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 and development by enabling the characterization of individual cells rather than bulk populations [1–4]. This approach has proven instrumental in uncovering rare cell types [4], tracking developmental trajectories, elucidating tumor evolution through clonal architecture analysis [3], and identifying somatic mutations that drive disease progression at unprecedented resolution. By resolving cellular mosaicism and enabling lineage tracing at single-cell resolution, these studies have fundamentally transformed our understanding of development, disease, and evolution. 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 [5–7].

To overcome this limitation, WGA has become essential for single-cell genomic studies [4, 7–10]. Various WGA techniques have been developed, each with distinct amplification mechanisms and characteristic error profiles. Multiple Displacement Amplification (MDA), introduced by Dean et al. [10], 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 [11, 12]. Degenerate Oligonucleotide-Primed PCR (DOP-PCR), pioneered by Telenius et al. [13], 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 [8], while Linear Amplification via Transposon Insertion (LIANTI) uses transposon insertion to create defined amplification origins, significantly improving uniformity and reducing artifacts [7]. More recently, Primary Template-directed Amplification (PTA) [14] and droplet-based MDA (dMDA) [15, 16] 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 [4, 17–21].

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 [3]. 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 [14]. However, false-positive SVs introduced during amplification can confound these analyses, leading to misinterpretation of genomic rearrangements and their biological significance [4, 22].

Despite its critical role, WGA introduces systematic artifacts that significantly impact downstream analyses [7, 11, 12, 22, 23]. 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 [11, 22]. 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 [11]. Critically, even with WGA technological advances, chimeric artifacts remain highly prevalent in single-cell long-read sequencing data [14, 15]. Lasken and Stockwell [11] 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, insertions, inversions, and translocations—that do not exist in the original cell [20, 22], posing substantial challenges for accurate SV detection in single-cell studies. Early work by Pinard et al. [12] 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 [24, 25], DeBreak [26], SVIM [27], and cuteSV [28]. These methods typically employ read alignment analysis, split-read detection, and local assembly strategies to identify SV signatures [29]. However, distinguishing genuine biological SVs from WGA-induced chimeric artifacts remains challenging [23, 30–32].

Current computational approaches for identifying WGA-induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [23, 30]. 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 [31, 32]. 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 [33–36].

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 [36, 37]. These models treat DNA sequences analogously to natural language, learning representations that capture both local motifs and long-range dependencies [33]. 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 [33–36], 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 more accurate analysis at single-cell resolution.

Results

ChimeraLM workflow, training strategy, and architecture

We developed ChimeraLM as an integrated component of single-cell genomic analysis pipelines to address WGA-induced chimeric artifacts (Fig. 1a). The workflow begins with standard single-cell procedures: cellular isolation through sorting technologies, DNA extraction, and WGA using established protocols. Amplified material is sequenced on long-read platforms such as ONT. ChimeraLM operates at a critical position in the analysis pipeline—after read alignment but before SV detection. Following standard quality filtering and mapping to the reference genome, ChimeraLM evaluates each chimeric read and classifies it as either biological or artificial. This binary classification enables selective retention of authentic genomic sequences while filtering out amplification artifacts upstream of variant calling. Filtered biological reads then proceed to conventional SV detection algorithms for identification of genuine genomic alterations. This design allows ChimeraLM to integrate with existing pipelines without requiring substantial modifications to established protocols.

A key innovation enabling ChimeraLM is our training data construction strategy, which leverages paired WGA and bulk sequencing from the same biological sample (Fig. 1b). This approach exploits a fundamental difference: while WGA data contains both biological reads and amplification-induced chimeric artifacts, bulk sequencing from non-amplified DNA contains only genuine biological sequences. Our ground truth labeling strategy compares each chimeric read from WGA against bulk sequencing data (see Methods). Reads that match bulk data are labeled as biological, indicating they represent authentic genomic sequences. Reads that fail to match bulk sequences are labeled as artificial chimeras generated during amplification.

Application of this matching strategy to PC3 WGA data (Extended Table 1) sequenced on PromethION revealed that 12,670,396 chimeric reads showed no matches in bulk data (classified as artificial), while 101,094, 190,309, and 1,777 reads showed

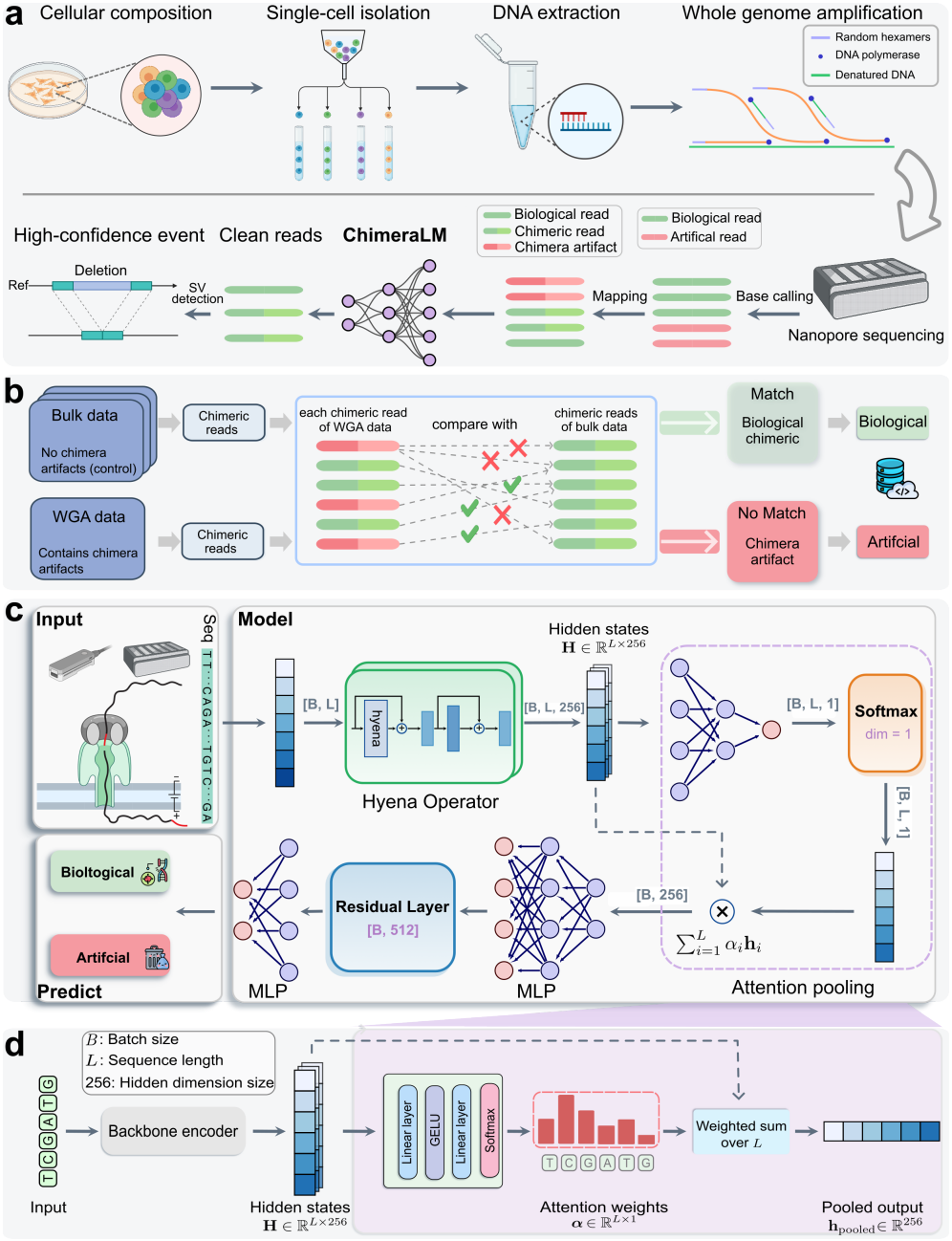


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated and sorted, followed by DNA extraction and WGA for genome amplification. WGA generates chimeric artifacts (red) through 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 neural network 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 backbone encoder (HyenaDNA [35]). 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 $\mathbf{H} \in \mathbb{R}^{L\times 256}$. 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.

1, 2, and 3 matches, respectively (classified as biological) (Extended Data Fig. 1). To create a balanced training dataset, we subsampled 293,180 reads from the no-match category as artificial chimeras and retained all reads with matches (293,180 total) as biological. Additionally, we subsampled 178,748 chimeric reads from bulk sequencing as biological controls, yielding a final dataset of 765,108 labeled reads partitioned into training (70%), validation (20%), and test (10%) sets using stratified splitting.

ChimeraLM's neural network architecture analyzes DNA sequences at single-nucleotide resolution to distinguish biological reads from chimeric artifacts (Fig. 1c,d). The architecture comprises three main components. First, input sequences are tokenized at single-nucleotide level, representing each base individually to preserve complete sequence information necessary for detecting chimeric junctions—the breakpoints where disparate genomic regions are artificially fused. These junctions often exhibit abrupt compositional changes requiring base-pair precision.

Second, the model employs Hyena operators [38], which efficiently process long DNA sequences. Traditional attention mechanisms scale quadratically with sequence length, making them computationally prohibitive for long reads. Hyena operators achieve subquadratic scaling, enabling ChimeraLM to analyze full-length reads without fragmentation, preserving structural context around chimeric junctions. We initialized the model with weights from HyenaDNA [35], a genomic foundation model pre-trained on diverse DNA sequences, allowing ChimeraLM to leverage general sequence patterns before fine-tuning.

Third, an attention pooling mechanism aggregates information across the entire read while learning which positions are most informative for classification (Fig. 1d). The attention module computes position-specific weights indicating each nucleotide's relevance to classification. This weighted aggregation produces a fixed-dimensional representation from variable-length inputs, which is then processed through MLP components with residual connections. The final softmax layer outputs probability scores for biological versus artificial classification (see Methods). This end-to-end architecture enables ChimeraLM to learn directly from raw sequence data without manual feature engineering, discovering complex patterns that may not be apparent through traditional bioinformatics approaches.

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

We first evaluated ChimeraLM's classification accuracy on held-out test data comprising reads with known biological or chimeric status (Fig. 2a). The model achieved an F1 score of 0.81, balancing sensitivity and specificity in artifact detection. ChimeraLM demonstrated high recall (0.95), successfully identifying 95% of true chimeric artifacts, which is critical for preventing false-positive structural variant calls. The model's precision of 0.70 indicates that 70% of reads flagged as chimeric were true artifacts.

This precision-recall tradeoff reflects a deliberate design choice prioritizing comprehensive artifact removal over perfect specificity. Retaining chimeric reads leads to false SV calls that misrepresent cellular genotypes, whereas removing some biological reads reduces sequencing depth but does not introduce false biological conclusions. For typical single-cell WGA samples with 20-30× coverage, the loss of 30% of reads in

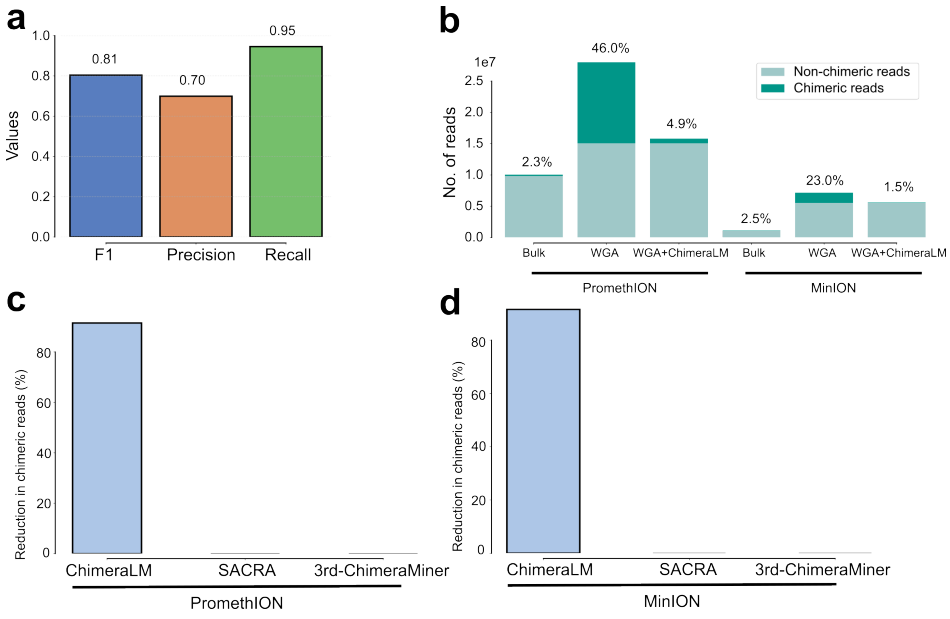


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves high recall (0.95) in identifying chimeric reads while maintaining acceptable precision (0.70), yielding an F1 score of 0.81 for binary classification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

chimera-dense regions maintains sufficient depth (14-21×) for reliable variant calling, while removal of 95% of chimeric artifacts substantially improves variant detection accuracy.

To assess practical effectiveness, we applied ChimeraLM to PC3 WGA data sequenced on PromethION and MinION platforms (Fig. 2b). ChimeraLM was trained using a subset of PromethION data, where chimeric artifacts were identified by comparison with bulk sequencing. Importantly, not all chimeric reads from PromethION were used during training, allowing evaluation on both seen and unseen examples, while MinION data was completely independent.

Bulk sequencing established baseline chimeric rates of 2.3% (PromethION) and 2.5% (MinION), representing low background artifacts in non-amplified samples. WGA dramatically increased contamination to 46.0% (PromethION) and 23.0% (MinION). When applied to the full PromethION dataset, ChimeraLM reduced chimeric content from 46.0% to 4.9%, retaining 15.8 million biological reads from 28.0 million total—a 10-fold reduction in artifacts. On the independent MinION platform,

ChimeraLM reduced chimeric reads from 23.0% to 1.5%, approaching bulk sequencing quality while preserving 5.6 million reads from 7.2 million total—a 15-fold reduction.

The strong performance on both unseen PromethION reads and the completely independent MinION platform demonstrates that ChimeraLM learns generalizable sequence features distinguishing biological reads from chimeric artifacts, rather than memorizing training examples or platform-specific signatures. This cross-platform generalization enables users to apply ChimeraLM to different datasets and sequencing platforms without retraining, making the method practical for routine single-cell genomic analyses.

We compared ChimeraLM to existing computational methods for detecting amplification-induced chimeric sequences: SACRA [30] and 3rd-ChimeraMiner [23] (Fig. 2c,d). Both tools were applied to PC3 WGA data sequenced on PromethION and MinION platforms using default parameters.

ChimeraLM achieved approximately 90% reduction in chimeric reads on both platforms, whereas SACRA and 3rd-ChimeraMiner showed no detectable reduction in chimeric content (0% reduction compared to unprocessed data). This substantial performance difference demonstrates ChimeraLM's effectiveness for detecting WGA-induced chimeric artifacts in long-read single-cell sequencing data. By training directly on WGA data, ChimeraLM learns sequence-level patterns specific to how WGA chimeras manifest in long-read sequencing, providing a practical solution for single-cell genomic quality control.

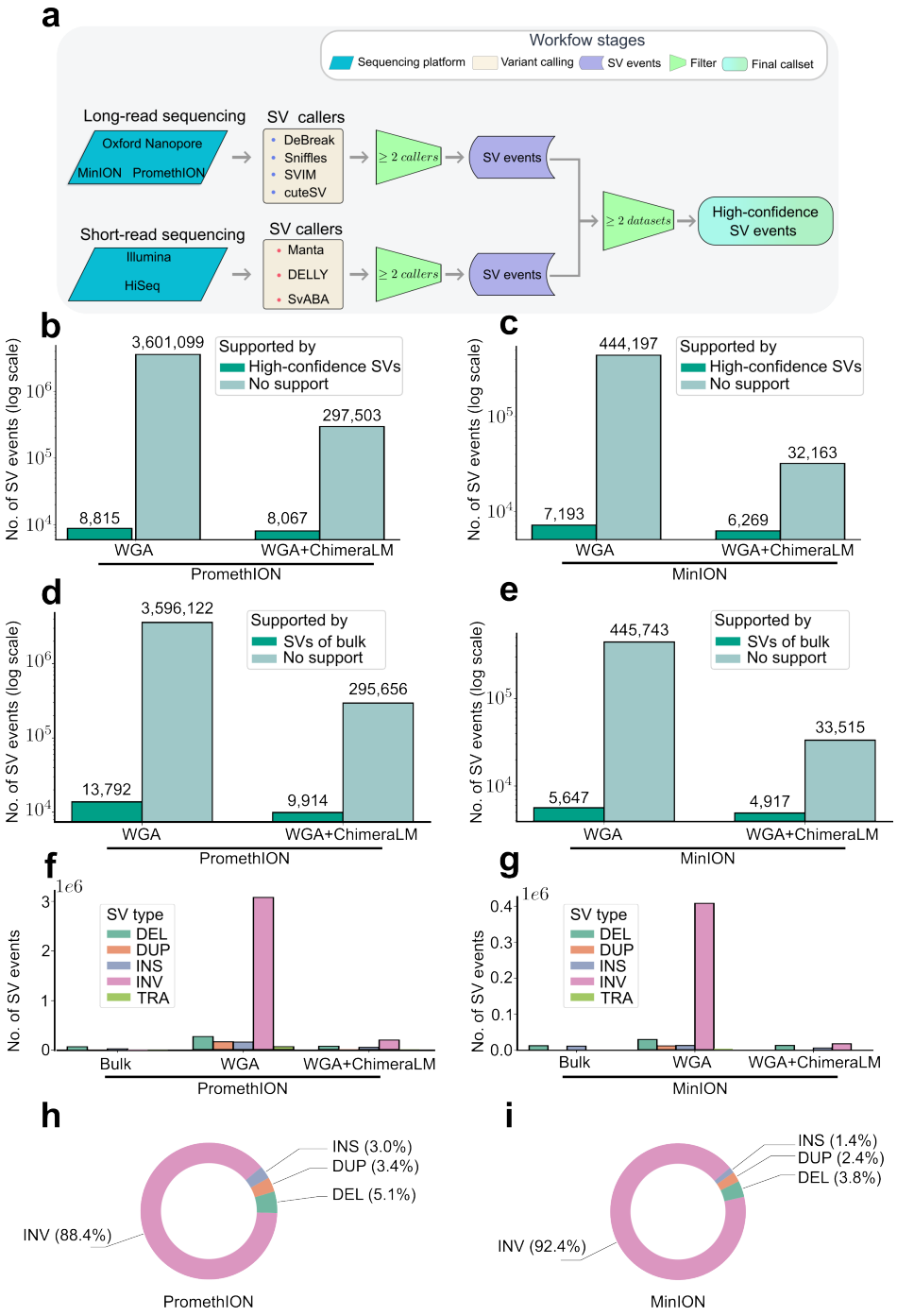


Fig. 3 ChimeraLM filtering improves structural variant detection accuracy. (a) Construction of high-confidence SV reference dataset. PC3 bulk DNA was sequenced on multiple platforms (ONT PromethION and MinION, Illumina HiSeq) and analyzed with multiple SV calling algorithms. SV events detected by ≥2 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 filtering 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 bulk-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 translocation (TRA) (light green) for PromethION (f) and MinION (g). Unfiltered WGA data shows elevated counts across all types, particularly inversions and translocations, 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 filtering.

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 downstream SV detection accuracy, we compared variant calls from unfiltered WGA data versus ChimeraLM-filtered data against two independent reference standards.

We first constructed a high-confidence gold standard SV dataset by integrating multiple sequencing platforms and detection algorithms (Fig. 3a). PC3 bulk DNA was sequenced using long-read (ONT PromethION and MinION) and short-read (Illumina HiSeq) technologies (Extended Table 1). SVs detected by $\geq$2 calling algorithms on the same platform and supported by both long-read and short-read data were designated as gold standard events. This multi-platform consensus approach ensures high specificity, as true SVs should be detectable across different sequencing technologies.

When comparing WGA samples against the gold standard, unfiltered data showed extensive false-positive SV calls (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—meaning 99.76% of calls were likely artifacts. ChimeraLM filtering reduced total calls to 305,570 while retaining 8,067 gold standard events, increasing the validation rate to 2.64% (11-fold improvement) and preserving 91.5% of true variants. MinION data showed similar results: WGA produced 451,390 calls with 1.59% validation rate, while ChimeraLM-filtered data yielded 38,432 calls with a 16.3% validation rate (10-fold improvement) and 87.2% true variant retention.

To complement the stringent gold standard validation, we also compared SV calls against long-read bulk sequencing from the same platform (Fig. 3d,e). This platform-matched validation captures true SVs that may be detectable specifically in long-read data but missed by short-read sequencing, providing a more inclusive reference for evaluating recall. On PromethION bulk validation, ChimeraLM filtering increased the validation rate from 0.38% to 3.24% (8.5-fold improvement) while retaining 71.9% of bulk-supported events. MinION bulk validation showed similar improvements, with validation rates increasing from 1.25% to 12.79% (10-fold improvement) and 87.1% retention of bulk-supported events.

These results have important practical implications. The 8-16 fold reduction in false-positive rate means researchers can focus on biologically relevant SVs without manually filtering thousands of artificial calls. The high retention of true variants (72-92% depending on validation stringency) ensures that ChimeraLM filtering does not compromise detection sensitivity for genuine genomic alterations. Together, these metrics demonstrate that ChimeraLM substantially improves the signal-to-noise ratio in single-cell SV detection, making downstream biological interpretation more reliable and efficient.

ChimeraLM normalizes SV type distributions and reveals artifact composition

We compared SV type distributions across bulk, unfiltered WGA, and ChimeraLM-filtered samples to assess whether artifact removal normalizes SV profiles (Fig. 3f,g). Bulk sequencing exhibited relatively balanced distributions across DELs, DUPs, INSs, INVs, and TRAs. In contrast, unfiltered WGA data showed dramatically skewed distributions dominated by INVs on both PromethION and MinION platforms. ChimeraLM filtering substantially normalized these distributions toward bulk profiles, dramatically reducing the excess inversions while maintaining other SV types at levels consistent with bulk sequencing. This normalization occurred through selective removal of artifact-supported inversions rather than indiscriminate elimination of all inversion calls.

To understand this normalization pattern, we leveraged ChimeraLM's read-level classification to systematically characterize which SV calls are associated with artificial versus biological reads—an analysis not possible without such a classifier (Fig. 3h,i). This deep learning approach reveals new insights into WGA artifact composition. Analysis of SVs supported exclusively by ChimeraLM-identified chimeric reads revealed that INVs strongly dominate (88.4% on PromethION, 92.4% on MinION), consistent with how chimeric junctions create alignment signatures resembling genuine INVs. Importantly, DELs (5.1% PromethION, 3.8% MinION), DUPs (3.4%, 2.4%), and INSs (3.0%, 1.4%) also appear in the artifact-associated landscape, suggesting that WGA-induced chimeras can manifest as multiple SV types, not only INVs. This finding provides a new perspective: WGA-induced artifacts are not limited to inversions but can manifest across multiple SV types, demonstrating how deep learning classification advances understanding of amplification artifact mechanisms.

This characterization has practical implications. While INVs are by far the most strongly associated with chimeric artifacts, the presence of other SV types in the artifact landscape indicates that comprehensive filtering—rather than INV-specific approaches—is necessary for accurate single-cell SV analysis. Without ChimeraLM filtering, single-cell genomic studies would be compromised not only by false-positive INVs but potentially by other artifact-associated SV types [31, 32]. The restoration of biologically representative SV type distributions enables accurate characterization of SV in single cells without confounding effects of amplification artifacts.

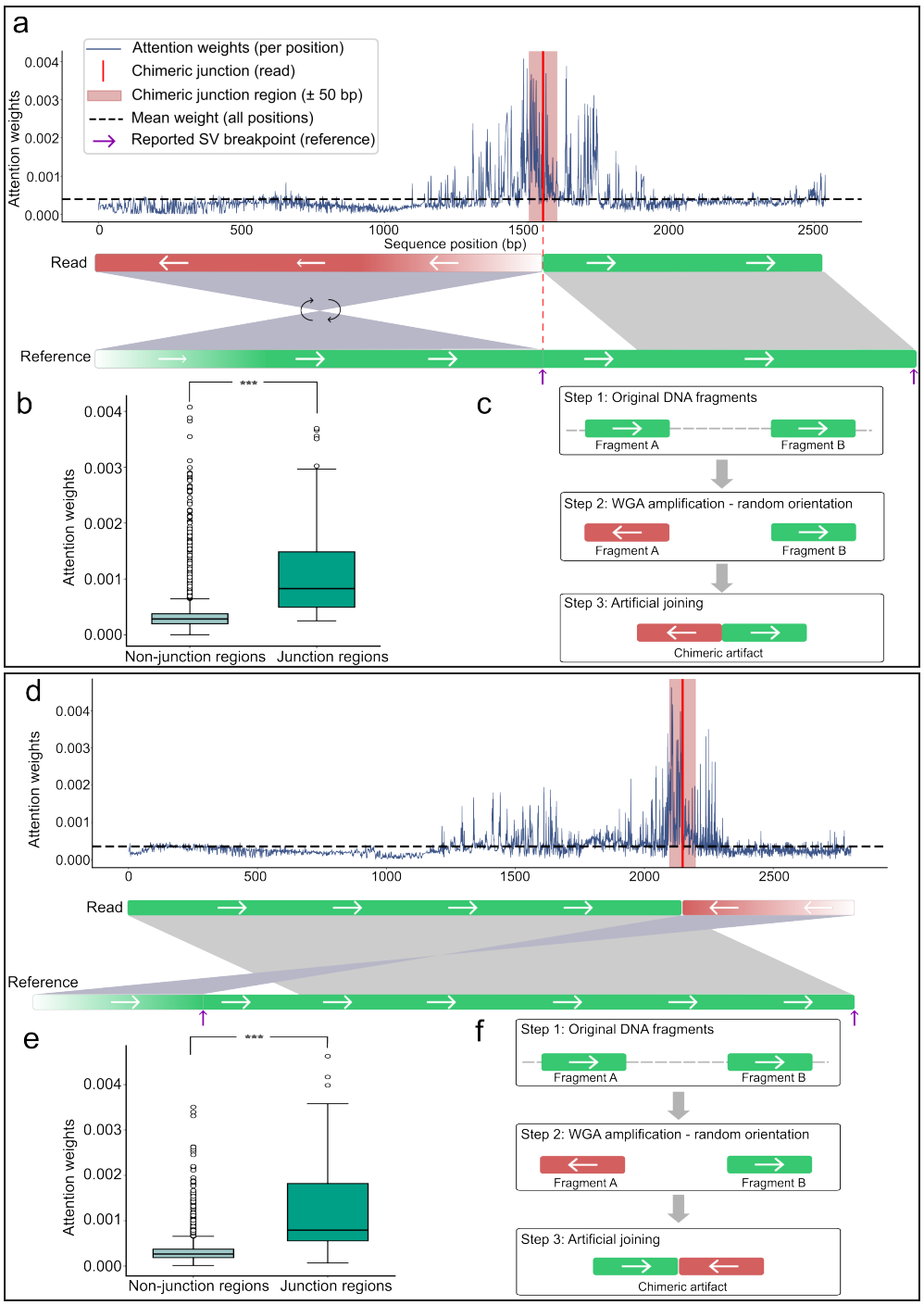


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions in representative examples. (a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per sequence position (blue line) and mean attention (dashed line). Red vertical lines mark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

A key advantage of ChimeraLM's architecture is interpretability—the ability to visualize which sequence regions drive classification decisions. Unlike black-box models, ChimeraLM's attention pooling mechanism assigns explicit weights to each nucleotide position, revealing what the model considers important. This interpretability provides insights into both the biological mechanisms of chimera formation and the model's decision-making process.

We examined attention weight distributions across chimeric reads to determine whether ChimeraLM focuses on mechanistically relevant features (Fig. 4a,d). In representative examples, attention profiles showed predominantly low baseline weights across most of the reads, with pronounced peaks at chimeric junction regions—the breakpoints where template switching artificially joins DNA fragments from different genomic loci. At these junctions, reads exhibit characteristic discordant alignment patterns: one segment aligns in reverse orientation while the adjacent segment aligns forward to a distant genomic location, creating the structural signature of WGA-induced chimeric artifacts.

Quantitative analysis confirmed that attention weights within the junction region ($\pm$50 bp) were significantly elevated compared to background regions ($p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$, Wilcoxon rank-sum test) (Fig. 4b,e). This localization pattern demonstrates that ChimeraLM learns mechanistically relevant features—specifically, the orientation discontinuities created when DNA fragments are artificially joined during amplification (Fig. 4c,f). The model's ability to focus on junction breakpoints validates that it captures the underlying mechanism of chimera formation rather than relying on spurious correlations.

Not all chimeric reads exhibit such clearly localized attention patterns. Some show more diffuse attention distributions, suggesting ChimeraLM integrates multiple complementary features for classification—including junction signatures when prominent, but also compositional biases, k-mer patterns, or context-dependent features. This feature diversity likely reflects the heterogeneous nature of chimeric artifacts, which vary in junction structure, fragment length, and local sequence context. Rather than limiting interpretability, this flexibility demonstrates the model's sophistication in handling diverse artifact types.

The interpretability provided by attention weights offers several practical advantages. First, it builds user trust by showing that classification decisions are based on biologically meaningful features rather than technical artifacts or dataset biases. Second, attention visualization enables manual inspection of individual predictions, particularly for high-confidence classifications where junction localization provides additional validation. ChimeraLM thus serves dual purposes: removing artifacts to improve data quality and providing analytical insights into amplification mechanisms.

Discussion

WGA has enabled genomic analysis from single cells but introduces chimeric artifacts that compromise structural variant 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 approach—removing problematic sequences at the read level rather than attempting to correct errors after variant calling—provides a practical solution for single-cell genomics laboratories.

Our results demonstrate several key advantages of ChimeraLM for long-read single-cell sequencing data. First, the method achieves approximately 90% reduction in chimeric reads across different nanopore sequencing platforms while retaining 87-92% of true SVs. Second, ChimeraLM reduces false-positive SV calls by 8-16 fold, enabling researchers to focus on biologically relevant variants without manual filtering of thousands of artificial calls. Third, the approach works across PromethION and MinION platforms without platform-specific retraining, indicating that the model learns generalizable sequence features of WGA-induced chimeras.

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 approaches rely on predefined criteria such as mapping quality or read depth [23, 30], which may not effectively distinguish chimeric artifacts from biological sequences. By learning directly from data, ChimeraLM discovers subtle compositional and structural features that distinguish authentic genomic sequences from amplification artifacts. The attention weight analysis provides evidence that the model can identify mechanistically relevant features such as orientation discontinuities at chimeric junctions, though the heterogeneity in attention patterns across reads suggests the model uses multiple complementary features for classification.

The improved reliability of SV detection has practical implications for single-cell genomics applications. Studies of chromosomal instability, clonal evolution, and structural variant burden in individual cells have been limited by high false-positive rates in WGA data [31, 32]. ChimeraLM enables more confident identification of genuine structural variants, supporting applications in cancer genomics, developmental biology, and aging research where single-cell resolution is essential for understanding cellular heterogeneity.

Several limitations warrant consideration. ChimeraLM was trained and validated on PC3 cell line data using MDA-based WGA and nanopore sequencing. Performance on other cell types and WGA chemistries (MALBAC, LIANTI and other MDA variants) remains to be systematically evaluated. Amplification biases may vary across genomic backgrounds with different chromatin states or DNA accessibility, potentially affecting model generalization. The requirement for bulk sequencing data to generate training labels limits immediate applicability to samples where bulk material is unavailable, though transfer learning from existing trained models may address this constraint.

ChimeraLM currently processes reads independently without considering genomic context or supporting read depth. Integrating additional features such as local coverage, mate-pair information, or phasing data could improve classification accuracy. The model requires Graphics Processing Unit (GPU) resources for efficient processing of large datasets (millions of reads), though Central Processing Unit (CPU) inference

remains feasible for smaller studies. Runtime optimization and model compression could improve accessibility for laboratories with limited computational infrastructure.

Future work should prioritize validation across diverse biological contexts in long-read single-cell sequencing. Testing on multiple cell types (primary cells, stem cells, immune cells) and WGA protocols will establish generalizability. Integration with additional long-read sequencing modalities (linked-reads, strand-seq) could provide complementary information for chimera detection. The interpretability of attention-based models could be leveraged to provide insights into WGA artifact formation mechanisms. Systematic analysis of attention patterns across thousands of chimeric reads may reveal common sequence motifs, structural features, or genomic contexts associated with template switching, informing development of improved amplification protocols.

More broadly, ChimeraLM demonstrates the utility of genomic language models for quality control applications [35]. The architectural innovations incorporated in ChimeraLM, particularly the use of Hyena operators for efficient long-range modeling [38], may have applications beyond chimeric detection. Similar deep learning approaches could address other data quality challenges in long-read single-cell genomics, including contamination detection, adapter artifact identification, or systematic error correction. As foundation models for biological sequences continue to advance, quality control and data preprocessing may emerge as important application domains alongside traditional prediction tasks.

ChimeraLM provides a practical and effective tool for improving long-read single-cell genomic data quality. By removing WGA-induced chimeric artifacts at the read level, the method enables more reliable SV detection and supports biological applications that require accurate characterization of genomic variation at single-cell resolution.

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` [39]. Reads with mean quality < 10 or length < 500 bp were removed. Residual adapters and concatemers were trimmed using Cutadapt v4.0 [40] in two-pass error-tolerant mode. Cleaned reads were aligned to the GRCh38.p13 reference genome using minimap2 v2.26 (`map-ont` preset) [41]. Resulting BAM files were sorted and indexed with SAMtools v1.16 [42]. Read length and mapping statistics were calculated using NanoPlot v1.46.1 [43]. 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) contained the SA tag, (2) were not unmapped, (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

Training data generation

We generated training data from PC3 cells using WGA sequencing on the PromethION P2 platform (ONT) and three independent bulk sequencing datasets: bulk sequenced

on PromethION P2, bulk sequenced on MinION Mk1c (ONT), and bulk sequenced on PacBio. WGA data sequenced on MinION Mk1c was reserved as a completely independent test set. Bulk sequencing from non-amplified genomic DNA serves as reference data containing only genuine biological sequences, while WGA sequencing from amplified DNA contains both authentic genomic sequences and amplification-induced chimeric artifacts.

Ground truth labeling

We first labeled chimeric reads from WGA PromethION P2 data by comparing them against chimeric reads from all three bulk datasets. For each WGA chimeric read, we extracted all alignment segments (defined by start-end positions on the reference genome) and compared them against alignment segments from bulk chimeric reads. A WGA read was labeled as biological if all its alignment segments matched corresponding segments from at least one bulk chimeric read within a 1 kb position threshold, indicating the chimeric structure exists in non-amplified DNA. WGA reads whose alignment segment patterns did not match any bulk chimeric reads across all three datasets were labeled as artificial chimeras generated during amplification. To augment the biological class, we subsampled chimeric reads from all three bulk datasets and labeled them as biological, as chimeric reads in non-amplified bulk DNA represent genuine biological events (e.g., true SVs). The final training dataset combined the labeled WGA PromethION P2 reads with the subsampled bulk chimeric reads.

Dataset partitioning and cross-platform validation

The combined labeled dataset was partitioned into training (70%), validation (20%), and test (10%) sets using stratified random sampling. The training set was used for model parameter optimization, the validation set for hyperparameter tuning and monitoring training progress, and the test set for final performance evaluation. The WGA MinION Mk1c dataset served as a completely independent test set for evaluating cross-platform generalization, as it was sequenced on a different nanopore platform and never exposed to the model during development. This design tests whether ChimeraLM learns generalizable sequence features of WGA-induced chimeric artifacts rather than platform-specific technical signatures.

Model architecture

Backbone encoder

ChimeraLM employs the pre-trained HyenaDNA model [35] 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 [38] 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:

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

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

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

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

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

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

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 [44] and PyTorch Lightning [45] 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 [46] 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 [47] 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 [24, 25], DeBreak v1.2 [26], SVIM v2.0.0 [27], and cuteSV v2.1.1 [28]. 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 [48], DELLY v1.5.0 [49], and SvABA v1.1.0 [50]. 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 [51]. 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 [52] 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 [30] (GitHub commit 9a2607e) and 3rd-ChimeraMiner [23] (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 [53], with statistical significance assessed at $p < 0.001$.

Data visualization

Figures were generated using Python with Matplotlib [54] and Seaborn [55].

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.

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

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 designed and implemented the model and computational tool. YL, QG and RY wrote the manuscript. RY supervised this work.

Data Availability.

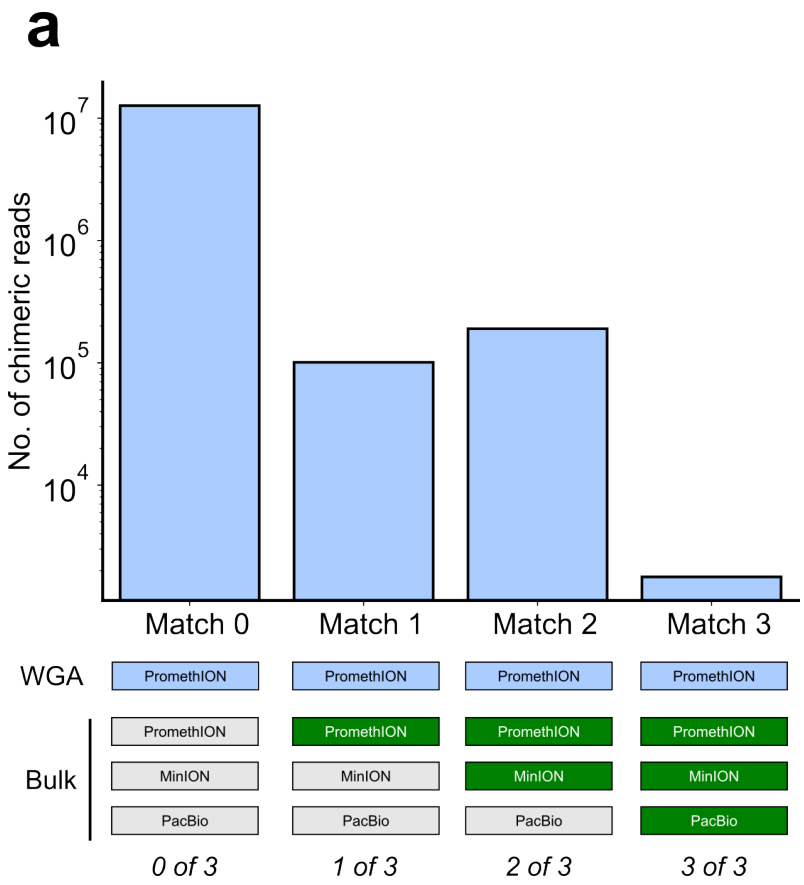
Code Availability. ChimeraLM, implemented in Python, is open source and available on GitHub (https://github.com/ylab-hi/ChimeraLM) under the Apache License, Version 2.0. The package can be installed via PyPI (https://pypi.org/project/chimeralm) using pip, with wheel distributions provided for Windows, 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 14

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



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) stratified by the number of matches found when comparing WGA chimeric reads against bulk sequencing data (x-axis). Match 0 indicates chimeric reads with no matches in bulk data (labeled as artificial chimeric artifacts, $\sim 10^7$ reads). Match 1, 2, and 3 indicate chimeric reads with 1, 2, or 3 matches in bulk data respectively (labeled as biological reads, $\sim 10^5$ reads each). This matching strategy forms the basis for ground truth labeling in supervised training.

GPU Graphics Processing Unit 14, 19, 21, 22

HPC High Performance Computing 21

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

LIANTI Linear Amplification via Transposon Insertion 2, 14

MALBAC Multiple Annealing and Looping-based Amplification Cycles 2, 14
MDA Multiple Displacement Amplification 2, 3, 14

MLP multilayer perceptron 5, 6, 18

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

PacBio Pacific Biosciences 3
PTA Primary Template-directed Amplification 2

SA Supplementary Alignment 16, 20, 21
SV Structural Variation 1–6, 9–12, 14, 15, 17, 20, 21

TRA translocation 9, 11

WGA Whole Genome Amplification 1–17, 19–23

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

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

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

[5] Leung, M. L. *et al.* Highly multiplexed targeted dna sequencing from single nuclei. *Nature Protocols* 214–235 (2016).

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

[7] 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).

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

[9] Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annual Review of Genomics and Human Genetics* 79–102 (2015).

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

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

[12] Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216 (2006).

[13] Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).

[14] 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).

[15] Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nature Communications* **14**, 5164 (2023).

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

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

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

[19] Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences* 11923–11928 (2015).

[20] 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).

[21] Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research* **11**, 1095–1099 (2001).

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

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

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

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

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

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

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

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

[30] 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).

[31] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[51] Guo, Q., Li, Y., Wang, T., Ramakrishnan, A. & Yang, R. Octopusv and tentaclesv: a one-stop toolkit for multi-sample, cross-platform structural variant comparison and analysis. *bioRxiv* (2025). URL https://www.biorxiv.org/content/10.1101/2025.03.24.645012v1.

[52] 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).

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

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

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