant comparison and analysis. *Bioinformatics* btaf599 (2025).

[44] English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined structural variant comparison preserves allelic diversity. *Genome Biology* **23**, 271 (2022).

[45] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 261–272 (2020).

[46] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 90–95 (2007).

[47] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* 3021 (2021).