ChimeraLM: A genomic language model that distinguishes true structural variants from artifacts in long-read whole genome amplification

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, yet this process introduces extensive chimeric artifacts that manifest as false-positive Structural Variation (SV) and compromise downstream analyses. Here, we developed ChimeraLM, a Genomic Language Model (GLM) that accurately identifies and removes WGA-induced chimeric artifacts. The model architecture integrates Hyena operators with attention mechanisms optimized for variable-length genomic sequences. ChimeraLM achieved state-of-the-art performance in chimeric detection (F1 score: **0.81**) and dramatically outperformed existing approaches, reducing chimeric contamination by $\sim$ **95**% compared to 0% for established tools. When applied to SV analysis, ChimeraLM improved the validation rate of detected events by **12-16** fold while preserving $>$ **87**% of genuine SV across multiple Nanopore sequencing platforms. Importantly, ChimeraLM processing normalized SV type distributions toward bulk sequencing profiles, eliminating the characteristic false-positive inversion (INV) bias of unprocessed WGA data. This approach directly addresses a fundamental bottleneck in single-cell genomics, enables more reliable genomic analysis of individual cells for applications spanning basic research, clinical diagnostics, and therapeutic development.

Keywords: Whole Genomics Amplification, 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, 2]. This approach has proven instrumental in uncovering rare cell types, tracking developmental trajectories, and identifying somatic mutations that drive disease progression. However, the limited DNA content in a single cell—typically only a few picograms—poses significant technical challenges for comprehensive genomic analysis [3, 4].

To overcome this limitation, WGA has become essential for single-cell genomic studies [5, 6]. Various WGA techniques, including Multiple Displacement Amplification (MDA), Multiple Annealing and Looping-based Amplification Cycles (MALBAC), and other emerging methods, can amplify the entire genome from a single cell by several orders of magnitude, generating sufficient DNA material for high-coverage sequencing [7–10]. This amplification enables the depth and breadth of coverage necessary for reliable variant calling, copy number analysis, and structural variation detection.

Despite its critical role, WGA introduces systematic artifacts that significantly impact downstream analyses [11, 12]. Among the most problematic are chimeric sequences—artificial DNA constructs formed when fragments from different genomic loci are erroneously joined during amplification [10–12]. These chimeric artifacts manifest as false-positive structural variations that do not exist in the original cell [11], posing substantial challenges for accurate SV detection and potentially leading to misinterpretation of genomic rearrangements and their biological significance.

Current computational approaches for identifying WGA-induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [12, 13]. However, these methods often fail to distinguish genuine structural variations from amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns or occur in repetitive genomic regions [14, 15]. This lack of robust artifact detection has limited the reliability of structural variant analysis in single-cell studies and hindered the full realization of single-cell genomics' potential.

Here, we developed ChimeraLM, a genomic language model specifically designed to detect chimeric artifacts introduced by whole genome amplification. By leveraging deep learning to capture sequence patterns and contextual information in genomic reads [16–18], 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.

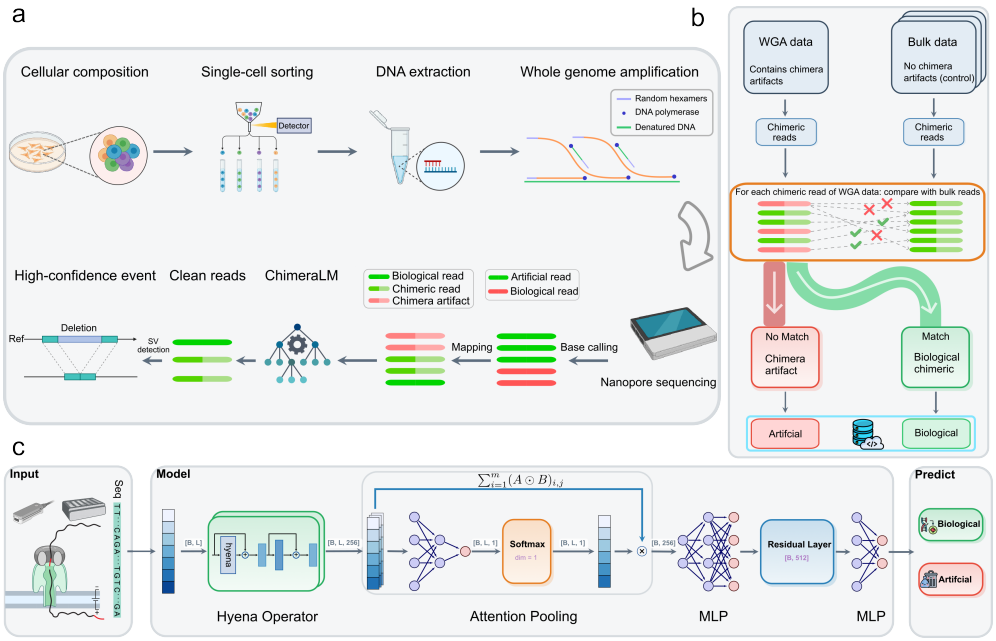


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. Created with BioRender.com. (a) Integration of ChimeraLM into single-cell genomic analysis workflows. Single cells are isolated from heterogeneous cellular populations through sorting technologies, followed by DNA extraction and WGA to generate sufficient material for sequencing. WGA introduces chimeric artifacts through random hexamers, DNA polymerase extension, and denatured DNA template switching. Sequencing reads from WGA-amplified samples contain both biological reads (green) and chimeric artifacts (red). ChimeraLM processes these reads to classify them as biological or artificial, enabling clean reads to proceed to SV analysis for high-confidence event detection such as deletions. (b) Dataset construction strategy for supervised learning. Training data is generated by comparing WGA sequencing reads against matched bulk sequencing data from the same biological sample. Bulk data contains only genuine biological sequences (no chimeric artifacts), while WGA data contains both biological reads and chimeric artifacts. Each WGA read is aligned against bulk data: reads that successfully match are labeled as "biological" (green pathway), while reads that fail to match are labeled as "artificial chimeric" (red pathway). This comparative approach provides reliable ground truth labels for training ChimeraLM in a supervised learning framework. (c) ChimeraLM neural network architecture. Input DNA sequences are tokenized and processed through a deep learning pipeline optimized for genomic sequence analysis. The architecture employs Hyena operators for efficient long-range dependency modeling, followed by attention pooling to aggregate variable-length sequence features. multilayer perceptron (MLP) components with residual connections process the pooled features to learn complex patterns distinguishing biological sequences from chimeric artifacts. The final output layer produces binary classification probabilities, predicting whether each input sequence represents a biological read or an artificial chimeric read.

Results

ChimeraLM integrates seamlessly into single-cell genomic workflows

To systematically address WGA-induced chimeric artifacts, we developed ChimeraLM as an integrated component of single-cell genomic analysis pipelines (Fig. 1a). Our

approach leverages the standard single-cell workflow, beginning with cellular isolation through Fluorescence-activated cell sorting (FACS) or microfluidics-based sorting, followed by DNA extraction and whole genome amplification using established protocols.

Amplified genomic material is then processed through long-read sequencing platforms such as Nanopore technology to generate comprehensive genomic coverage. ChimeraLM operates at a critical juncture in the analysis pipeline, positioned between initial read processing and downstream analyses such as structural variant detection (Fig. 1a). Following standard quality filtering and read cleaning procedures, ChimeraLM evaluates each sequencing read to classify it as either biological or chimeric artifact. This binary classification enables selective retention of authentic genomic sequences while filtering out amplification artifacts before they impact downstream analyses.

The filtered, high-quality biological reads are subsequently processed through conventional structural variant detection algorithms, enabling identification of genuine genomic alterations such as deletion (DEL), duplication (DUP), and other rearrangements. By removing chimeric sequences upstream of variant calling, ChimeraLM ensures that detected SV represent true biological events rather than technical artifacts (Fig. 1a).

This workflow design allows ChimeraLM to integrate with existing single-cell genomic pipelines without requiring substantial modifications to established protocols, providing a versatile solution for improving the accuracy of genomic studies across diverse research applications.

Training dataset construction enables supervised learning of chimeric patterns

To train ChimeraLM for accurate chimeric artifact detection, we developed a dataset construction strategy that leverages paired WGA and bulk sequencing data from the same biological samples (Fig. 1b). This approach exploits the fundamental difference between these datasets: while WGA data contains both biological reads and chimeric artifacts introduced during amplification, bulk sequencing from the same sample contains only genuine biological sequences.

Our ground truth labeling strategy compares each WGA read against the bulk sequencing dataset (Fig. 1b). Reads that successfully match bulk data are classified as "biological," indicating they represent authentic genomic sequences present in the original sample. Conversely, reads that fail to match bulk sequences are labeled as "artificial chimeric" artifacts, representing artificial constructs generated during WGA rather than genuine genomic content (Fig. 1b).

Application of this matching strategy to the PC3 cell line dataset revealed that the majority of chimeric reads ($\sim 10^7$ reads) showed no matches in bulk data and were classified as artificial, while smaller subsets ($\sim 10^5$ reads each) showed 1, 2, or 3 matches and were classified as biological (Extended Data Fig. 1). This comparative approach generates a comprehensive labeled dataset where each WGA read receives binary classification based on its presence or absence in the matched bulk control,

capturing the full spectrum of chimeric artifacts naturally occurring during WGA while providing reliable ground truth labels for model training.

Following dataset construction, we partitioned the labeled reads into training (70%), validation (20%), and test (10%) sets to ensure robust model development and unbiased performance evaluation. The training set was used for model parameter optimization, the validation set for hyperparameter tuning and model selection, and the test set was reserved for final performance assessment. This rigorous data splitting strategy ensures that ChimeraLM's reported performance metrics reflect its ability to generalize to previously unseen WGA data.

ChimeraLM architecture leverages modern genomic language modeling advances

ChimeraLM employs a neural architecture specifically designed for genomic sequence analysis and chimeric artifact detection (Fig. 1c). The model operates at single-base pair resolution and accepts DNA sequences as input, which are tokenized and encoded into numerical representations suitable for deep learning processing. This encoding preserves the sequential nature of genomic information while enabling efficient computation.

The core architecture consists of Hyena operators [19], a recent advancement in sequence modeling that provides computational advantages over traditional transformer attention mechanisms while maintaining the ability to capture long-range dependencies in genomic sequences. Hyena operators enable ChimeraLM to process variable-length sequencing reads efficiently while learning complex patterns that distinguish biological sequences from chimeric artifacts. The Hyena operators are initialized with HyenaDNA [18], a pre-trained long-context genomic language model, providing a strong foundation for learning genomic sequence features relevant to chimeric detection.

Following the Hyena operator layers, ChimeraLM incorporates an attention pooling mechanism that aggregates sequence-level features. This pooling strategy allows the model to handle reads of varying lengths while focusing computational attention on the most informative regions for chimeric detection. The attention weights learned during training provide interpretability into which sequence features contribute most strongly to classification decisions.

The aggregated features are then processed through multiple MLP components arranged in a residual architecture. This design enables gradient flow optimization during training while allowing the model to learn both low-level sequence motifs and high-level compositional patterns indicative of chimeric artifacts. Residual connections help prevent vanishing gradients and improve model convergence during training on large genomic datasets.

The final output layer produces a binary classification predicting whether each input sequence represents a biological read or an artificial chimeric artifact. This end-to-end architecture enables ChimeraLM to learn directly from raw sequence data without requiring manual feature engineering, allowing the model to discover complex patterns that may not be apparent through traditional bioinformatics approaches.

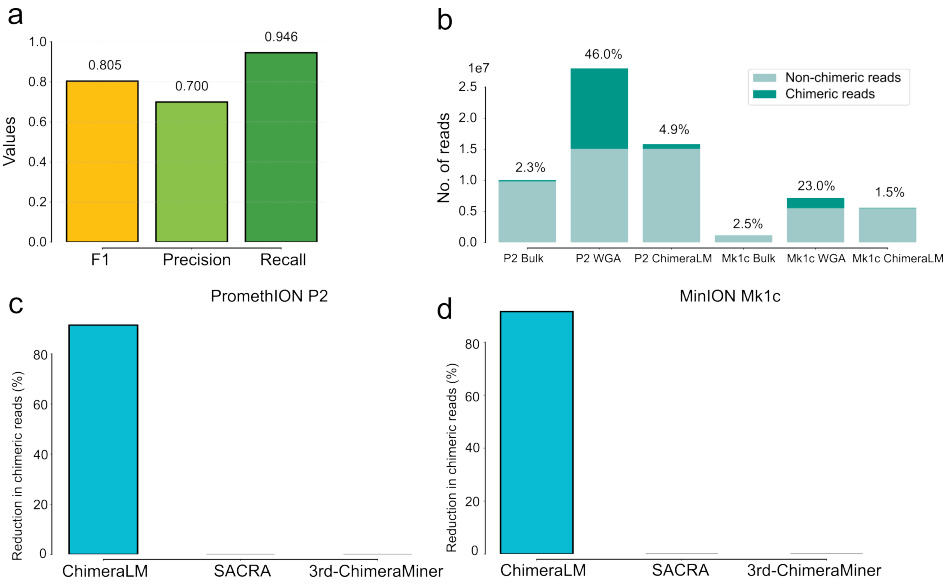


Fig. 2 ChimeraLM performance evaluation and benchmarking against existing methods. (a) ChimeraLM performance metrics on test dataset. Bar chart showing F1 score (0.805), precision (0.700), and recall (0.946) achieved by ChimeraLM on held-out test data for binary classification of biological versus chimera artifact. (b) Chimeric read reduction across sequencing platforms. Stacked bar charts comparing chimeric read proportions (dark teal) versus non-chimeric reads (light teal) across bulk sequencing, WGA, and ChimeraLM-processed samples for PC3 cell line data. Left side shows PromethION P2 platform data with chimeric rates of 2.3% (bulk), 46.0% (WGA), and 4.9% (ChimeraLM). Right side shows MinION Mk1c platform data with chimeric rates of 2.5% (bulk), 23.0% (WGA), and 1.5% (ChimeraLM). (c) Benchmarking on P2 platform data. Percentage reduction in chimeric reads achieved by ChimeraLM (90%) compared to existing computational methods SACRA and 3rd-ChimeraMiner (both 0% reduction). (d) Benchmarking on Mk1c platform data. Percentage reduction in chimeric reads achieved by ChimeraLM (90%) compared to SACRA and 3rd-ChimeraMiner (both 0% reduction) using MinION Mk1c sequencing data.

ChimeraLM achieves high performance in chimeric artifact detection

We evaluated ChimeraLM's performance on held-out test data to assess its ability to accurately classify biological and chimeric reads (Fig. 2a). ChimeraLM demonstrated robust performance across key classification metrics, achieving an F1 score of 0.805, which balances precision and recall for the binary classification task. The model exhibited high recall of 0.946, indicating that it successfully identified 94.6% of true chimeric artifacts in the test dataset, minimizing the risk of retaining false-positive structural variants in downstream analyses. The precision of 0.700 demonstrates that 70.0% of reads classified as chimeric were indeed artifacts, representing a trade-off that prioritizes comprehensive artifact removal while maintaining reasonable specificity.

ChimeraLM reduces chimeric artifact burden across sequencing platforms

To evaluate ChimeraLM's practical impact, we applied the trained model to PC3 cell line data generated on two Nanopore long-read sequencing platforms: PromethION P2 and MinION Mk1c (Fig. 2b). The analysis revealed substantial differences in chimeric read proportions between bulk sequencing, WGA samples, and ChimeraLM-processed data.

For P2 platform data, bulk sequencing exhibited a low baseline chimeric rate of 2.3% among 10,065,403 total reads, consistent with the expected minimal artifact rate in non-amplified samples. In contrast, WGA amplification dramatically increased the chimeric burden to 46.0% of reads in a dataset containing 28,027,667 total sequences. ChimeraLM processing effectively reduced this chimeric proportion to 4.9% while retaining 15,833,834 high-quality biological reads, representing a ∼10-fold reduction in artifact contamination compared to unprocessed WGA data.

Similar results were observed for Mk1c platform data, where bulk sequencing contained 2.5% chimeric reads among 1,140,363 total sequences. WGA amplification increased chimeric contamination to 23.0% of 7,193,945 total reads. ChimeraLM processing reduced the chimeric fraction to 1.5% while preserving 5,610,252 biological reads, achieving nearly complete artifact removal that approached the quality of bulk sequencing data.

These results demonstrate ChimeraLM's effectiveness across different Nanopore sequencing platforms and highlight the substantial chimeric artifact burden introduced by WGA, which varies between platforms but is consistently and dramatically reduced by ChimeraLM processing.

ChimeraLM outperforms existing chimera artifact detection tools

We benchmarked ChimeraLM against established computational methods for chimeric sequence detection, including SACRA [13] and 3rd-ChimeraMiner [12], using both P2 and Mk1c datasets (Fig. 2c,d). ChimeraLM achieved superior performance compared to existing approaches across both sequencing platforms.

For P2 data, ChimeraLM reduced chimeric read contamination by ∼92%, demonstrating substantial improvement in data quality (Fig. 2c). In contrast, both SACRA and 3rd-ChimeraMiner failed to achieve meaningful chimeric read reduction, showing 0% improvement over unprocessed WGA data. This stark difference highlights the limitations of existing rule-based and alignment-based approaches for detecting complex chimeric artifacts in long-read sequencing data.

Similar performance advantages were observed for Mk1c data, where ChimeraLM again achieved ∼91% reduction in chimeric reads while SACRA and 3rd-ChimeraMiner provided no detectable improvement (Fig. 2d). These results demonstrate that ChimeraLM's deep learning approach captures complex sequence patterns that are not effectively identified by traditional computational methods.

The consistent superior performance across different sequencing platforms establishes ChimeraLM as a significant advance over existing tools, providing researchers

with a reliable method for improving single-cell genomic data quality regardless of the specific long-read sequencing technology employed.

ChimeraLM improves structural variant calling accuracy by reducing false positives

To evaluate ChimeraLM's impact on downstream structural variant analysis, we constructed a gold-standard SV dataset using bulk sequencing data (Fig. 3a) and assessed SV detection performance on both P2 and Mk1c platforms with and without ChimeraLM filtering (Fig. 3b,c). The gold standard was constructed by integrating long-read (Nanopore P2 and Mk1c) and short-read (Illumina HiSeq) sequencing platforms with multiple SV callers, retaining only high-confidence events supported by $\geq$2 callers per platform and $\geq$2 datasets (Fig. 3a). This analysis directly measures ChimeraLM's ability to improve the biological relevance of SV calls by comparing detected events against high-confidence reference data.

For P2 platform data, unprocessed WGA samples yielded 3,609,914 total structural variant calls, of which only 8,815 (0.24%) were supported by the gold standard (Fig. 3b). The remaining 3,601,099 calls represented unsupported events likely arising from chimeric artifacts and other amplification biases. ChimeraLM processing dramatically improved this ratio, reducing total SV calls to 305,570 while maintaining 8,067 supported events (2.64% of total calls). This represents a $\sim$12-fold increase in validation rate while preserving 91.5% of true positive events, compared with WGA data.

Similar improvements were observed for Mk1c data, where raw WGA processing identified 451,390 total SV with 7,193 supported events (1.59%) (Fig. 3c). ChimeraLM filtering reduced the total to 38,432 calls while retaining 6,269 supported events, achieving a 16.3% validation rate ($\sim$10-fold increase) and preserving 87.2% of true positive SV compared with WGA data. The consistent performance across platforms demonstrates ChimeraLM's robust ability to eliminate false-positive SV while maintaining detection sensitivity for genuine genomic alterations.

ChimeraLM normalizes structural variant type distributions toward bulk sequencing profiles

We analyzed the distribution of SV types to assess whether ChimeraLM processing restores SV profiles characteristic of high-quality bulk sequencing data (Fig. 3d,e). This analysis tests the hypothesis that WGA introduces systematic biases in the apparent frequency of different SV classes, which should be corrected by effective chimeric artifact removal.

In both P2 and Mk1c datasets, bulk sequencing exhibited balanced distributions across DEL, DUP, INS, INV, and TRA events, with relatively modest numbers reflecting the stringent quality filtering applied to establish the gold standard (Fig. 3d,e). Unprocessed WGA data showed dramatically skewed distributions dominated by spurious INV events, consistent with literature reports that INV are frequently artificial in amplified samples due to template switching during WGA [11, 12].

ChimeraLM processing substantially normalized these distributions, reducing the overwhelming preponderance of false-positive INV while maintaining more balanced representation of other SV types. The filtered data profiles more closely resembled bulk sequencing distributions, indicating that ChimeraLM successfully identifies and removes the systematic biases introduced by WGA without eliminating legitimate SV of other classes.

Characterization of chimeric artifact-supported false-positive structural variants

To understand the specific types of false-positive SV that would be retained without ChimeraLM filtering, we analyzed the SV events in unfiltered WGA data that were specifically supported by reads classified as chimeric artifacts by ChimeraLM (Fig. 3f,g). This analysis profiles the "false-positive landscape" that researchers would encounter without effective chimeric artifact removal.

For P2 data, chimeric artifact-supported SV were overwhelmingly dominated by INV (88.4% of events), with smaller contributions from DEL (5.1%), DUP (3.4%), and INS (3.0%) (Fig. 3f). This extreme bias toward inversions confirms that template switching during WGA predominantly manifests as apparent INV events in downstream SV calling pipelines.

Mk1c data showed a similar but even more pronounced pattern, with INV comprising 92.4% of chimeric artifact-supported events, followed by DEL (3.8%), DUP (2.4%), and INS (1.4%) (Fig. 3g). The consistency of this pattern across platforms indicates that inversion artifacts represent the primary mode of false-positive SV generation in WGA workflows.

These results demonstrate that without ChimeraLM, genomic studies would be severely compromised by false-positive INV calls, potentially leading to misinterpretation of chromosomal instability, copy number profiles, and other key genomic features. ChimeraLM's ability to identify and remove these specific artifacts represents a critical advancement for accurate SV analysis.

ChimeraLM predictions correlate with chimeric alignment complexity

To validate ChimeraLM's classification accuracy, we analyzed the distribution of chimeric alignments per chimeric read, comparing how ChimeraLM classified these known chimeric sequences as biological versus artificial (Extended Data Fig. 2a,b). This provides independent validation by examining whether ChimeraLM correctly identifies the most structurally complex chimeric artifacts.

Among chimeric reads, those classified as "artificial" by ChimeraLM predominantly exhibited 2 chimeric alignments per read ($\sim 1.0 \times 10^7$ and $\sim 1.4 \times 10^6$ reads in P2 and Mk1c data, respectively), representing simple two-part chimeric structures. Smaller fractions showed 3+ alignments ($\sim 2.1 \times 10^6$ and $\sim 0.2 \times 10^6$ reads in P2 and Mk1c data, respectively), indicating more complex multi-fragment chimeras from sequential template switching events.

Importantly, chimeric reads classified as "biological" by ChimeraLM showed minimal representation across all alignment complexity categories, suggesting these may represent genuine structural variants or less disruptive chimeric events that preserve biological relevance.

This pattern demonstrates that ChimeraLM successfully prioritizes the most structurally complex and potentially problematic chimeric artifacts for removal, while preserving chimeric reads that may still retain biological information. The consistency across both datasets validates ChimeraLM's ability to distinguish between different classes of chimeric complexity.

ChimeraLM demonstrates capacity to learn biologically relevant sequence features

To investigate whether ChimeraLM can capture biologically meaningful features for chimeric artifact detection, we examined the attention weight distributions from the model's pooling mechanism. Attention weights indicate which sequence regions contribute most strongly to individual classification decisions, providing potential insight into learned patterns.

We present representative examples where ChimeraLM's attention mechanism shows focused activity at chimeric junction sites (Fig. 4 and Extended Data Fig. 3). In these chimeric reads, the attention profiles exhibited predominantly low baseline weights with pronounced peaks coinciding with chimeric junctions where reads transition between reverse-complemented and forward-oriented sequences. These junctions represent artificial joining points where DNA fragments from different genomic loci were ligated during WGA.

The alignment pattern illustrates the structural signature present in these examples (Fig. 4a). In the first example, the read portion aligns in reverse orientation (red), while the downstream portion aligns in forward orientation to a distant genomic location (green). This discordant orientation pattern represents a characteristic feature of WGA-induced chimeric artifacts.

For these specific examples, quantitative analysis showed that attention weights within 100 bp windows ($\pm$50 bp) centered on chimeric junctions exhibited significantly higher values compared to background regions (Fig. 4b). The median attention weight in the junction window was approximately 2.7-fold higher than background (window median: 0.0008; background median: 0.0003; $p < 0.001$, Wilcoxon rank-sum test).

These examples align with the proposed mechanism of chimera formation during WGA (Fig. 4c). Original DNA fragments from distant genomic loci undergo random orientation changes during amplification. Template switching events cause these independently oriented fragments to be artificially joined, creating chimeric constructs with orientation discontinuities at junction sites.

These case studies demonstrate that ChimeraLM has the capacity to learn biologically interpretable features related to chimeric junction sites, though the prevalence and consistency of this attention pattern across the full dataset remains to be systematically characterized. The observation that the model can focus on mechanistically relevant sequence features in at least some cases provides evidence that ChimeraLM's

learned representations may incorporate structural signatures of WGA artifacts rather than relying solely on other sequence characteristics.

Discussion

ChimeraLM addresses a fundamental bottleneck that has limited the widespread adoption and reliability of single-cell genomic approaches. While WGA has enabled genomic analysis of individual cells, the systematic introduction of chimeric artifacts has remained an unsolved challenge that compromises downstream interpretations and limits biological insights.

Traditional approaches to managing WGA artifacts have focused on post-hoc filtering of SV calls or coverage-based correction methods [12, 13]. ChimeraLM represents a paradigm shift toward proactive identification of problematic sequences before they impact downstream analyses. This upstream intervention strategy addresses the root cause of analytical errors rather than attempting to correct their consequences after variant calling.

The success of this approach demonstrates the power of modern genomic language models to capture complex sequence patterns that are not readily apparent through traditional bioinformatics methods. Unlike rule-based approaches that rely on predefined criteria, ChimeraLM learns directly from data, enabling discovery of subtle features that distinguish authentic biological sequences from amplification artifacts [10, 12, 18]. This data-driven approach is particularly valuable for complex genomic phenomena where explicit rules may be insufficient or unknown.

The demonstrated effectiveness of ChimeraLM has broader implications for single-cell genomics methodology. The ability to substantially improve data quality through computational approaches reduces the experimental burden of optimizing amplification protocols and may enable researchers to focus on biological questions rather than technical optimization. This could accelerate adoption of single-cell genomic approaches in laboratories with limited specialized expertise in amplification chemistry.

Furthermore, improved reliability of SV detection opens new avenues for applications that have been previously constrained by high false-positive rates. Studies of chromosomal instability, copy number evolution, and SV burden in individual cells become more feasible when researchers can have confidence in the authenticity of detected events.

ChimeraLM's success exemplifies the transformative potential of language model approaches in genomics. The recent emergence of foundation models for biological sequences has demonstrated remarkable capabilities across diverse tasks, but most applications have focused on prediction of molecular phenotypes or functional annotations. ChimeraLM represents one of the first applications of GLM to quality control and data preprocessing, suggesting that these approaches may have broader utility for improving experimental data quality than previously recognized.

The architectural innovations incorporated in ChimeraLM, particularly the use of Hyena operators for efficient long-range modeling, may have applications beyond

chimeric detection [18, 19]. Similar approaches could potentially address other quality control challenges in genomics, such as contamination detection, adapter artifact identification, or systematic error correction in diverse sequencing technologies.

While ChimeraLM represents a significant advance, several limitations merit consideration. The requirement for paired bulk sequencing data for training constrains the initial application scope, though this limitation may be addressable through transfer learning approaches as the method matures. The current focus on Nanopore platforms, while representing the most common long-read technology for single-cell applications, leaves questions about broader platform compatibility.

More fundamentally, the binary classification approach assumes a clear distinction between biological and artificial sequences. In reality, some chimeric events may represent genuine biological phenomena, such as chromothripsis or complex structural rearrangements. Future developments may need to incorporate more nuanced classification schemes that can distinguish between different types of chimeric events based on their likely biological relevance.

The success of ChimeraLM suggests several promising directions for future development. Integration with real-time sequencing platforms could enable immediate quality assessment and adaptive sampling strategies. Extension to other single-cell genomic applications, such as chromatin accessibility or methylation analysis, could address analogous quality control challenges in emerging single-cell methods.

The interpretability features of modern language models could be leveraged to provide insights into the sequence features that distinguish chimeric artifacts, potentially informing development of improved amplification protocols. This feedback loop between computational analysis and experimental optimization could drive continuous improvement in single-cell genomic methods.

ChimeraLM demonstrates that sophisticated computational approaches can effectively address fundamental technical challenges that have limited single-cell genomics. By providing a robust solution to chimeric artifact detection, this work removes a significant barrier to reliable single-cell genomic analysis and opens new possibilities for biological discovery and clinical application. As single-cell approaches become increasingly central to modern biology and medicine, computational tools like ChimeraLM will be essential for realizing their full potential.

Methods

MDA sequencing

Training data construction

Paired sequencing dataset generation

Training data for ChimeraLM was constructed using a novel comparative approach that leverages paired WGA and bulk sequencing datasets from identical biological samples (PC3 cell line). This methodology exploits the fundamental difference between amplified and non-amplified samples to establish reliable ground truth labels for supervised learning. Bulk sequencing data serves as a reference standard containing only genuine biological sequences, while WGA data contains both authentic genomic

content and amplification-induced chimeric artifacts. Sample preparation involved processing PC3 cell line to generate matched WGA and bulk sequencing datasets.

Ground truth labeling methodology

The core innovation of our training data construction lies in the comparative labeling strategy that automatically generates binary classification labels. Each chimeric read from the WGA dataset was systematically compared against the corresponding bulk sequencing data through sequence alignment analysis to determine its biological authenticity.

Chimeric reads from WGA data were compared against the bulk sequencing dataset to identify chimera present in both datasets. WGA reads that successfully aligned to bulk sequences were classified as biological, indicating they represent authentic genomic content preserved through the amplification process. Conversely, WGA reads that failed to align to any chimeric sequences in the bulk dataset were labeled as artificial chimeric artifacts, representing spurious sequences generated during the amplification process. The comparison-based classification approach ensures that training labels reflect objective evidence of sequence authenticity. This methodology captures the full spectrum of naturally occurring chimeric artifacts while providing reliable positive examples of biological sequences that successfully traverse the amplification process.

Dataset partitioning and stratification

The complete labeled dataset was partitioned into training, validation, and test sets using stratified sampling to maintain balanced representation of biological and artificial sequences across all data splits. The training set comprised 70% of the total data and was used for model parameter optimization during supervised learning. The validation set contained 20% of the data and served for hyperparameter tuning, model selection, and monitoring training progress. The remaining 10% was reserved as a held-out test set for final performance evaluation and remained completely isolated from the training process.

Stratification ensured that each data partition maintained similar proportions of biological and artificial sequences, preventing training bias that could arise from imbalanced class distributions. Random sampling within strata was employed to minimize systematic biases while maintaining statistical representativeness of the overall dataset characteristics.

Model architecture

ChimeraLM uses a pre-trained HyenaDNA which provides robust genomic sequence representations learned from large-scale genomic data. The backbone model handles DNA sequence tokenization using single base pair resolution, converting nucleotide sequences into token representations.

Input sequences are processed through the Hyena operators. The backbone generates contextualized hidden representations for each position in the input sequence, capturing both local sequence motifs and long-range dependencies essential for distinguishing biological sequences from chimeric artifacts.

ChimeraLM employs a binary sequence classifier designed specifically for genomic sequence classification tasks. The classification head processes the hidden states through a multi-stage architecture optimized for variable-length sequence classification. The classifier implements attention-based pooling to aggregate variable-length sequence representations into fixed-size feature vectors.

$$\text{attention\_weights} = \text{softmax}(\text{Linear}(\text{GELU}(\text{Linear}(\text{hidden\_states}))))$$

The attention pooling mechanism computes position-wise attention weights through a two-layer neural network where the first linear layer projects from the hidden dimension to 256 dimensions, followed by Gaussian Error Linear Unit (GELU) activation and a second linear layer projecting to a single attention score per position. The attention weights are normalized using *softmax* and applied to compute the weighted sum of hidden states across sequence positions. When attention masks are provided for variable-length sequences, the attention weights are masked to exclude padding tokens and renormalized to ensure proper probability distribution over valid sequence positions. This approach accommodates the natural variability in sequencing read lengths without introducing computational inefficiency or artificial sequence information.

The pooled sequence representation is processed through a MLP with residual connections. The MLP uses two hidden layers, each consisting of linear transformation to 512 hidden dimensions, GELU activation function, and dropout regularization with rate 0.1. This residual design enables stable gradient flow during training and improved representation learning for complex genomic patterns.

The final classification layer maps the processed features to binary classification logits using a linear transformation with output dimension 2. The model outputs two logits corresponding to biological and artificial classes rather than a single probability, enabling the use of cross-entropy loss during training. Classification predictions are made by selecting the class with the highest logit value.

Model training and optimization

Training data preparation

ChimeraLM was trained using the paired WGA and bulk sequencing dataset constructed from PromethION P2 platform data. The tokenizer was initialized from the pre-trained HyenaDNA model to maintain consistency with the backbone architecture and preserve the benefits of pre-training on large-scale genomic data. This approach ensures that input sequences are processed using the same tokenization scheme employed during the backbone model's original training, enabling effective transfer learning for the chimeric detection task.

Model training framework and optimization

Model training was implemented within PyTorch [20] and PyTorch Lightning framework [21], providing a standardized interface for supervised classification tasks with automatic handling of training loops, validation procedures, and model checkpointing. The training process employed mixed-precision computation using bf16-mixed

precision to accelerate training while maintaining numerical stability for gradient computation.

The optimization procedure utilized the AdamW optimizer [22], an extension of the Adam optimizer [23] that incorporates decoupled weight decay regularization. The optimizer was configured with a learning rate of $1 \times 10^{-}4$ and weight decay coefficient of 0.01. The AdamW optimizer updates model parameters θ according to the following formulation:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right)$$

where α represents the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates, ϵ is a small constant for numerical stability, and λ denotes the weight decay coefficient. The decoupled weight decay term $\lambda\theta_t$ provides regularization that is independent of the gradient-based updates.

Early stopping was implemented with a patience of 10 epochs to prevent over-fitting and automatically terminate training when validation performance ceased to improve. This approach ensures optimal model generalization by identifying the point of best validation performance rather than training to completion. The training process used a fixed random seed of 12345 to ensure reproducibility across multiple training runs and facilitate comparison of different model configurations.

Loss function and objective formulation

The training objective employed cross-entropy loss for the binary classification task, providing probabilistically grounded optimization that encourages the model to produce well-calibrated probability estimates. For a training example with true class label $y \in \{0, 1\}$ and model output logits $z = [z_0, z_1]$, the cross-entropy loss is computed as:

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

where the softmax function converts logits to normalized probability distributions over the two classes (biological and artificial). The cross-entropy formulation provides strong gradients for misclassified examples while allowing confident predictions to contribute minimal loss, enabling efficient learning of the decision boundary between biological sequences and chimeric artifacts.

Data loading and computational configuration

Training employed a batch size of 48 sequences per batch, balancing computational efficiency with memory constraints and gradient stability. Data loading was optimized using 30 parallel workers to minimize I/O bottlenecks and ensure continuous data supply to the Graphics Processing Unit (GPU) during training.

The training infrastructure utilized GPU acceleration (two A100 GPUs) to enable efficient processing of the large-scale genomic sequence datasets. Mixed-precision training with brain floating-point 16-bit (bf16) format was employed to reduce memory requirements and accelerate computation while maintaining numerical precision sufficient for stable gradient computation and model convergence.

Model validation and evaluation

Model performance was continuously monitored throughout training using the validation dataset. The model evaluation metrics included accuracy, precision, recall and the F1 score, calculated using the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Validation metrics were computed at the end of each epoch to track model improvement and trigger early stopping when performance plateaued. The validation process used the same data preprocessing and tokenization procedures as training to ensure consistent evaluation conditions.

The final model selection was based on optimal validation performance as determined by the early stopping callback. This approach ensures that the reported model represents the configuration with the best generalization performance rather than the final training epoch, preventing overfitting and providing realistic performance estimates for unseen data.

Training infrastructure and implementation

The training process was managed using Hydra [24] configuration framework to enable reproducible experimentation and systematic hyperparameter management. Default callbacks were employed for standard training monitoring, including metrics logging, model checkpointing, and training progress tracking. The configuration system allowed for systematic exploration of hyperparameters while maintaining experimental reproducibility and version control.

All training experiments were conducted using consistent hardware configurations and software environments to ensure reliable performance comparisons. The model training process typically required 24-48 hours depending on dataset size and convergence characteristics, with automatic checkpointing enabled to prevent loss of progress due to hardware failures or interruptions.

SV evaluation

Sniffles2 [25, 26], Manta [27], DELLY [28], DeBreak [29], SVIM [30], cuteSV [31], and SvABA [32] were used.

Construction of gold-standard SV dataset

Figure plotting and visualization

The figures were created using Python with libraries including Matplotlib [33] and Seaborn [34].

Computing resource

All computations were performed on a High Performance Computing (HPC) server equipped with a 64-core Intel(R) Xeon(R) Gold 6338 CPU and 256 GB of RAM. The server was also configured with two NVIDIA A100 GPUs, each with 80 GB of memory, enabling efficient processing of both CPU-intensive tasks and GPU-accelerated deep learning workloads.

Supplementary information.

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

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. An interactive demo is available on Hugging Face (https://huggingface.co/spaces/yangliz5/chimeralmr), allowing users to test ChimeraLM's functionality without local installation. For large-scale analyses, we recommend using ChimeraLM on systems with GPU acceleration. Detailed system requirements and optimization guidelines are available in the repository's documentation.

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

DEL deletion 4, 8, 9, 22
DUP duplication 4, 8, 9, 22

FACS Fluorescence-activated cell sorting 4

GELU Gaussian Error Linear Unit 14
GLM Genomic Language Model 1, 12
GPU Graphics Processing Unit 16–18

HPC High Performance Computing 17

INS insertion 8, 9, 22
INV inversion 1, 8, 9, 22

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

SV Structural Variation 1–4, 8, 9, 11, 22

TRA translocation 8, 22

WGA Whole Genome Amplification 1–11, 13, 15, 22

References

[1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. *Communications Biology* **7**, 1288 (2024).

[2] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons. *Nature Communications* **15**, 4220 (2024).

[3] Leung, M. L. *et al.* Highly multiplexed targeted DNA sequencing from single nuclei. *Nature Protocols* **11**, 214–235 (2016).

[4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**, 175–188 (2016).

[5] 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* **338**, 1622–1626 (2012).

[6] 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* **16**, 79–102 (2015).

[7] de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).

[8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. *Scientific Reports* **11**, 17171 (2021).

[9] Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences* **112**, 11923–11928 (2015).

[10] Agyabeng-Dadzie, F. *et al.* Evaluating the Benefits and Limits of Multiple Displacement Amplification With Whole-Genome Oxford Nanopore Sequencing. *Molecular Ecology Resources* **25**, e14094 (2025).

[11] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal* **21**, 1688–1696 (2023).

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

[13] 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).

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

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

[16] Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025).

[17] Zhou, Z. *et al. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes*, 1–24 (2024).

[18] Nguyen, E. *et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution*, Vol. 36, 43177–43201 (2023).

[19] Poli, M. *et al. Hyena Hierarchy: Towards Larger Convolutional Language Models*, Vol. 202, 28043–28078 (2023).

[20] Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* , Vol. 32, 8024–8035 (2019).

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

[22] Loshchilov, I. & Hutter, F. *Decoupled Weight Decay Regularization* (2019).

[23] Kingma, D. P. & Ba, J. L. *Adam: A Method for Stochastic Optimization* (2015).

[24] Yadan, O. Hydra - A framework for elegantly configuring complex applications. GitHub repository (2019). URL https://github.com/facebookresearch/hydra. Accessed: 2025-10-01.

[25] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **15**, 461–468 (2018).

[26] Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* **42**, 1571–1580 (2024).

[27] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications **32**, 1220–1222.

[28] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis **28**, i333–i339.

[29] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak **14**, 283.

[30] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads **35**, 2907–2915.

[31] Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV **21**, 189.

[32] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly **28**, 581–591.

[33] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

[34] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

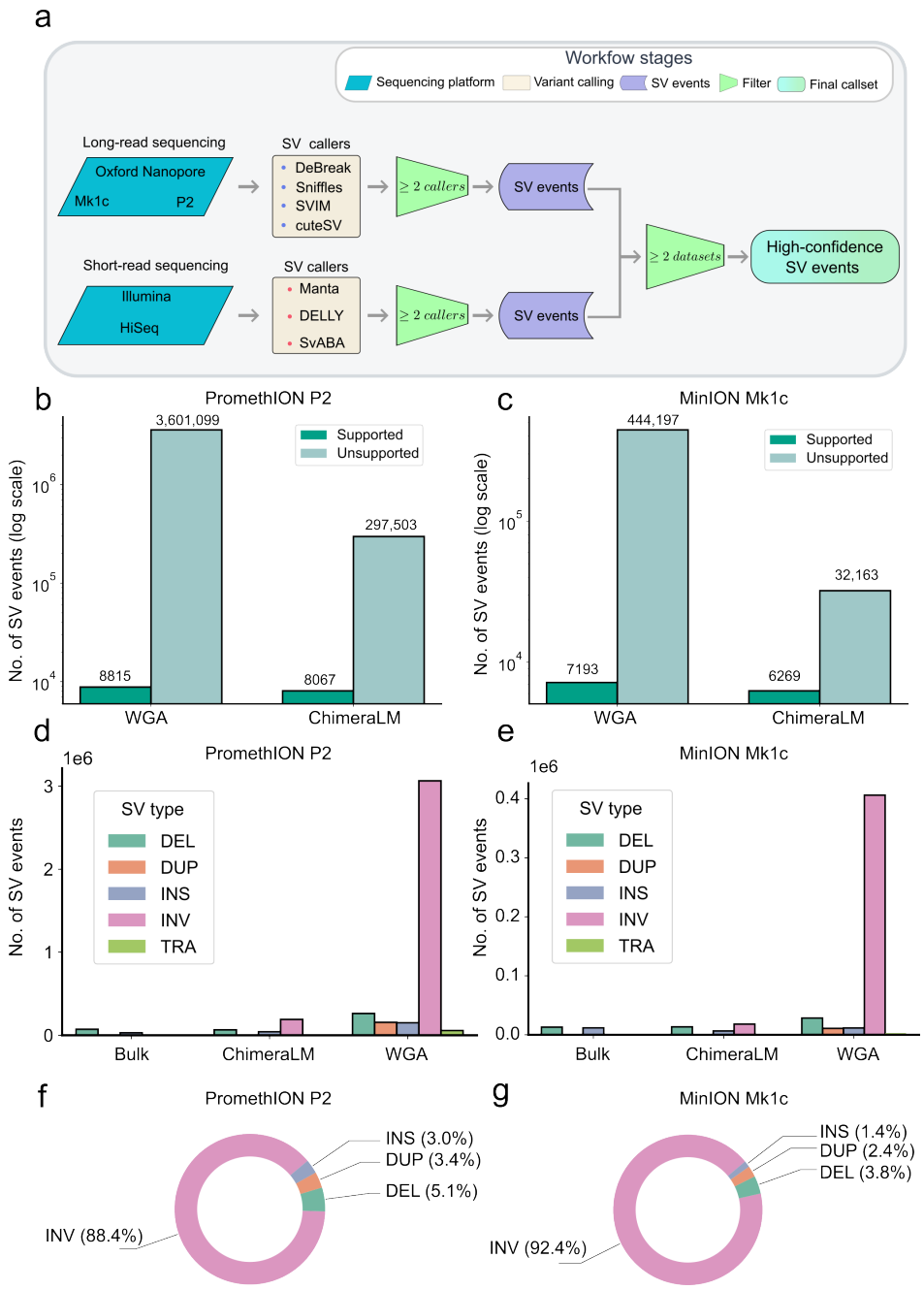


Fig. 3 SV detection accuracy and type distribution analysis. (a) Workflow for constructing gold standard SV dataset from bulk sequencing data. Long-read sequencing (Oxford Nanopore Mk1c and P2) and short-read sequencing (Illumina HiSeq) platforms are used with multiple SV callers. SV events detected by ≥2 callers per platform are filtered, and events supported by ≥2 datasets (both long-read and short-read) are retained as high-confidence SV events for gold standard. (b,c) SV validation using bulk sequencing gold standard. Stacked bar charts showing total SV calls (log scale) classified as supported (dark teal) or unsupported (light teal) events when compared against gold standard. Panel (b) shows PromethION P2 results comparing WGA vs ChimeraLM-filtered data; panel (c) shows MinION Mk1c results. Numbers above bars indicate absolute counts of supported/unsupported events. (d,e) SV type distributions across sample processing methods. Bar charts displaying the number of detected structural variants by type: DEL (green), DUP (orange), insertion (INS) (blue), INV (pink), and translocation (TRA) (light green). Panel (d) shows P2 platform data; panel (e) shows Mk1c platform data. Data compared across bulk sequencing, ChimeraLM-filtered, and unfiltered WGA samples. (f,g) Composition of chimeric artifact-supported SV. Pie charts showing the proportion of different SV types among events supported specifically by reads classified as chimeric artifacts by ChimeraLM in unfiltered WGA data. Panel (f) shows P2 data; panel (g) shows Mk1c data. Percentages indicate relative frequency of each SV class within the chimeric artifact-supported subset.

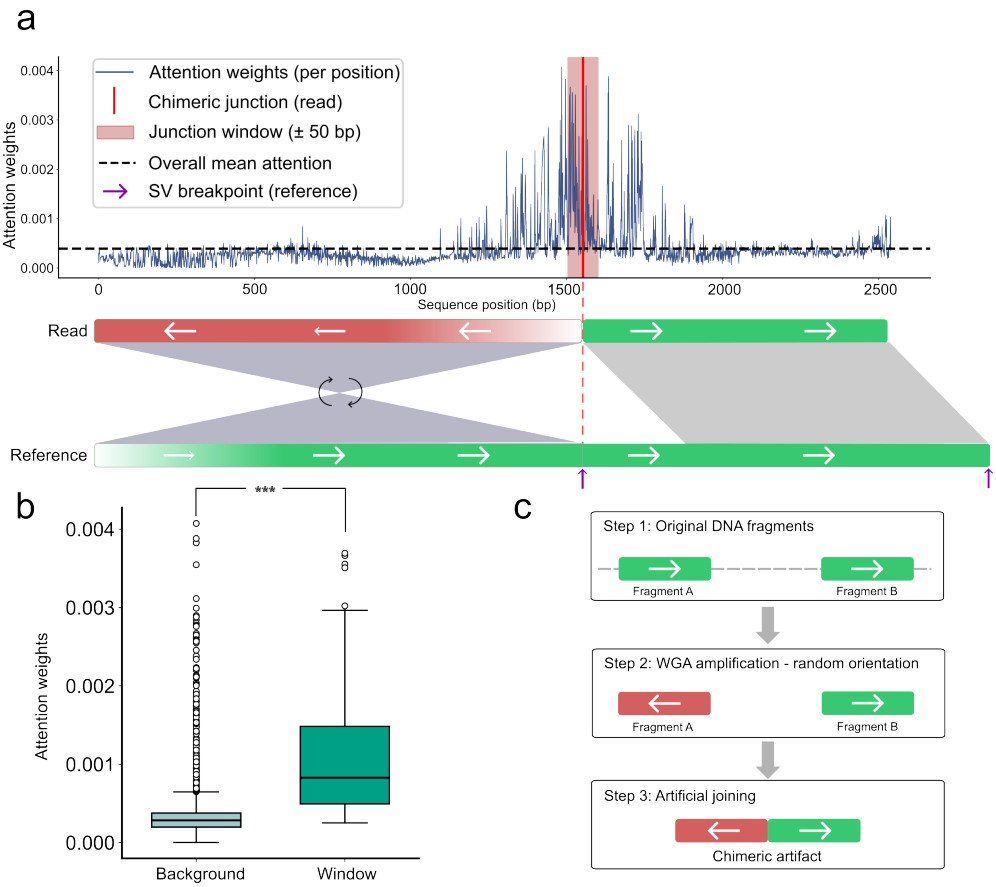
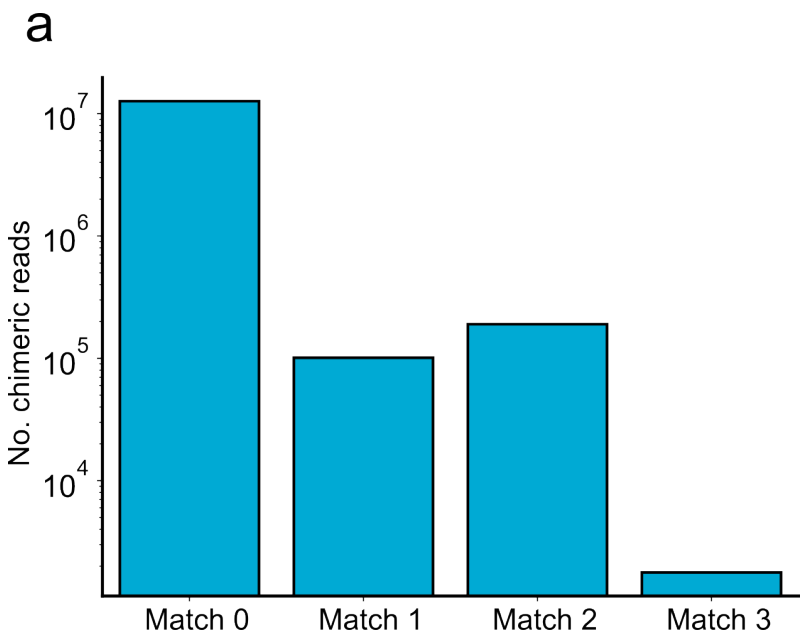
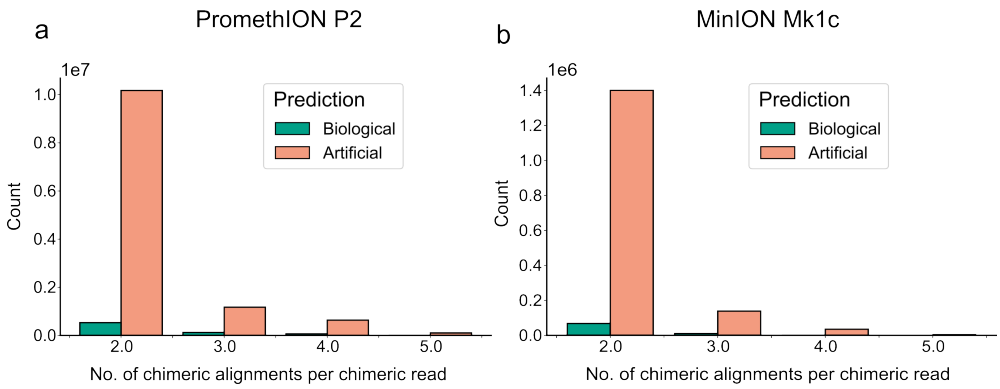


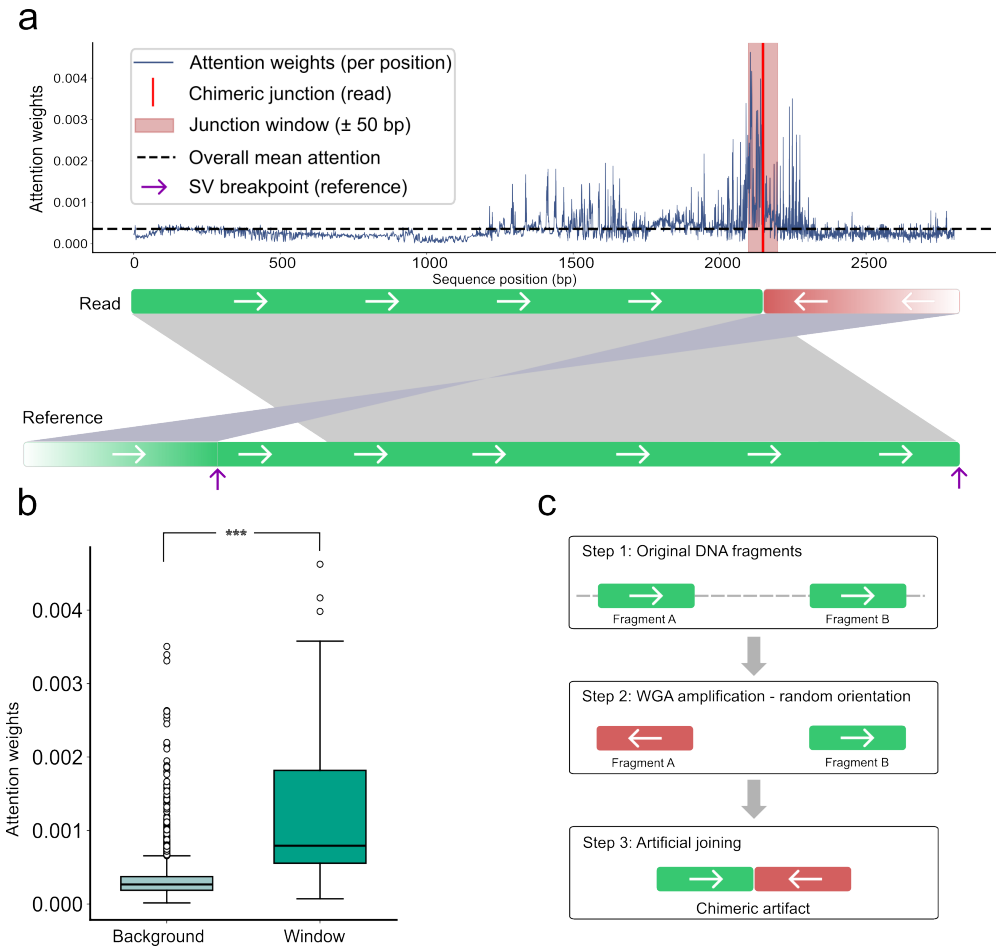
Fig. 4 Attention-based interpretability reveals ChimeraLM's capacity to focus on chimeric junction regions. (a) Attention weight profile across a representative chimeric read. Upper panel shows attention weights per sequence position (blue line) with overall mean attention (dashed line). Red vertical line indicates the chimeric junction position in the read, with pink shading marking the junction window (±50 bp). Purple arrow indicates the corresponding SV breakpoint position in the reference genome. Lower panel illustrates read alignment: the read (top bar) shows reverse-complemented sequence (red with leftward arrows) transitioning to forward sequence (green with rightward arrows) at the junction. Reference genome (bottom bar) shows continuous forward orientation, with gray regions indicating alignment relationships. (b) Quantitative comparison of attention weights between junction window and background regions. Box plots show significantly elevated attention weights in the junction window (median ∼0.0008) compared to background regions (median ∼0.0003), with statistical significance indicated (***, $p < 0.001$, Wilcoxon rank-sum test). (c) Proposed mechanism of chimera formation during WGA. Step 1: Original DNA fragments from distant genomic loci (Fragment A and Fragment B) exist in forward orientation. Step 2: During WGA, Fragment A undergoes random reverse-complementation while Fragment B maintains forward orientation. Step 3: Template switching causes artificial joining of the two fragments, creating a chimeric artifact with discordant orientation patterns.



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, $\sim 10^5$-10^5 reads each). This matching strategy forms the basis for ground truth labeling in supervised training.



Extended Data Fig. 2 Distribution of chimeric alignments per chimeric read stratified by ChimeraLM prediction. (a) PromethION P2 platform chimeric alignment analysis. Bar chart showing the distribution of chimeric reads based on the number of chimeric alignments per read (x-axis: 2, 3, 4+ alignments) and total read count (y-axis, log scale). Bars are colored by ChimeraLM's binary classification: biological (dark teal) and artificial (coral). Analysis includes only reads identified as chimeric (minimum 2 alignments per read). (b) MinION Mk1c platform chimeric alignment analysis. Bar chart showing the distribution of chimeric reads based on the number of chimeric alignments per read (x-axis: 2, 3, 4+ alignments) and total read count (y-axis, log scale). Bars are colored by ChimeraLM's binary classification: biological (dark teal) and artificial (coral). Analysis includes only reads identified as chimeric (minimum 2 alignments per read).



Extended Data Fig. 3 Attention