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 genomic analysis relies on Whole Genome Amplification (WGA) to generate sufficient DNA for sequencing, but this process introduces chimeric artifacts that manifest as false-positive Structural Variations (SVs) and compromise downstream interpretation. Here we present ChimeraLM, a genomic language model that identifies and removes WGA-induced chimeric reads from long-read sequencing data. ChimeraLM uses a model architecture based on Hyena operators to analyze DNA sequences at single-nucleotide resolution, learning generalizable sequence features that distinguish genuine biological sequences from amplification-induced artifacts. When applied to nanopore sequencing data from WGA-amplified cells, ChimeraLM reduces chimeric read content by approximately 90% while retaining 87-92% of true SVs. This filtering improves SV validation rates 10-16 fold and normalizes SV type distributions toward bulk sequencing profiles, eliminating the characteristic false-positive inversion (INV) bias in unprocessed WGA data. Attention weight analysis reveals that ChimeraLM can focus on chimeric junction regions, learning biologically interpretable sequence features. ChimeraLM addresses a fundamental bottleneck in single-cell genomics, enabling more confident detection of chromosomal instability and SV in applications across cancer biology, developmental biology, and neuroscience. The software is 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 understanding of cellular heterogeneity by enabling characterization of individual cells rather than bulk populations [? ? ? ?]. This approach has proven instrumental in uncovering rare cell types [?], tracking developmental trajectories [?], and elucidating tumor evolution through clonal architecture analysis. However, the limited DNA content in a single cell—typically only 6-7 picograms containing approximately two copies of the 3-billion-base-pair human genome—poses significant technical challenges for comprehensive genomic analysis [? ? ?]. To overcome this limitation, WGA has become essential for single-cell genomic studies [? ? ? ? ?]. Various WGA techniques have been developed, each with distinct amplification mechanisms and characteristic error profiles. Multiple Displacement Amplification (MDA), introduced by Dean et al. [?], utilizes the highly processive phi29 DNA polymerase to achieve isothermal amplification with products exceeding 10 kb, though it suffers from pronounced amplification bias and chimera formation [? ?]. Degenerate Oligonucleotide-Primed PCR (DOP-PCR), pioneered by Telenius et al. [?], employs thermocycling with degenerate primers to achieve more uniform coverage but generates shorter amplicons. Multiple Annealing and Looping-based Amplification Cycles (MALBAC) combines quasi-linear preamplification with exponential amplification to reduce bias [?], while Linear Amplification via Transposon Insertion (LIANTI) uses transposon insertion to create defined amplification origins, significantly improving uniformity and reducing artifacts [?]. More recently, Primary Template-directed Amplification (PTA) [?] and droplet-based MDA (dMDA) [? ?] have emerged as promising alternatives that modify reaction conditions to suppress chimera formation, though these methods require specialized equipment and protocols that have limited their widespread adoption. These amplification methods can increase DNA content by several orders of magnitude (typically 1,000- to 10,000-fold), generating sufficient material for high-coverage sequencing necessary for reliable variant calling, copy number analysis, and SV detection [? ? ? ? ? ?].

Accurate single-cell genomics is particularly critical for multiple applications where false-positive SVs can lead to incorrect biological conclusions. In cancer research, distinguishing genuine clonal evolution patterns from amplification artifacts is essential for understanding tumor heterogeneity and therapeutic resistance [?]. In developmental biology, accurate detection of somatic mosaicism enables the reconstruction of lineage relationships and identification of pathogenic mutations in rare cell populations. For CRISPR-based genome editing, single-cell analysis with reliable SV detection is crucial for comprehensive assessment of off-target effects and ensuring genomic stability [?]. However, false-positive SVs introduced during amplification can confound these analyses, leading to misinterpretation of genomic rearrangements and their biological significance [? ?].

Despite its critical role, WGA introduces systematic artifacts that significantly impact downstream analyses [? ? ? ? ?]. Among the most problematic are chimeric sequences—artificial DNA constructs formed through template switching during amplification. During MDA, the highly processive phi29 polymerase can dissociate from one genomic template and reinitiate synthesis on a spatially proximate but genomically distant template [? ?]. This phenomenon is exacerbated by the branching nature of MDA, where multiple DNA synthesis reactions occur simultaneously in a densely packed reaction environment, increasing the probability of illegitimate template switching [?]. Critically, even with WGA technological advances, chimeric artifacts remain highly prevalent in single-cell long-read sequencing data [? ?]. Lasken and Stockwell [?] demonstrated that chimera formation occurs through both strand displacement and branch migration mechanisms, with chimeric junctions often occurring at sites of microhomology. These chimeric artifacts manifest as apparent SVs—including deletions (DELs), insertions (INSs), INVs, and translocations (TRAs)—that do not exist in the original cell [? ?], posing substantial challenges for accurate SV detection in single-cell studies. Early work by Pinard et al. [?] documented significant amplification bias and the presence of chimeric products in MDA, demonstrating that certain genomic regions can be over- or under-represented by orders of magnitude.

The advent of long-read sequencing technologies, particularly Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms, has transformed SV detection by enabling direct observation of structural rearrangements that span kilobases to megabases. Numerous computational tools have been developed to detect SVs from long-read data, including Sniffles2 [? ?], DeBreak [?], SVIM [?], and cuteSV [?]. These methods typically employ read alignment analysis, split-read detection, and local assembly strategies to identify SV signatures [?]. However, distinguishing genuine biological SVs from WGA-induced chimeric artifacts remains challenging [? ? ? ?].

Current computational approaches for identifying WGA-induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [? ?]. However, these heuristic methods often fail to distinguish genuine SVs from amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns, occur in repetitive genomic regions, or involve multiple genomic loci [? ?]. This lack of robust, automated artifact detection has limited the reliability of SV analysis in single-cell studies and hindered the full realization of single-cell genomics' potential for studying somatic mosaicism, tumor evolution, and rare cell populations.

The emergence of deep learning, particularly language models based on transformer architectures, has demonstrated remarkable success in genomics applications [? ? ? ?]. Recent genomic language models have shown the ability to learn complex sequence patterns and contextual relationships in DNA sequences, enabling improved performance in tasks such as regulatory element prediction, variant effect prediction, and functional annotation [? ? ?]. These models treat DNA sequences analogously to natural language, learning representations that capture both local motifs and long-range dependenies [**?**]. By training on large-scale genomic datasets, such models can internalize patterns of genuine biological sequences, including characteristic features of repetitive elements, chromatin structure, and sequence composition biases.

Here, we developed ChimeraLM, a genomic language model specifically designed to detect chimeric artifacts introduced by WGA. By leveraging deep learning to capture sequence patterns, structural features, and contextual information in genomic reads [**? ? ? ? ?**], ChimeraLM effectively distinguishes genuine biological sequences from WGA-induced chimeric artifacts. We demonstrate that ChimeraLM achieves superior performance compared to existing methods and substantially improves the reliability of SV detection in single-cell genomic studies, thereby enabling accurate SV analysis at single-cell resolution.

Results

Overview of ChimeraLM workflow and model architecture

Single-cell genomics relies on whole-genome amplification (WGA) to obtain sufficient DNA for sequencing (Fig. 1a). The standard workflow includes single-cell isolation, DNA extraction, WGA, long-read sequencing (e.g., ONT), base calling, and alignment to the reference genome. During amplification, template-switching events introduce artificial chimeric reads, resulting in alignment files that contain a mixture of authentic and artifactual sequences. In downstream analysis, these artifacts can mimic SV and confound variant detection. To address this challenge, we developed ChimeraLM, a Genomic Language Model (GLM) designed to integrate directly into this analysis pipeline and distinguish biological reads from amplification-induced artifacts.

ChimeraLM functions as a pre-processing filter, operating after read alignment but before SV detection. 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 decision enables the retention of authentic genomic sequences while removing amplification artifacts prior to variant calling. The resulting high-confidence biological reads are then passed to conventional SV detection algorithms for accurate identification of genomic rearrangements.

A high-confidence labeled dataset was required for supervised training of the model (Fig. 1b). 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 PacBio; see Methods). WGA reads whose chimeric structures were absent from all three bulk datasets were labeled artificial. Conversely, WGA reads with structures validated in one or more bulk datasets were labeled biological.

Application of this labeling strategy to the PC3 WGA data (Extended Table 1) quantified the read distribution across these categories (Extended Data Fig. 1). We identified 12,670,396 chimeric reads with zero matches in the bulk reference, which

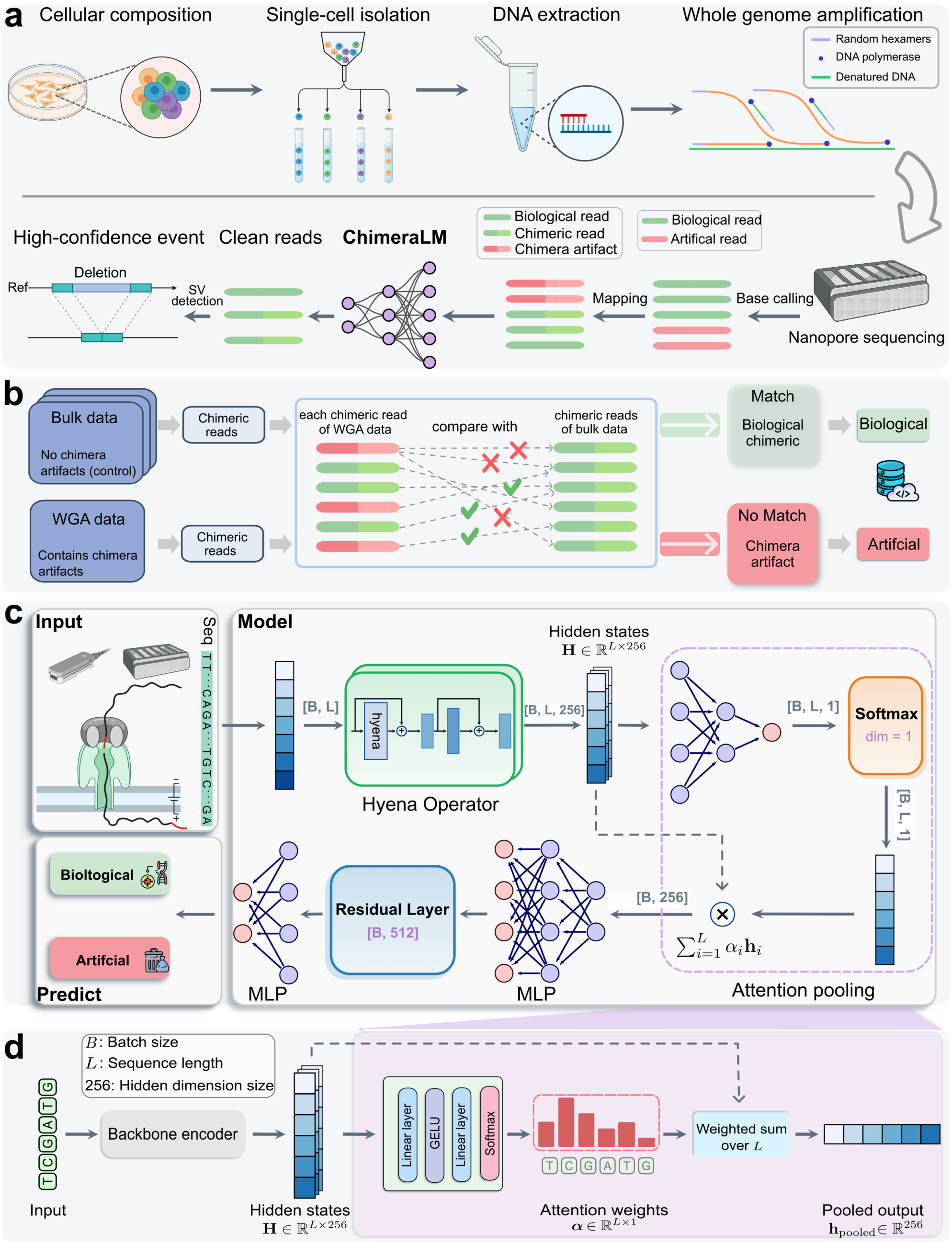


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA for genome amplification. WGA generates chimeric artifacts (red) through template 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 $\mathbf{H} \in \mathbb{R}^{L\times256}$ through a backbone 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 backbone encoder (HyenaDNA) transforms input sequences into hidden state $\mathbf{H} \in \mathbb{R}^{L\times256}$. Attention weights $\boldsymbol{\alpha} \in \mathbb{R}^{L\times1}$ 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.

were classified as artificial. Conversely, we identified a total of 293,180 reads validated as biological. This biological set was composed of reads matching one (Match 1: 101,094 reads), two (Match 2: 190,309 reads), or all three (Match 3: 1,777 reads) of the bulk reference datasets. To construct a balanced training dataset, we retained all 293,180 biological reads (combining Match 1, 2, and 3) and subsampled an equal number (293,180) of artificial reads from the no-match category. This set was augmented with 178,748 chimeric reads subsampled from the bulk datasets as positive controls. The final dataset of 765,108 labeled reads was partitioned into training (70%), validation (20%), and internal test (10%) sets using stratified splitting.

The architecture of ChimeraLM (Fig. 1c) was specifically designed to learn from this dataset by operating directly on raw DNA sequences, bypassing conventional, feature-based classifiers. This design must address three primary technical challenges: (1) efficiently processing variable-length sequences of many kilobases, and (2) simultaneously maintaining single-nucleotide resolution to detect the precise, abrupt compositional changes that define chimeric junctions, and (3) aggregating variable-length sequence representations into a consistent classification output.

ChimeraLM first addresses the need for high resolution by tokenizing input sequences at the single-nucleotide level. This base-pair precision is required to preserve the complete sequence information necessary for detecting chimeric junctions—the breakpoints where disparate genomic regions are artificially fused and which often exhibit abrupt compositional changes.

The architecture's core employs Hyena operators [**?**], selected specifically to overcome the challenge of processing long DNA sequences. Traditional attention mechanisms scale quadratically with sequence length, making them computationally prohibitive for long-read data. Hyena operators, by contrast, achieve subquadratic scaling, enabling ChimeraLM to analyze full-length reads without fragmentation and thus preserve the structural context around chimeric junctions. To leverage existing genomic knowledge, we initialized the model with weights from HyenaDNA [**?**], a genomic foundation model pre-trained on diverse DNA sequences.

Finally, to produce a classification, the model employs an attention pooling mechanism to aggregate information across the entire variable-length read (Fig. 1d). This module computes learned, position-specific weights to identify which nucleotides—such as those at the junction boundary—are most informative for the classification decision. This weighted aggregation produces a fixed-dimensional representation, which is then processed through MLP components with residual connections. A final softmax layer outputs the probability scores for the biological versus artificial classes (see Methods). This end-to-end architecture enables ChimeraLM to learn directly from raw sequence data, discovering complex patterns that may not be apparent through rule-based algorithms.

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

We first evaluated ChimeraLM's classification accuracy on the held-out test set (derived from the PromethION training data), which comprised reads with known biological or artificial status (Fig. 2a). The model achieved an F1 score of 0.81, reflecting

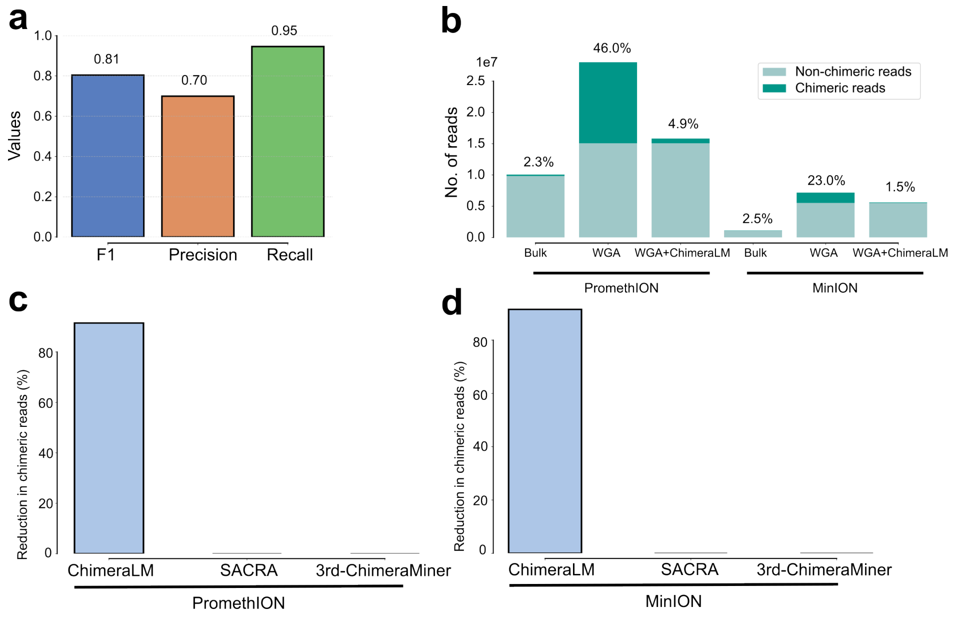


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves high recall (0.95) in identifying chimera artifacts while maintaining acceptable precision (0.70), yielding an F1 score of 0.81 for binary classification 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.

balanced sensitivity and specificity in artifact detection. A recall of 0.95 indicates that 95% of true chimeric reads were correctly identified—critical for minimizing downstream false-positive structural variant calls—while a precision of 0.70 shows that the majority of reads flagged as chimeric were true artifacts. These results establish the model's reliability for identifying amplification-induced artifacts in long-read data.

We next assessed its practical effectiveness on the full PC3 WGA datasets, comparing performance on the PromethION and MinION platforms (Fig. 2b). Bulk sequencing established a low baseline chimeric read rate (2.3% for PromethION; 2.5% for MinION). WGA dramatically increased this artifact load to 46.0% (PromethION) and 23.0% (MinION). After ChimeraLM filtering, chimeric content dropped to 4.9% on PromethION and 1.5% on MinION—representing 10- to 15-fold reductions—while retaining 15.8 million and 5.6 million biological reads. This restoration to near-bulk quality demonstrates that ChimeraLM effectively separates genuine genomic reads from WGA-induced artifacts.

We then benchmarked ChimeraLM against existing computational tools for detecting amplification-induced chimeras, SACRA [**?**] and 3rd-ChimeraMiner [**?**] (Fig. 2c,d). When applied to the same PromethION and MinION WGA data, ChimeraLM achieved an approximately 90% reduction in chimeric reads on both platforms. In stark contrast, neither SACRA nor 3rd-ChimeraMiner showed any detectable reduction in chimeric content (0% reduction).

Together, these results demonstrate a robust and generalizable performance. The strong filtering on the MinION dataset (Fig. 2b) is particularly noteworthy, as this dataset served as a completely independent test set; the model was trained exclusively on PromethION data. This cross-platform generalization, combined with the high recall on the internal test set (Fig. 2a) and the clear superiority over existing tools (Fig. 2c,d), indicates that ChimeraLM learns fundamental, generalizable sequence features of WGA-induced artifacts rather than platform-specific signatures.

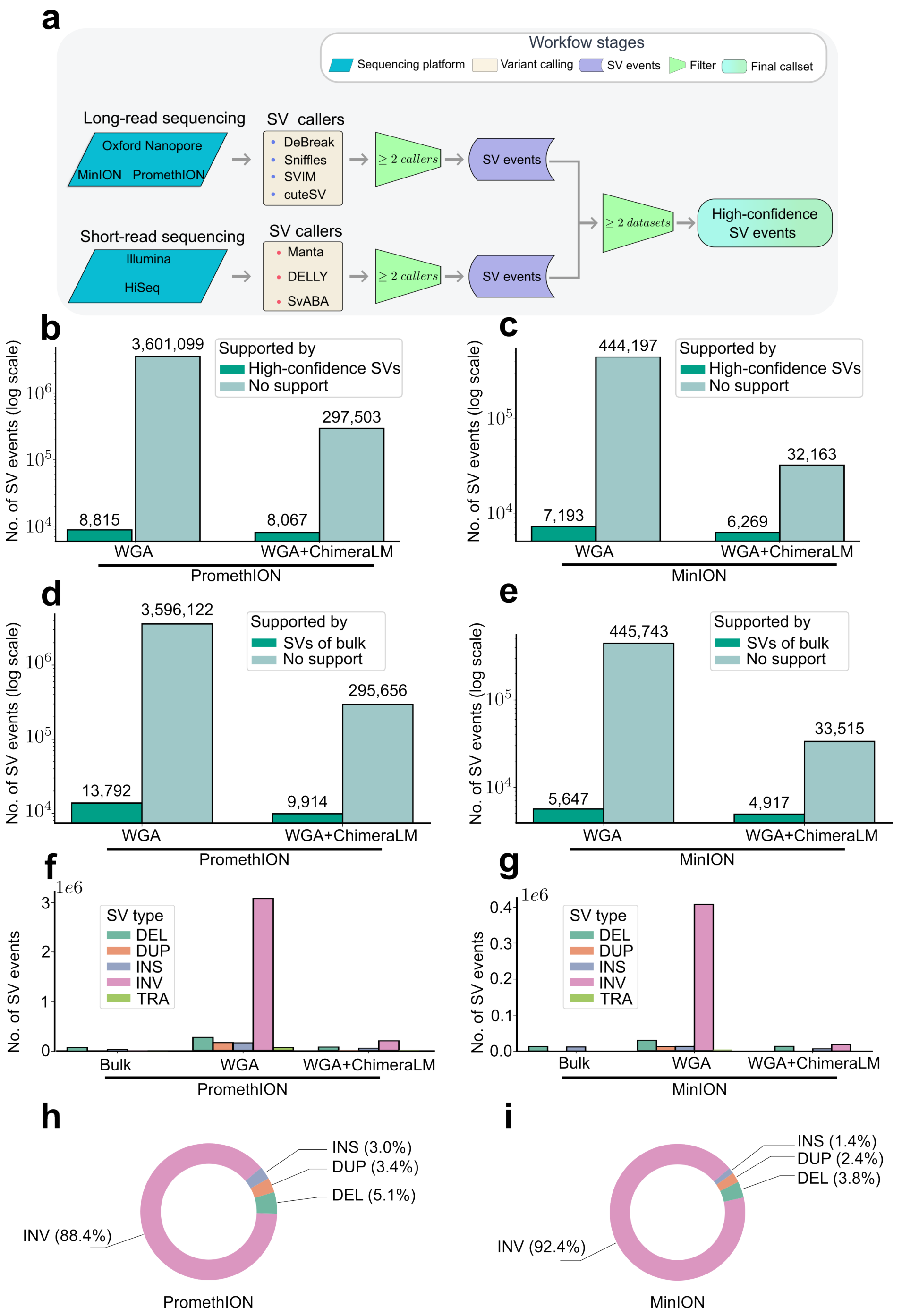


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of high-confidence SV reference dataset. PC3 bulk DNA was sequenced on multiple platforms (ONT PromethION and MinION, Illumina HiSeq) and analyzed with multiple SV calling algorithms. SV events detected by ≥2 callers on 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: DEL (green), duplication (DUP) (orange), 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

Accurate SV detection is essential for understanding genomic diversity and disease mechanisms in single cells. However, WGA-induced chimeric artifacts can be misidentified as genuine SVs, leading to incorrect biological conclusions. To quantify ChimeraLM's impact on SV calling accuracy, we compared variant calls from unfiltered WGA data and ChimeraLM-filtered data against two independent reference standards (Fig. 3).

We first established a high-confidence gold standard SV dataset by integrating results from bulk PC3 DNA sequenced on multiple platforms (ONT PromethION, ONT MinION, and Illumina HiSeq) and analyzed with multiple SV callers (Fig. 3a; Extended 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, ensuring high specificity across technologies.

Comparison against this gold standard revealed that unfiltered WGA data contained extensive false-positive SVs (Fig. 3b,c). On PromethION, raw WGA data produced 3.6 million SV calls, of which only 8,815 (0.24%) matched gold standard events—indicating that over 99% were artifacts. After ChimeraLM filtering, total calls dropped to 305,570 while retaining 8,067 true events, raising the validation rate to 2.64% (11-fold improvement) and preserving 91.5% of true variants. MinION data showed similar results, with calls reduced from 451,390 to 38,432 and the validation rate increasing from 1.59% to 16.3% (10-fold improvement) while retaining 87.2% of true variants. These results highlight ChimeraLM's ability to remove spurious SV calls while maintaining biological sensitivity.

To complement this stringent validation, we next performed platform-matched bulk 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. Under this benchmark, 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. The consistent improvements across independent datasets demonstrate that ChimeraLM effectively suppresses WGA-induced artifacts without sacrificing detection of genuine SVs.

Together, these analyses demonstrate that ChimeraLM reduces false-positive SV calls by 8–16 fold while preserving 72–92% of true variants, resulting in a substantial enhancement of the signal-to-noise ratio in single-cell SV discovery. By restoring near-bulk specificity and maintaining robust sensitivity, ChimeraLM enables more accurate and interpretable downstream genomic analyses.

ChimeraLM restores unbiased SV-type distributions and characterizes artifact composition

Amplification artifacts can distort the apparent spectrum of SVs, often inflating specific SV types. To evaluate whether ChimeraLM effectively corrects such distortions, we compared SV type distributions across bulk, unfiltered WGA, and ChimeraLM-filtered datasets (Fig. 3f,g). Bulk sequencing showed relatively balanced proportions of DELs, DUPs, INSs, INVs, and TRAs. In contrast, unfiltered WGA data exhibited a dramatic overrepresentation of INVs on both PromethION and MinION platforms, consistent with pervasive amplification artifacts. After ChimeraLM filtering, these distributions were largely restored toward bulk-like profiles: excessive INVs were markedly reduced while other SV categories remained stable. This shift reflects selective removal of artifact-supported INVs rather than indiscriminate loss of genuine inversion signals, demonstrating high specificity in distinguishing chimeric from biological reads.

To investigate the basis of this normalization, we analyzed SV calls supported exclusively by reads classified as chimeric by ChimeraLM (Fig. 3h,i). These artifact-supported events were overwhelmingly dominated by INVs, comprising 88.4% on PromethION and 92.4% on MinION. This pattern is consistent with template-switching junctions that produce 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%) were also observed, demonstrating that WGA-induced chimeras can mimic diverse SV categories rather than only INVs.

This characterization has important implications for single-cell genomics. Although INVs are the predominant artifact type, the coexistence of DELs, DUPs, and INSs among chimeric events indicates that comprehensive filtering—rather than inversion-specific correction—is essential for accurate SV detection. Without ChimeraLM filtering, single-cell SV analyses would be confounded not only by false-positive INVs but also by other artifact-associated variants [**? ?**]. By restoring biologically representative SV type distributions, ChimeraLM enables robust and interpretable characterization of structural variation in single cells without distortion from WGA-induced artifacts.

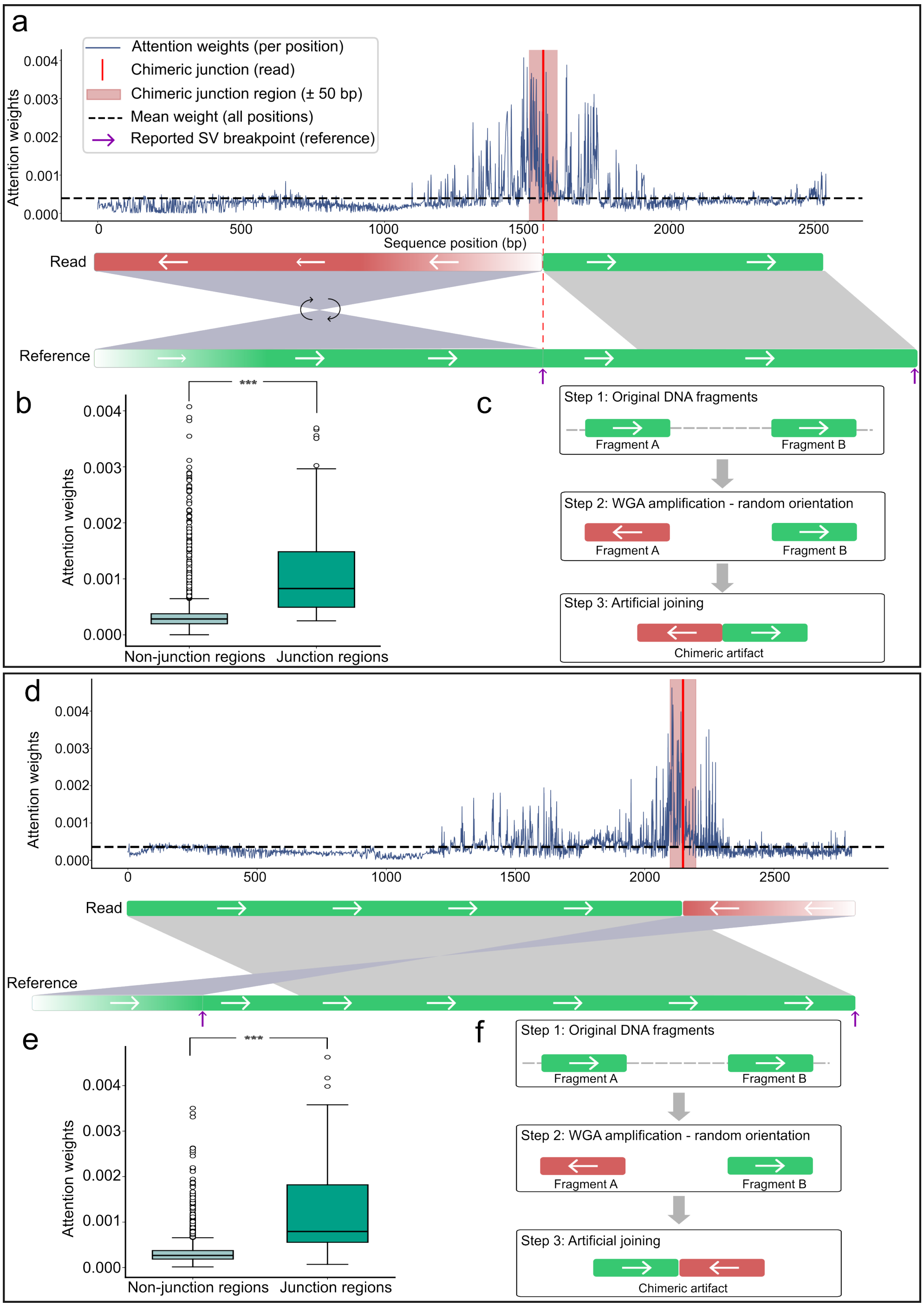


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions. (a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per sequence position (blue line) and mean attention (dashed line). Red vertical lines mark chimeric junction positions, 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.

ChimeraLM provides interpretable classification through attention visualization

We next investigated whether ChimeraLM's attention mechanism highlights biologically meaningful regions within sequencing reads (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 artificially joins DNA fragments from distinct genomic loci (Fig. 4a,d). These peaks coincided precisely with alignment breakpoints characterized by orientation changes between adjacent read segments—the defining signature of WGA-induced chimeric artifacts.

Quantitative analysis confirmed that attention weights within junction regions (±50 bp) were significantly higher than those in non-junction regions (Wilcoxon rank-sum test, $p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$) (Fig. 4b,e). Such localization indicates that ChimeraLM learns mechanistically relevant features associated with artificial junction formation rather than relying on spurious correlations.

Schematic reconstruction of the amplification process further supports this interpretation (Fig. 4c,f). During WGA, DNA fragments from distant genomic loci may undergo random strand orientation changes before being joined by template switching. This process produces artificial junctions with discordant orientations—forward-to-reverse or reverse-to-forward—that generate inversion-like alignment signatures and are effectively recognized by the model's attention peaks.

Together, these analyses demonstrate that ChimeraLM's attention mechanism can localize chimeric junctions at single-base resolution and capture the underlying orientation discontinuities that define WGA-induced artifacts.

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 87–92% of true SVs. It reduces false-positive SV calls by 8–16 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 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 contexts in long-read single-cell sequencing. Testing on multiple cell types (primary, stem, or immune cells) and WGA protocols will establish generalizability. The interpretability of attention-based models could also 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.

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

Backbone encoder

ChimeraLM employs the pre-trained HyenaDNA model [**?**] as its backbone 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 backbone 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 1×10^{-4} and weight decay 0.01. 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 $z = [z_0, z_1]$, the loss is:

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

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

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).

Extended Data Table 1 Sequencing and alignment statistics of PC3

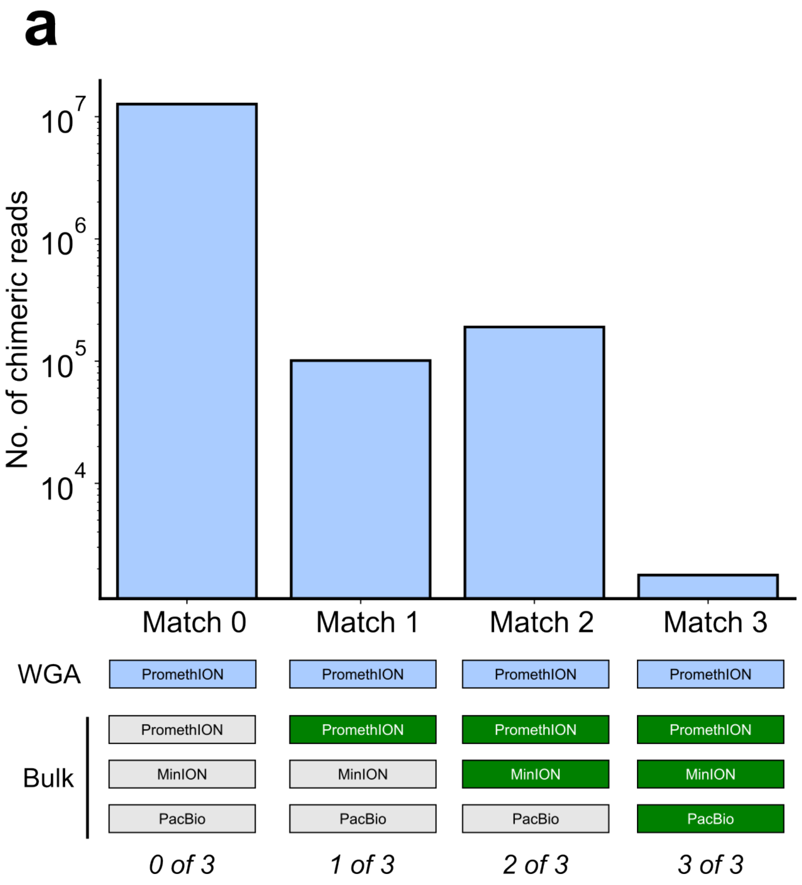
Sample	**Platform**	**Reads** ($\times 10^6$)	**Total bases** (Gb)	**Total bases aligned** (Gb)	**Fraction aligned**	**Mean length** (bp)	**Mean quality** (Q)	**Average identity** (%)
WGA	MinION	9.11	14.6	10.4	0.7	1,603	14.3	97.6
WGA	PromethION	44.69	128.2	69.2	0.5	2,869	14.5	96.1
Bulk	MinION	0.97	8.1	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

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



Extended Data Fig. 1 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). Symbols below the bars represent the number of bulk datasets with a match (0 of 3 – 3 of 3). This matching scheme forms the basis for ground-truth labeling in supervised training.

implemented the model. YL built the command-line tool and documentation. YL, QG and RY wrote the manuscript. RY supervised this work.

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. 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 15

DEL deletion 3, 9, 11
dMDA droplet-based MDA 2
DOP-PCR Degenerate Oligonucleotide-Primed PCR 2
DUP duplication 9, 11

GLM Genomic Language Model 15
GPU Graphics Processing Unit 15, 20, 22, 24

HPC High Performance Computing 22

INS insertion 3, 9, 11
INV inversion 1, 3, 9, 11

LIANTI Linear Amplification via Transposon Insertion 2, 14, 15

MALBAC Multiple Annealing and Looping-based Amplification Cycles 2, 14, 15
MDA Multiple Displacement Amplification 2, 3, 14, 15
MLP multilayer perceptron 5, 6, 19

ONT Oxford Nanopore Technologies 3, 4, 9, 10, 16, 17

PacBio Pacific Biosciences 3
PTA Primary Template-directed Amplification 2

SA Supplementary Alignment 16, 17, 20–22
SV Structural Variation 1–5, 9–15, 17, 20, 21

TRA translocation 3, 9, 11

WGA Whole Genome Amplification 1–15, 17–23