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 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 WGA-induced chimeric artifacts. We trained ChimeraLM using a novel strategy that leverages paired WGA and bulk sequencing data from the same sample to establish ground truth labels for supervised learning. The model architecture integrates Hyena operators with attention pooling mechanisms optimized for variable-length genomic sequences. ChimeraLM achieved high performance in chimeric detection (F1 score: 0.805, recall: 0.946) and dramatically outperformed existing methods (SACRA, 3rd-ChimeraMiner), reducing chimeric contamination by ∼90% compared to 0% for these established tools. When applied to SV analysis, ChimeraLM improved the validation rate of detected events ∼12-fold (from 0.24% to 2.64%) while preserving 91.5% of genuine SV on PromethION P2 data, with similar performance on MinION Mk1c platform. ChimeraLM processing normalized SV type distributions toward bulk sequencing profiles, eliminating the characteristic false-positive inversion (INV) bias (88-92% of chimeric artifact-supported SV) in unprocessed

WGA data. Attention weight analysis revealed that ChimeraLM can focus on chimeric junction regions in representative examples, demonstrating capacity to learn biologically interpretable features. This improvement enables confident detection of genuine chromosomal instability and copy number evolution in single cells, facilitating applications in cancer biology, developmental biology, and neuroscience that were previously limited by high false-positive rates. This approach addresses a fundamental bottleneck in single-cell genomics and enables more reliable structural variant analysis of individual cells.

Keywords: Whole Genome 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 [**?** **?**]. 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 [**?** **?**].

To overcome this limitation, WGA has become essential for single-cell genomic studies [**?** **?**]. 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 [**?** **?** **?** **?**]. This amplification enables the depth and breadth of coverage necessary for reliable variant calling, copy number analysis, and SV detection. Accurate single-cell genomics is particularly critical for studying somatic mosaicism in development and disease, tracking clonal evolution in cancer, and characterizing rare cell populations—applications where false-positive SVs can lead to incorrect biological conclusions about genomic stability and cellular identity.

Despite its critical role, WGA introduces systematic artifacts that significantly impact downstream analyses [**?** **?**]. Among the most problematic are chimeric sequences—artificial DNA constructs formed when fragments from different genomic loci are erroneously joined during amplification [**?** **?** **?**]. These chimeric artifacts manifest as false-positive SVs that do not exist in the original cell [**?**], 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 [**?** **?**]. However, these methods often fail to distinguish genuine SVs from amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns or occur in repetitive genomic regions [**?** **?**]. This lack of robust artifact detection has limited

the reliability of SV 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 WGA. By leveraging deep learning to capture sequence patterns and contextual information in genomic reads [**? ? ?**], ChimeraLM effectively distinguishes genuine biological sequences from WGA-induced chimeric artifacts. We demonstrate that ChimeraLM achieves superior performance compared to existing methods and substantially improves the reliability of SV detection in single-cell genomic studies.

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 WGA 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 after read mapping and before downstream analyses such as SV detection (Fig. 1a). Following standard quality filtering and read alignment to the reference genome, ChimeraLM evaluates each chimeric 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 SV 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

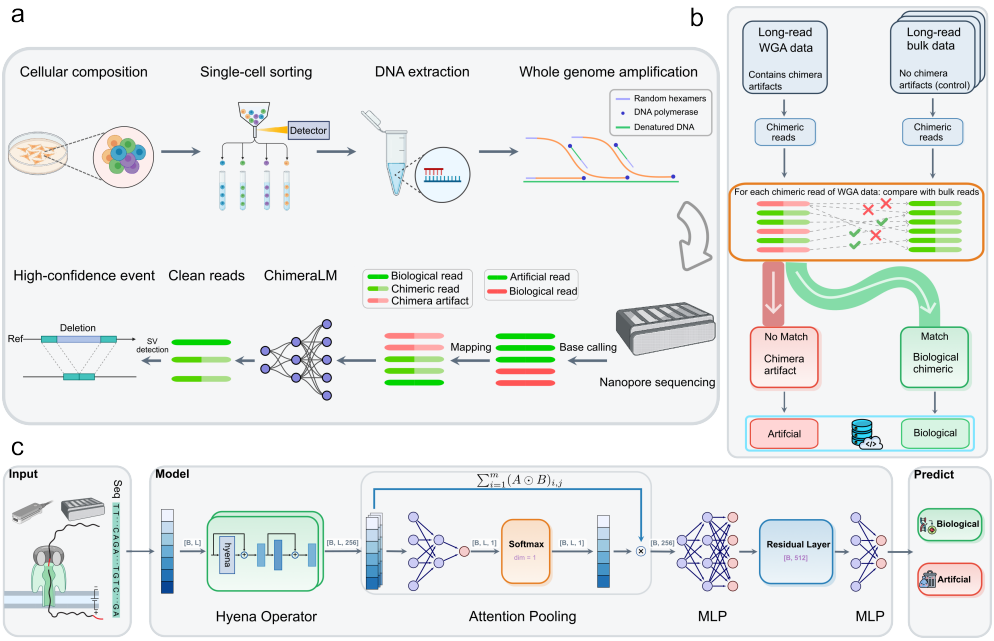


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. Created with BioRender.com. Overview: ChimeraLM removes artificial DNA sequences created during WGA, enabling accurate detection of genuine structural variants in single cells. (a) Integration of ChimeraLM 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.

chimeric artifacts introduced during amplification, bulk sequencing from the same sample contains only genuine biological sequences.

Our ground truth labeling strategy compares each chimeric read of WGA against the bulk sequencing dataset (Fig. 1b, see Methods for details). 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 12,670,396 chimeric reads showed no matches in bulk data and were classified as artificial, while 101,094, 190,309, and 1,777 reads showed 1, 2, and 3 matches respectively and were classified as biological (Extended Data Fig. 1). Each chimeric read of WGA received a binary label based on this comparison. To create a balanced training dataset, we subsampled 293,180 reads from the no-match category as artificial chimeric artifacts and retained all reads with 1, 2, or 3 matches (293,180 total) as biological. Additionally, we subsampled 178,748 reads from the bulk sequencing data as biological controls, yielding a final dataset of 765,108 labeled reads.

We then partitioned this dataset into training (70%), validation (20%), and test (10%) sets using stratified splitting 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 deep learning architecture designed to distinguish biological sequences from chimeric artifacts by analyzing DNA sequences at single-base pair resolution (Fig. 1c). This design choice is critical for detecting chimeric junctions—the breakpoints where disparate genomic regions are artificially joined during WGA—since these junctions often exhibit abrupt changes in sequence characteristics that require nucleotide-level precision to identify.

The architecture consists of three main components: (1) single-nucleotide tokenization following the HyenaDNA framework [**?**], which represents each DNA base individually to preserve complete sequence information; (2) the HyenaDNA backbone, a genomic foundation model pre-trained on diverse DNA sequences that efficiently captures long-range dependencies through Hyena operators [**?**]; and (3) a classification head with attention pooling that weights sequence positions by their relevance to the chimeric/biological distinction (Fig. 1c).

The HyenaDNA backbone provides two key advantages for chimeric artifact detection. First, it processes full-length sequencing reads (up to 32 kb) efficiently through subquadratically-scaled operations, avoiding the computational bottleneck of traditional transformer attention while maintaining the ability to capture long-range sequence dependencies. Second, by initializing with weights pre-trained on diverse genomic sequences, ChimeraLM benefits from learned representations of general DNA sequence patterns, which it then fine-tunes specifically for distinguishing biological

from chimeric sequences. The attention pooling mechanism aggregates information across the entire read length while learning to focus on informative regions, providing interpretability into which sequence features most strongly contribute 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 classification layer applies a softmax function to produce probability scores for the binary classification task. For each read, the model outputs two probabilities corresponding to the "biological" and "artificial chimeric" categories, and the class with the higher probability determines the final prediction (see Methods for details). 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.

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 0.80, which balances precision and recall for the binary classification task. The model exhibited high recall of 0.95, indicating that it successfully identified 95% of true chimeric artifacts in the test dataset, minimizing the risk of retaining false-positive SVs in downstream analyses. The precision of 0.70 demonstrates that 70% 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

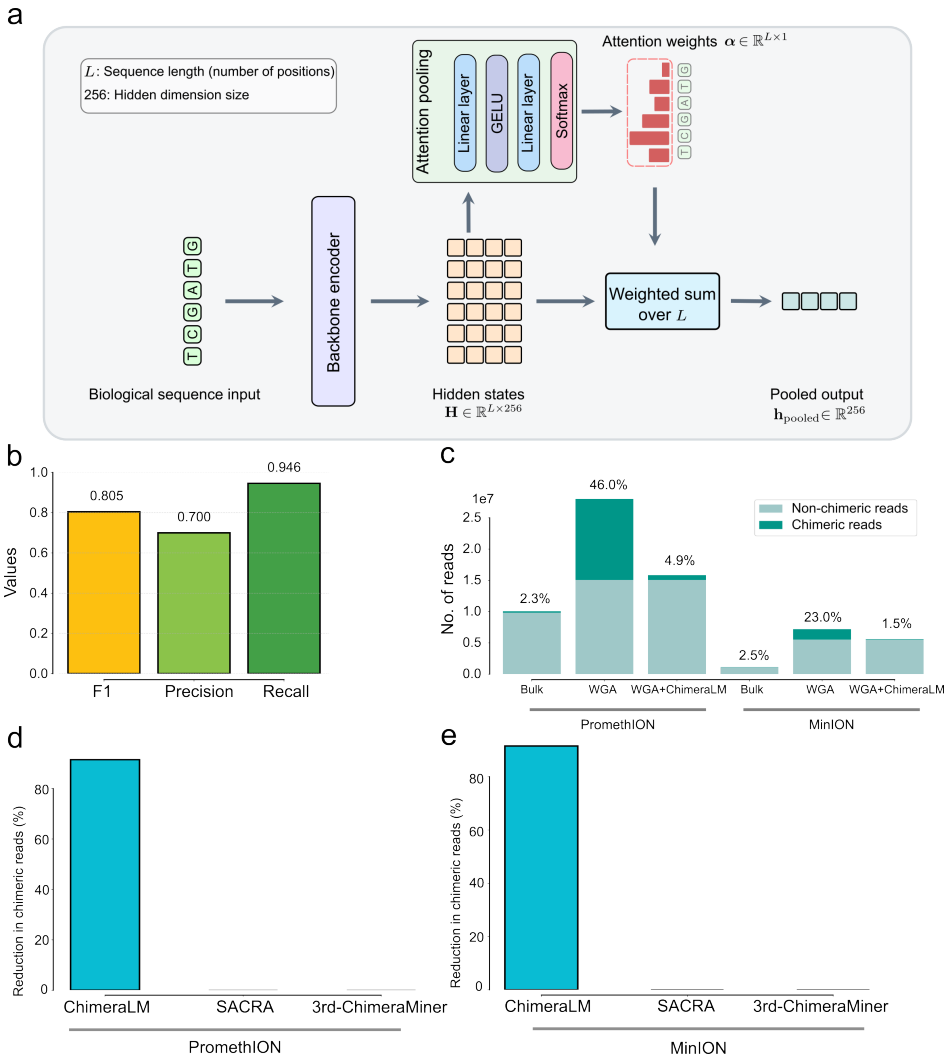


Fig. 2 ChimeraLM performance evaluation and benchmarking against existing methods. (a) ChimeraLM performance metrics on test dataset. Bar chart showing F1 score (0.805), precision (0.700), and recall (0.946) achieved by ChimeraLM on held-out test data for binary classification of biological versus chimeric artifacts. (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 panel shows PromethION P2 platform data with chimeric rates of 2.3% (bulk), 46.0% (WGA), and 4.9% (ChimeraLM). Right panel 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. Bar chart showing percentage reduction in chimeric reads achieved by ChimeraLM (∼92%) compared to existing computational methods SACRA and 3rd-ChimeraMiner (both 0% reduction). (d) Benchmarking on Mk1c platform data. Bar chart showing percentage reduction in chimeric reads achieved by ChimeraLM (∼91%) compared to SACRA and 3rd-ChimeraMiner (both 0% reduction).

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 [**?**] and 3rd-ChimeraMiner [**?**], 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

The ultimate value of chimeric artifact removal lies in improving biological interpretation of structural variation. 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 ≥2 callers per platform and ≥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 ∼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 (∼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 [**? ?**].

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$50 bp) centered on chimeric junctions exhibited significantly higher values compared to background regions. For the representative read shown in Fig. 4 (read ID: 6f4568e5-3543-48c7-a64b-f4010c03804c), the difference was highly significant ($p = 5.3 \times 10^{-14}$, Wilcoxon rank-sum test). Similarly, the supplementary example (Extended Data Fig. 3, read ID: 1c66d41b-de5a-4e3d-b54d-69d41bfc3160) showed comparable statistical significance ($p = 6.8 \times 10^{-15}$, 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 [**? ?**]. 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 [**? ? ?**]. 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 [**? ?**]. 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.

Additionally, ChimeraLM's performance depends on the quality of WGA protocols—extremely degraded samples or novel amplification methods may exhibit artifact patterns not represented in the training data. The model currently requires Graphics Processing Unit (GPU) resources for efficient processing of large datasets, though Central Processing Unit (CPU) inference remains possible for smaller studies. Future work should evaluate performance across diverse WGA chemistries (PicoPLEX, LIANTI, etc.) and validate on additional cell types beyond the PC3 cell line used for training, as amplification biases may vary across different genomic backgrounds and chromatin states. The reliance on supervised learning with labeled training data means that ChimeraLM's effectiveness may be reduced when applied to WGA artifacts that differ substantially from those in the training dataset. Cross-validation on multiple cell lines and WGA protocols will be essential to establish the generalizability of the approach.

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.

Practical recommendations for users

For researchers planning to apply ChimeraLM to single-cell genomic studies, we recommend the following workflow: (1) Generate both WGA and bulk sequencing data from the same sample when possible to enable model training or fine-tuning on sample-specific artifact patterns; (2) Apply ChimeraLM filtering before SV calling rather than post-hoc filtering of variant calls; (3) Use the filtered reads for all downstream analyses, not just SV detection, as chimeric artifacts can also impact Single Nucleotide Polymorphism (SNP) calling, copy number analysis, and other applications; (4) When GPU resources are limited, process data in batches to manage memory constraints; (5) For novel WGA protocols or cell types, validate ChimeraLM performance on a subset of data with orthogonal validation (e.g., FISH, Polymerase Chain Reaction (PCR) confirmation) before large-scale application. These guidelines ensure optimal performance and appropriate application of ChimeraLM across diverse research contexts.

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

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 Oxford Nanopore Ligation Sequencing Kit V14 (SQK-LSK114) and sequenced on MinION Mk1C or PromethION P2 Solo devices with R10.4.1 flow cells according to the manufacturer's genomic DNA workflow. Because all single-cell samples originated from the same monoclonal lineage, observed differences between amplified and bulk data primarily reflect MDA-induced artifacts rather than biological variation, providing a controlled experimental setting for downstream analyses.

Basecalling and read processing

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

Resulting BAM files were sorted and indexed with SAMtools v1.16 [**?**]. 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

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

Biological rationale for deep learning approach

Traditional bioinformatics methods for chimeric detection rely on alignment metrics, coverage patterns, and mate-pair orientation—features that may overlap between genuine structural variants and WGA artifacts. Deep learning enables ChimeraLM to integrate multiple sequence-level features simultaneously, including nucleotide composition, k-mer patterns, and positional context, to distinguish artifacts from biology. The attention mechanism allows the model to focus on chimeric junctions where sequence orientation changes abruptly, a hallmark of template-switching artifacts. This data-driven approach discovers patterns that may not be obvious to rule-based algorithms, particularly for complex multi-fragment chimeras.

ChimeraLM uses a pre-trained HyenaDNA model (hyenadna-small-32k-seqlen) which provides robust genomic sequence representations learned from large-scale genomic data. The backbone model handles DNA sequence tokenization using single base pair resolution with a maximum sequence length of 32,768 base pairs, converting nucleotide sequences into token representations.

Tokenization

The tokenizer employs character-level encoding where each nucleotide is mapped to a unique token ID. For a DNA sequence $S = (s_1, s_2, \ldots, s_L)$ where $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, \mathrm{N}\}$ and L is the sequence length, the tokenization function τ maps each nucleotide to an

integer token:

$$\tau(s_i) = \begin{cases} 7 & \text{if } s_i = \text{A} \\ 8 & \text{if } s_i = \text{C} \\ 9 & \text{if } s_i = \text{G} \\ 10 & \text{if } s_i = \text{T} \\ 11 & \text{if } s_i = \text{N} \\ 6 & \text{otherwise (unknown)} \end{cases}$$

Additional special tokens include [CLS]=0, [SEP]=1, [BOS]=2, [MASK]=3, [PAD]=4, and [RESERVED]=5. The tokenized sequence is represented as $\mathbf{x} = (\tau(s_1), \tau(s_2), \ldots, \tau(s_L))$, with truncation applied when $L > 32{,}768$ and padding tokens appended when $L < 32{,}768$ to enable batch processing.

Backbone encoding

Input sequences are processed through the Hyena operators. For a tokenized input sequence $\mathbf{x} \in \mathbb{Z}^L$, the backbone model generates contextualized hidden representations:

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

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L)$ represents position-wise hidden states with dimension $d_{\text{hidden}} = 256$. 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.

Classification head with attention pooling

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.

The attention pooling mechanism computes position-wise attention weights through a two-layer neural network. For hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$, attention scores are computed as:

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

where the first linear layer projects from dimension 256 to 256, followed by Gaussian Error Linear Unit (GELU) activation, and a second linear layer projects to a single attention score per position.

When attention masks $\mathbf{m} \in \{0,1\}^L$ are provided for variable-length sequences (to handle padding), the attention weights are masked and renormalized:

$$\boldsymbol{\alpha}_{\text{masked}} = \frac{\boldsymbol{\alpha} \odot \mathbf{m}}{\sum_{i=1}^{L}(\alpha_i \cdot m_i) + \epsilon}$$

where $\odot$ denotes element-wise multiplication and $\epsilon = 10^{-9}$ ensures numerical stability.

The pooled representation is computed as the weighted sum of hidden states:

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

This approach accommodates the natural variability in sequencing read lengths without introducing computational inefficiency or artificial sequence information.

Multi-layer perceptron with residual connections

The pooled sequence representation is processed through a MLP with residual connections. For the pooled representation $\mathbf{h}_{\text{pooled}} \in \mathbb{R}^{256}$, the first layer performs dimensionality expansion:

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

For subsequent layers when dimensions match, residual blocks are employed. A residual block $\mathcal{R}$ with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ computes:

$$\begin{aligned}
\mathbf{z}_1 &= \text{Linear}_{512\to512}(\mathbf{f}_{\text{in}}) \\
\mathbf{z}_2 &= \text{Dropout}_{0.1}(\text{GELU}(\mathbf{z}_1)) \\
\mathbf{z}_3 &= \text{Linear}_{512\to512}(\mathbf{z}_2) \\
\mathbf{f}_{\text{out}} &= \text{Dropout}_{0.1}(\mathbf{z}_3) + \mathbf{f}_{\text{in}}
\end{aligned}$$

where the final addition represents the skip connection that adds the input to the transformed output. This residual design enables stable gradient flow during training and improved representation learning for complex genomic patterns.

Output layer and classification

The final classification layer maps the processed features to binary classification logits. For the final feature representation $\mathbf{f}_{\text{final}} \in \mathbb{R}^{512}$ from the MLP, the output logits are computed as:

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

where z_0 represents the logit for the "biological" class and z_1 represents the logit for the "artificial chimeric" class.

The model outputs two logits rather than a single probability, enabling the use of cross-entropy loss during training. During inference, classification predictions are

made by selecting the class with the highest logit:

$$\hat{y} = \underset{i \in \{0,1\}}{\operatorname{argmax}} z_i$$

Equivalently, after applying softmax, the predicted class corresponds to $\hat{y} = 1$ (artificial chimeric) if $z_1 > z_0$, otherwise $\hat{y} = 0$ (biological).

Complete model architecture

The complete ChimeraLM architecture can be summarized as a composition of functions:

$$\hat{y} = \text{argmax}(\text{Linear}_{512 \to 2}(\mathcal{R}(\text{MLP}(\text{AttentionPool}(\text{HyenaDNA}(\tau(S)))))))$$

where τ denotes tokenization, HyenaDNA is the pre-trained backbone encoder, AttentionPool performs attention-based pooling, MLP is the first feedforward layer with expansion, $\mathcal{R}$ represents residual blocks, and the final linear layer produces binary classification logits.

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 (hyenadna-small-32k-seqlen) using the Hugging Face transformers library, loading the pre-configured tokenizer with maximum sequence length of 32,768 bp, automatic truncation, and padding enabled. 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. Sequences longer than the maximum length are truncated, while shorter sequences are padded to enable efficient batch processing.

Model training framework and optimization

Model training was implemented within PyTorch [**?**] and PyTorch Lightning framework [**?**], 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 [**?**], an extension of the Adam optimizer [**?**] that incorporates decoupled weight decay regularization. The optimizer was configured with a learning rate of 1×10^{-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.

Learning rate scheduling

A learning rate scheduler (ReduceLROnPlateau) was employed to dynamically adjust the learning rate during training based on validation loss performance. The scheduler monitors validation loss $\mathcal{L}_{\text{val}}$ and reduces the learning rate by a factor of 0.1 when no improvement is observed for 10 consecutive epochs:

$$\alpha_{t+1} = \begin{cases} 0.1 \cdot \alpha_t & \text{if } \min(\mathcal{L}_{\text{val}}^{t-10}, \ldots, \mathcal{L}_{\text{val}}^{t}) \geq \mathcal{L}_{\text{val}}^{t-11} \\ \alpha_t & \text{otherwise} \end{cases}$$

where α_t denotes the learning rate at epoch t. This adaptive learning rate adjustment enables fine-grained optimization as training progresses, allowing the model to escape plateaus and achieve better convergence.

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 16 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 GPU during training.

The training infrastructure utilized GPU acceleration 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:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \end{aligned}$$

where TP (true positives) represents chimeric reads correctly classified as artificial, TN (true negatives) represents biological reads correctly classified as biological, FP (false positives) represents biological reads incorrectly classified as artificial, and FN (false negatives) represents chimeric reads incorrectly classified as biological. 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 [**?**] 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.

Model inference and application

Inference pipeline

For applying ChimeraLM to new WGA sequencing datasets, the model takes a BAM file as input. Chimeric reads are first identified using the same SA tag-based filtering criteria described in the data processing section: reads must contain the SA tag and must not be unmapped, secondary, or supplementary alignments. Identified chimeric reads are extracted with their sequence information (read ID and DNA sequence). Each read sequence is tokenized using the HyenaDNA tokenizer with single base pair resolution, applying truncation for sequences exceeding 32,768 bp and padding for shorter sequences to enable batch processing. The trained ChimeraLM model processes sequences in batches, generating two logits $[z_0, z_1]$ for each read, where z_0 corresponds to the "biological" class and z_1 corresponds to the "artificial chimeric" class.

Classification decision

The final classification is determined by applying the softmax function to convert logits to probabilities:

$$p_i = \frac{\exp(z_i)}{\exp(z_0) + \exp(z_1)}, \quad i \in \{0, 1\}$$

The predicted class is the one with the higher probability. If $p_1 > p_0$ (i.e., $z_1 > z_0$), the read is classified as "artificial chimeric"; otherwise, it is classified as "biological."

Output filtering

ChimeraLM outputs a filtered BAM file containing only reads classified as "biological," effectively removing chimeric artifacts from the dataset. The filtered BAM file retains all original alignment information and can be directly used for downstream genomic analyses, including structural variant calling. Reads classified as "artificial chimeric" are excluded from the output file.

Test set evaluation

Final model performance was evaluated on the held-out test set that remained completely isolated from the training process. The test set evaluation provides an unbiased estimate of ChimeraLM's generalization performance on previously unseen data. For the binary classification task, true positives (TP) represent chimeric reads correctly classified as artificial, true negatives (TN) represent biological reads correctly classified as biological, false positives (FP) represent biological reads incorrectly classified as artificial, and false negatives (FN) represent chimeric reads incorrectly classified as biological. All metrics (Precision, Recall, F1 score, Accuracy) were computed using the same formulations as described in the Model validation and evaluation section.

SV evaluation

SV calling

Multiple SV calling tools were employed to ensure comprehensive detection of structural variants across different sequencing platforms. For long-read sequencing data (Oxford Nanopore PromethION P2 and MinION Mk1c), we used Sniffles2 [**?** **?**], DeBreak [**?**], SVIM [**?**], and cuteSV [**?**]. For short-read sequencing data (Illumina HiSeq), we used Manta [**?**], DELLY [**?**], and SvABA [**?**]. Each tool was run with default parameters optimized for their respective sequencing technologies.

Construction of gold-standard SV dataset

To construct a high-confidence gold-standard SV dataset, we integrated SV calls from both long-read and short-read sequencing platforms using bulk sequencing data from the PC3 cell line (Fig. 3a). This multi-platform, multi-caller approach ensures robust identification of genuine structural variants while minimizing false positives.

For each sequencing platform, SV events were filtered to retain only those detected by $\geq$2 independent SV callers, reducing platform-specific false positives. Subsequently, SV events supported by $\geq$2 sequencing datasets (requiring support from both long-read and short-read platforms) were retained as high-confidence events. This stringent filtering strategy ensures that the gold-standard dataset contains only structural variants with strong multi-platform evidence, providing a reliable reference for evaluating ChimeraLM's impact on SV detection accuracy.

The resulting gold-standard dataset represents high-confidence structural variants that can be reliably detected across different sequencing technologies and calling algorithms, serving as ground truth for assessing the biological relevance of SV calls in WGA samples before and after ChimeraLM filtering.

Benchmarking against existing methods

ChimeraLM was benchmarked against two existing computational methods for chimeric artifact detection: SACRA [**?**] and 3rd-ChimeraMiner [**?**]. Both tools were run on the same WGA datasets (PromethION P2 and MinION Mk1c) using default parameters as recommended in their respective documentation. Performance was evaluated by comparing the percentage reduction in chimeric reads achieved by each method relative to unprocessed WGA data. Chimeric reads were identified using alignment-based criteria, and reduction rates were calculated as the proportion of chimeric reads successfully removed by each method.

Attention weight analysis

To investigate the interpretability of ChimeraLM's decision-making process, we analyzed attention weights from the attention pooling mechanism for representative chimeric reads. Attention weights indicate the relative importance assigned to each sequence position during classification. For selected chimeric reads, we extracted per-position attention weights and visualized them alongside read alignments to reference genome coordinates.

Chimeric junction positions were identified from alignment data, and a window region of ±50 bp surrounding each junction was defined. Attention weights within the junction window were compared to background regions using the Wilcoxon rank-sum test. Statistical significance was assessed at $p < 0.001$ threshold to identify cases where the model exhibited significantly elevated attention at chimeric junctions.

Statistical analysis

Statistical comparisons were performed using non-parametric tests implemented in Python's SciPy library [**?**]. The Wilcoxon rank-sum test (Mann-Whitney U test) was used to compare attention weight distributions between junction windows and background regions, as attention weights do not follow normal distributions. All statistical tests were two-sided, and p-values less than 0.001 were considered statistically significant.

Figure plotting and visualization

All figures were created using Python with libraries including Matplotlib [**?**] and Seaborn [**?**]. Genomic alignment visualizations were generated using custom scripts to illustrate read-to-reference mappings, orientation patterns, and chimeric junction positions. Statistical plots including box plots, bar charts, and pie charts were generated with standard plotting functions, with colors selected to ensure accessibility and clarity.

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.

Software and reproducibility

All analyses were conducted using Python 3.10 with PyTorch 2.0 [**?**] and PyTorch Lightning 2.0 [**?**]. Key dependencies include NumPy [**?**] for numerical computing, Pandas [**?**] for data manipulation, and Scikit-learn [**?**] for machine learning utilities. The HyenaDNA model [**?**] was obtained from the official repository and adapted for chimeric detection.

Random seeds were fixed (seed=12345) for all stochastic operations including data shuffling, weight initialization, and dropout to ensure reproducibility of results. Model checkpoints were saved at the end of each training epoch, and the best-performing model based on validation metrics was selected for final evaluation.

All custom scripts for data processing, model training, and evaluation are documented with inline comments and structured to facilitate replication of the reported results. Computational workflows were managed using configuration files (Hydra

framework [**?**]) to enable systematic tracking of experimental parameters and ensure reproducible execution across different computing environments.

Supplementary information.

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

Declarations

Author Contributions. YL, QG and RY designed the study. YL and QG performed the analysis. QG performed the experiments. YL 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

CPU Central Processing Unit 13

DEL deletion 3, 9, 10, 27
DUP duplication 3, 9, 10, 27

FACS Fluorescence-activated cell sorting 3

GELU Gaussian Error Linear Unit 17
GLM Genomic Language Model 1, 12
GPU Graphics Processing Unit 13, 20, 21, 24, 25

HPC High Performance Computing 24

INS insertion 9, 10, 27

INV inversion 1, 9, 10, 27

MALBAC Multiple Annealing and Looping-based Amplification Cycles 2
MDA Multiple Displacement Amplification 2
MLP multilayer perceptron 4, 6, 18

PCR Polymerase Chain Reaction 13

SA Supplementary Alignment 15
SNP Single Nucleotide Polymorphism 13
SV Structural Variation 1–4, 6, 8–10, 12, 13, 23, 27, 28, 30

TRA translocation 9, 27

WGA Whole Genome Amplification 1–13, 15, 16, 19, 22, 23, 27–30

References

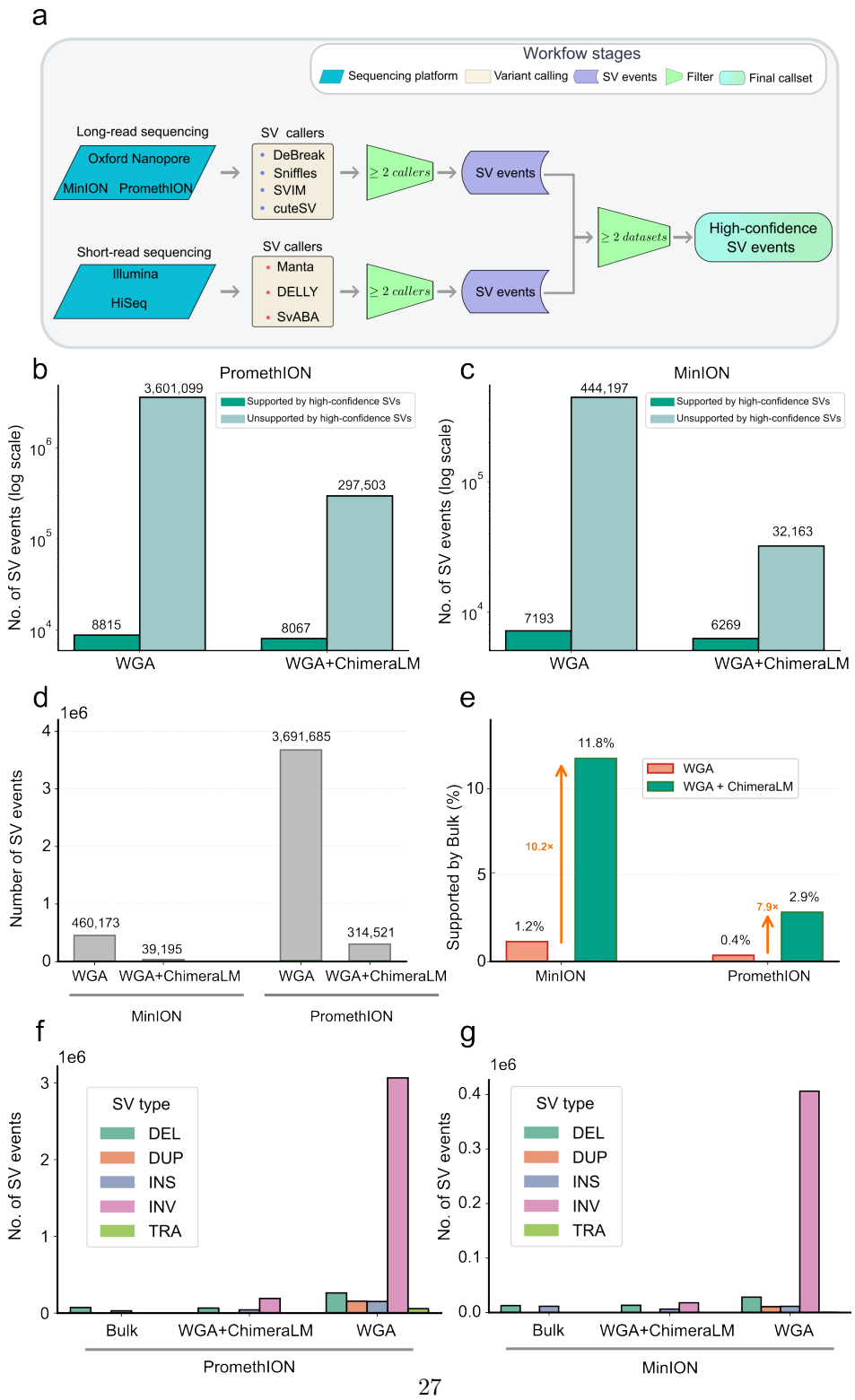


Fig. 3 SV detection accuracy and type distribution analysis. (a) Workflow for constructing gold standard SV dataset from bulk sequencing data. Long-read sequencing (Oxford Nanopore Mk1c and P2) and short-read sequencing (Illumina HiSeq) platforms are used with multiple SV callers. SV events detected 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

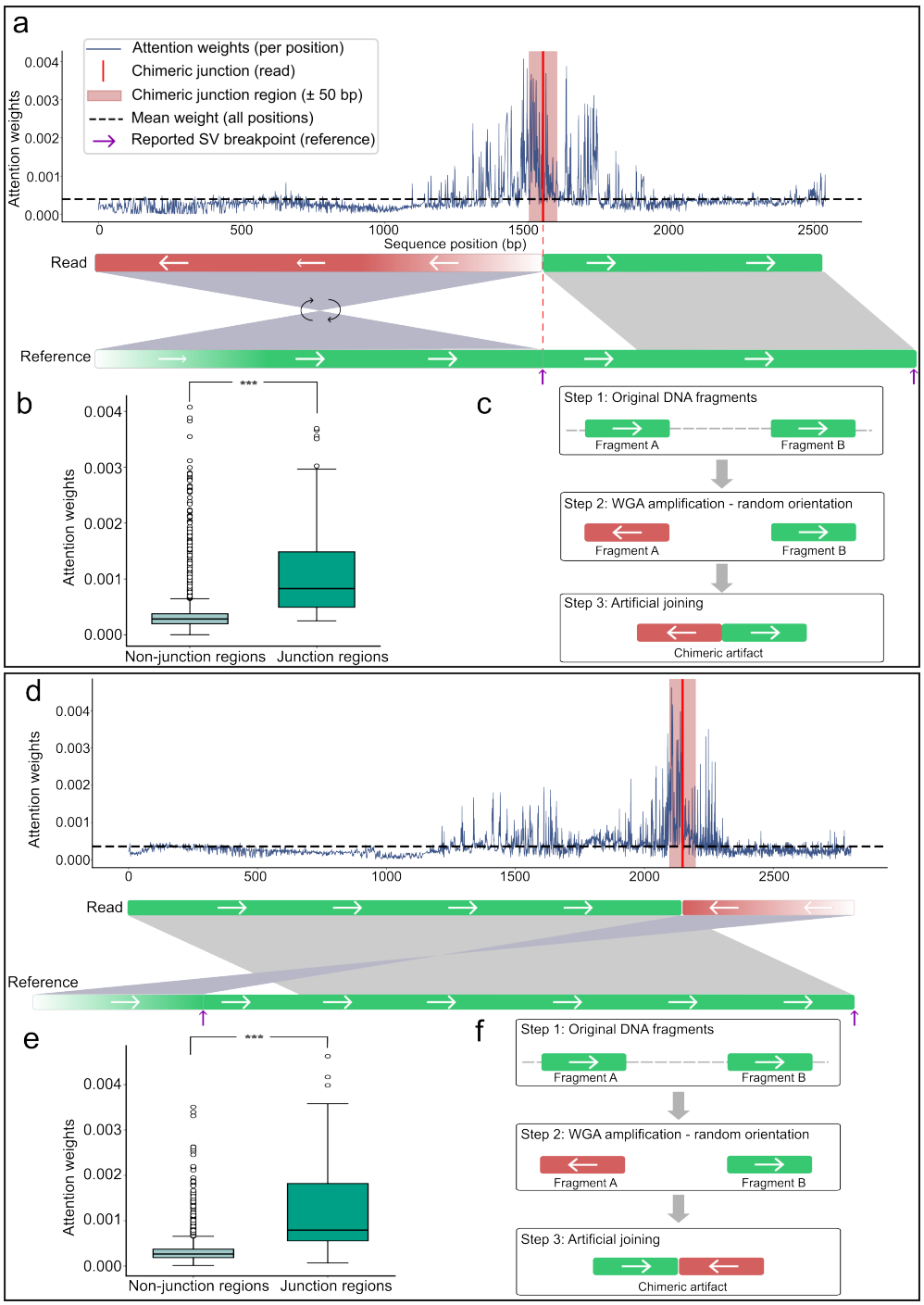
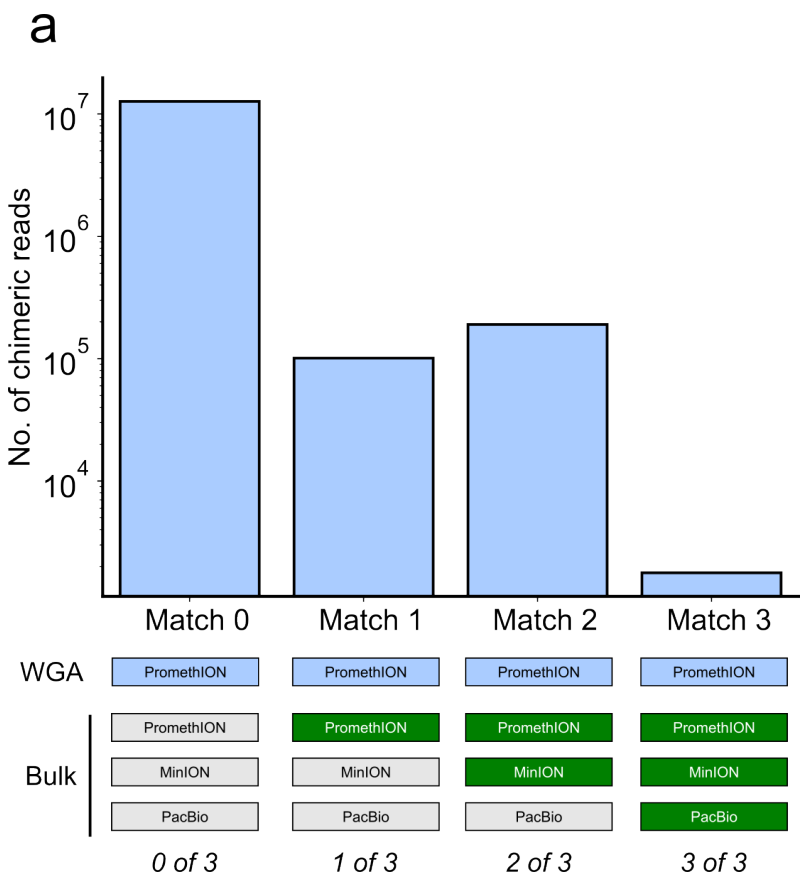
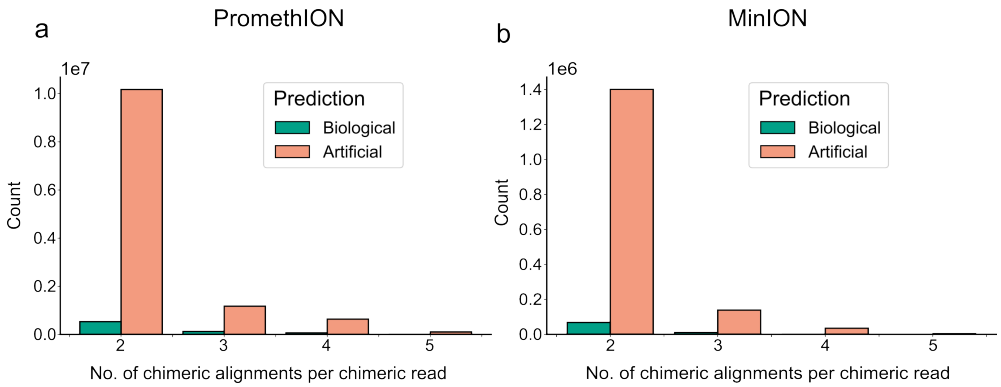


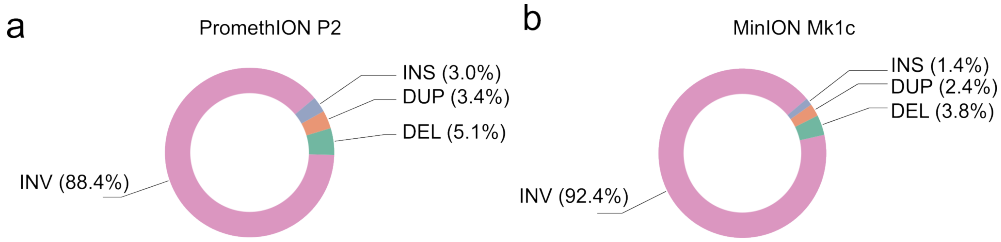
Fig. 4 Attention-based interpretability reveals ChimeraLM's capacity to focus on chimeric junction regions. (a) Attention weight profile across a representative chimeric read (read ID: 6f4568e5-3543-48c7-a64b-f4010c03804c). Upper panel shows attention weights per sequence position (blue line) with overall 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 compared to background regions ($p = 5.3 \times 10^{-14}$, 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, $\sim 10^5$ 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 Additional example of attention-based interpretability showing ChimeraLM's focus on chimeric junction. (a) Attention weight profile across another representative chimeric read (read ID: 1c66d41b-de5a-4e3d-b54d-69d41bfc3160). 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 arrows indicate the corresponding SV breakpoint positions in the reference genome. Lower panel illustrates read alignment: the read (top bar) shows forward sequence (green with rightward arrows) transitioning to reverse-complemented sequence (red with leftward 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 compared to background regions ($p = 6.8 \times 10^{-15}$, Wilcoxon rank-sum test). (c) Chimera formation mechanism illustrated for this example. Step 1: Original DNA fragments from distant genomic loci (Fragment A and Fragment B) exist in forward orientation. Step 2: During WGA, Fragment A maintains forward orientation while Fragment B undergoes random reverse-complementation. Step 3: Template switching causes artificial joining of the two fragments with opposite orientations, creating a chimeric artifact.