

ChimeraLM: a language model enables accurate
structural variant detection in whole-genome
amplified long-read sequencing

Yangyang Li^{1†}, Qingxiang Guo^{1†}, Rendong Yang^{1,2*}

¹Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, 60611, IL, USA.

²Robert 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 chimera artifacts that manifest as false-positive [structural variation \(SV\)](#) and compromise downstream analyses. Here we present ChimeraLM, an interpretable [genomic language model \(GLM\)](#) that identifies [WGA](#)-induced chimera artifacts directly from sequence information. ChimeraLM is trained on matched [WGA](#) and bulk long-read sequencing from the same sample, using bulk support to label chimeric reads as amplification-induced artifacts or genuine genomic events. To capture long-range dependencies in variable-length reads, the model integrates Hyena operators with attention pooling. Evaluated on matched [WGA](#) and bulk nanopore datasets, ChimeraLM eliminated ~90% of chimeric reads from [WGA](#) data, restoring chimeric read proportion to near-bulk levels, whereas existing approaches achieved at most an 8% reduction. Using high-confidence [SV](#) call sets derived from matched bulk data as a reference, ChimeraLM removed 92-93% of unsupported [SV](#) calls from [WGA](#) datasets while retaining 72-92% of bulk-supported [SVs](#). ChimeraLM further normalized [SV](#) type distributions toward bulk profiles by suppressing the characteristic inversion bias observed in unprocessed [WGA](#) data. Attention-based interpretation indicates that ChimeraLM

047 concentrates classification evidence at chimeric junctions, demonstrating its ability
048 to learn biologically interpretable features. ChimeraLM provides a general
049 approach for suppressing amplification-induced artifacts, enabling more reliable
050 single-cell **SV** analysis across long-read platforms.

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

056 Main

058 Single-cell and low-input genomics have transformed our ability to resolve biological
059 heterogeneity, enabling the discovery of rare cell states and the reconstruction of clonal
060 evolution in cancer and development [1–3]. However, the limited DNA input (on the
061 order of picograms per cell) makes comprehensive genome-wide profiling technically
062 challenging [4, 5]. **Whole genome amplification (WGA)** therefore remains a prerequisite
063 for high-coverage sequencing [6–8], yet it introduces systematic errors that compromise
064 genomic fidelity, particularly for **structural variation (SV)** detection [9–11].

065 A major source of error in **WGA** is amplification-induced chimera formation.
066 During this process, highly processive polymerases such as phi29, which is used in **mul-**
067 **tiple displacement amplification (MDA)**, can switch templates and join discontinuous
068 genomic loci into a single molecule [9–13]. As a result, **WGA**-based sequencing often
069 produces chimeric reads that constitute a substantial fraction of the data [9]. These
070 artificial sequences frequently create alignment patterns that closely resemble those
071 generated by genuine **SVs**, including translocations and inversions [10]. Consequently,
072 **SV** callers that rely on alignment-based signals (e.g., split-read and supplementary
073 alignments) and coverage-derived evidence often misinterpret these amplification arti-
074 facts as true rearrangements, inflating false positives and distorting **SV** spectra [14–22].
075 This problem is particularly consequential for **WGA**-based long-read sequencing, which
076 is otherwise well suited for resolving complex **SVs** at sing-cell resolution [23, 24].

077 Distinguishing genuine genomic rearrangements from **WGA**-induced chimera arti-
078 facts remains a major computational challenge. Existing quality-control approaches
079 typically rely on handcrafted rules or alignment-derived features, such as read ori-
080 entation signatures or local coverage deviations [11, 13, 25]. However, these heuristics
081 are often sensitive to platform- and protocol-specific variation. Moreover, they cannot
082 capture sequence-level patterns or long-range dependencies within reads. As a result,
083 low-input long-read sequencing remains difficult to deploy in settings where high preci-
084 sion is essential, including somatic mosaicism profiling [26] and validation of CRISPR
085 off-target effects [27].

086 To address this challenge, we present ChimeraLM, an interpretable **genomic lan-**
087 **guage model (GLM)** for identifying and filtering **WGA**-induced artifacts at the
088 single-read level. Unlike existing approaches that rely on handcrafted rules derived
089 from read alignments or sequence-level [11, 13, 25], ChimeraLM formulates artifact
090 detection as a sequence-modeling task and learns discriminative features directly from
091 raw reads [28]. Building on advances in DNA foundation models [29–32], it captures
092

latent motifs and structural dependencies that generalize across long-read sequencing platforms provided by Oxford Nanopore Technologies (ONT). On ONT WGA datasets, ChimeraLM reduces chimeric reads by ~90% while preserving 72–92% of bulk-supported SVs, improving SV validation rates by 8.5- to 11.0-fold and mitigating WGA-induced biases in SV type distributions. Together, ChimeraLM provides an effective and interpretable filter for WGA long-read data, enabling robust SV discovery in single-cell and low-input genomics.

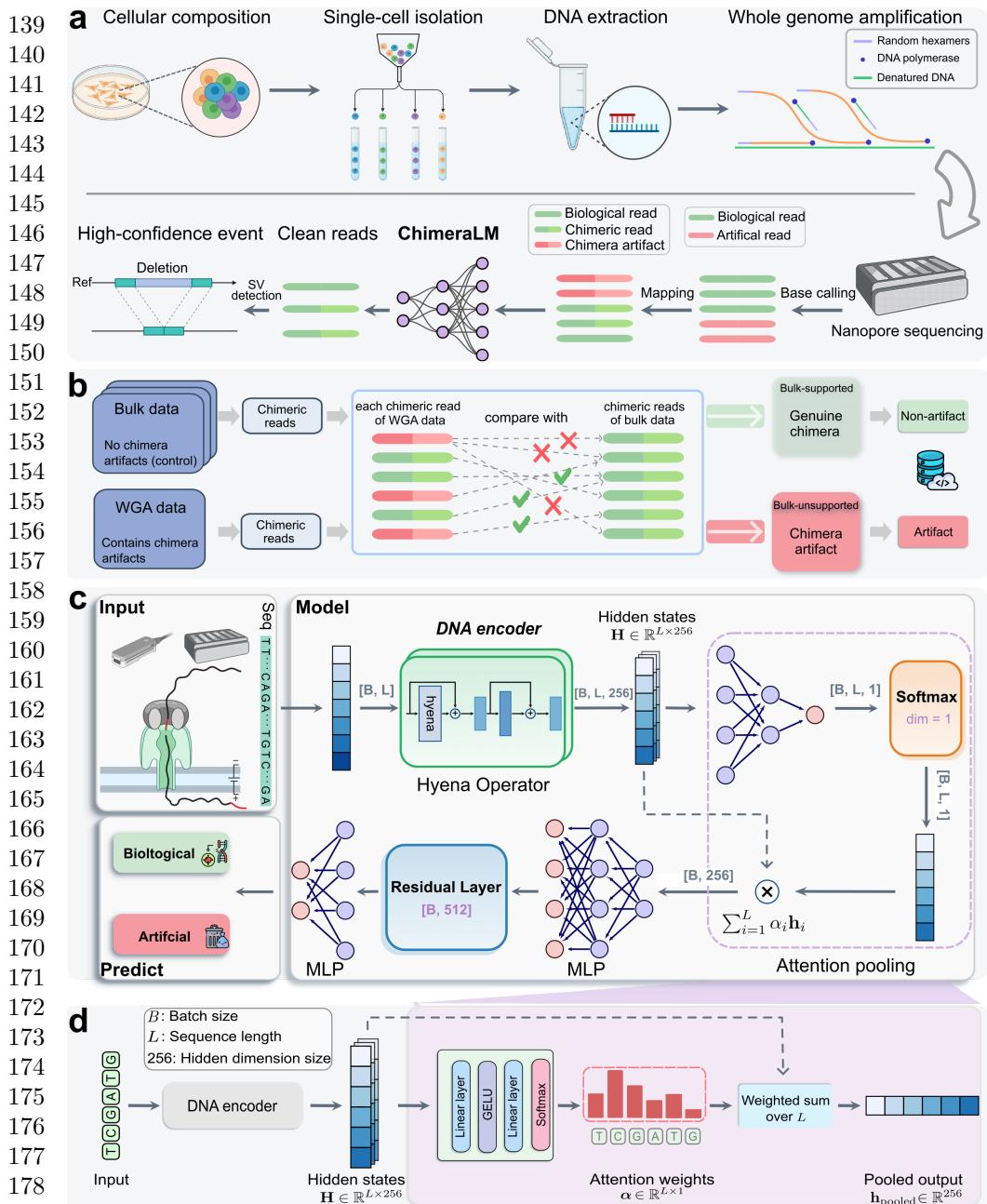
Results

Overview of ChimeraLM workflow and model architecture

ChimeraLM operates as a post-alignment filtering module within the single-cell long-read sequencing pipeline (Fig. 1a). After base calling and alignment to the reference genome, the resulting read set typically contains a mixture of true chimeric reads that arise from authentic genomic rearrangements and artificial chimeras introduced during WGA. ChimeraLM evaluates all reads with chimeric alignments before variant calling. For each read, the model determines whether the observed chimera reflects a genuine genomic event or a WGA-induced artifact. This binary classification enables selective removal of artificial chimeric reads while preserving true rearrangements, which improves the accuracy and sensitivity of downstream SV analyses.

To construct a robust supervised training set, we generated WGA long-read sequencing data from the human prostate cancer cell line PC3 using the ONT PromethION platform. For ground-truth calibration, we acquired three matched bulk long-read datasets from unamplified genomic DNA across diverse technologies: ONT PromethION, ONT MinION, and Pacific Biosciences (PacBio) Sequel II. Leveraging these references, we established a bulk-supported labeling framework by cross-referencing each WGA chimeric read against the chimeric alignment structures identified in the unamplified bulk data (Methods; Fig. 1b; Extended Data Fig. 1a). Under this classification scheme, WGA reads with alignment architectures corroborated by bulk evidence were labeled as genuine genomic events, while those lacking bulk support were classified as WGA-induced artifacts. To evaluate the model’s ability to generalize across sequencing hardware, we generated an independent WGA dataset on the MinION platform, which was reserved exclusively for testing (Extended Data Fig. 1a).

This labeling procedure classified 12,963,576 chimeric reads from the WGA PromethION dataset into two groups (genuine events and WGA-induced artifacts) based on bulk support (Extended Data Fig. 1b). Among these, 12,670,396 reads (97.7%) showed no matching alignment structures in any bulk dataset and were labeled as WGA-induced artifacts. The remaining 293,180 reads (2.3%) had matching structures in at least one bulk dataset, indicating they represent genuine genomic events rather than amplification artifacts, and were labeled as genuine chimeric reads. To construct a balanced training dataset, we retained all 293,180 genuine chimeric reads and randomly subsampled an equal number of WGA-induced artifacts. We further added 178,748 chimeric reads sampled from the bulk datasets to the genuine-event set, expanding the diversity of bulk-supported chimeric alignment structures used for training. The final



181 **Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts.** (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA. During amplification, WGA-induced chimeric artifacts (red) are generated alongside genuine chimeric reads (green). After base calling and mapping, ChimeraLM classifies reads with chimeric alignments as genuine events or WGA-induced artifacts, enabling downstream SV detection on filtered data. (b) Bulk-supported label generation. Chimeric reads from WGA data are compared against bulk sequencing from the same cell line. Reads with bulk-supported alignment structures are labeled as genuine events (green); reads with no bulk match are labeled as WGA-induced artifacts (red). (c) ChimeraLM architecture. Input DNA sequences (batch size B , sequence length L) are tokenized at single-nucleotide resolution and encoded into hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$ through DNA encoder (HyenaDNA [29]). Hyena operators capture long-range dependencies. Attention pooling aggregates position-specific features, and multilayer perceptron (MLP) layers with residual connections process pooled representations for binary classification of genuine events and WGA-induced artifacts. (d) Attention pooling mechanism. Attention weights $\alpha \in \mathbb{R}^{L \times 1}$ are computed through linear layers with GELU activation and softmax normalization, assigning importance scores to each position. The weighted sum produces a fixed-dimensional representation $\mathbf{h}_{\text{pooled}} \in \mathbb{R}^{256}$. Created with BioRender.com.

labeled dataset comprised 765,108 reads and was split into training (70%), validation (20%), and test (10%) sets using stratified sampling (Extended Data Fig. 1a).

To model these labeled reads, ChimeraLM must process long, variable-length DNA sequences at single-nucleotide resolution (Fig. 1c). We therefore built ChimeraLM on HyenaDNA [29], a genomic foundation model pre-trained on diverse DNA sequences. Each read is tokenized at nucleotide resolution and encoded by Hyena operators [33], which capture long-range sequence context without splitting the input. The encoder produces a sequence of hidden states across the full read. To obtain a fixed-length representation for classification, ChimeraLM uses an attention-pooling module that learns position-specific weights and computes a weighted sum over the hidden states (Fig. 1d). The pooled representation is then passed through residual MLP blocks, and a final softmax outputs the probability that a read reflects a genuine event versus a WGA-induced artifact.

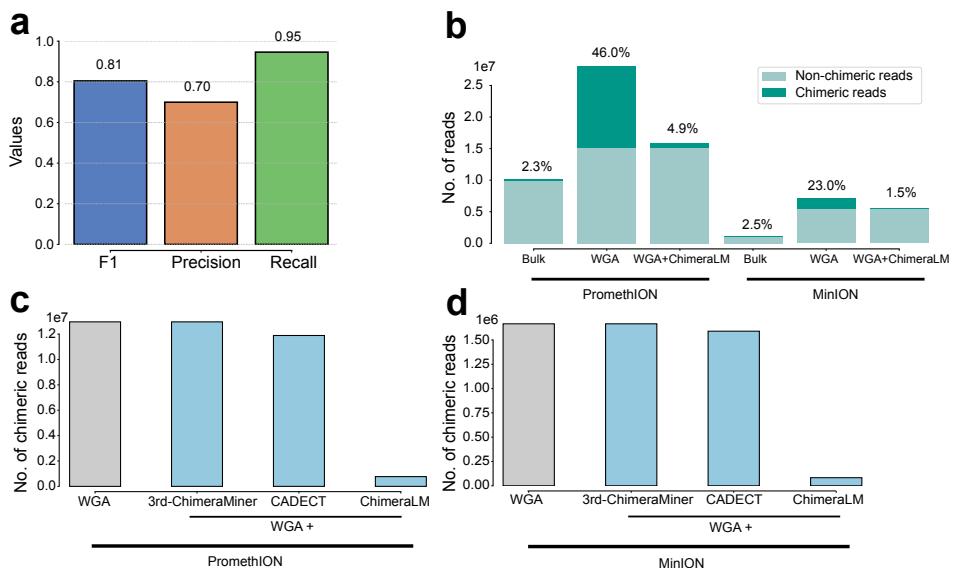


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves recall of 0.95, precision of 0.70, and F1 score of 0.81. (b) Chimeric read reduction across sequencing platforms. Stacked bars show proportions of chimeric (dark teal) and non-chimeric (light teal) reads in bulk, WGA, and ChimeraLM-filtered samples. ChimeraLM reduces chimeric read frequencies from 46.0% to 4.9% (PromethION) and from 23.0% to 1.5% (MinION), approaching bulk levels (2.3% and 2.5%, respectively). (c,d) Benchmarking against existing methods on PromethION (c) and MinION (d). The gray bar indicates the total number of chimeric read on unfiltered WGA data. The blue bar represents the total number of chimeric reads remaining after filtering by each method. ChimeraLM achieves approximately 90% reduction in chimeric reads on both platforms, while 3rd-ChimeraMiner shows no detectable reduction and CADECT shows 8.3% and 4.6% reduction on PromethION and MinION, respectively. SACRA failed to complete due to memory exhaustion (> 500 GB RAM required).

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

233 We first benchmarked ChimeraLM using the held-out test split (10%) derived from
234 the bulk-supported labeled dataset (Extended Data Fig. 1a). This test set consists of
235 chimeric reads labeled as genuine events or **WGA**-induced artifacts based on whether
236 their chimeric alignment structures are supported by matched bulk sequencing.
237 ChimeraLM achieved an F1 score of 0.81, with 0.95 recall and 0.70 precision (Fig. 2a).
238 The high recall indicates that most **WGA**-induced artifacts are correctly identified for
239 removal, which is important for limiting downstream false-positive **SV** calls, while the
240 precision confirms that most reads flagged as artifacts are true amplification-induced
241 chimeras rather than genuine genomic events.
242

243 We next examined whether applying ChimeraLM filtering would restore chimeric
244 read rates in full PC3 **WGA** datasets toward the levels observed in bulk sequencing
245 across both PromethION and MinION platforms (Fig. 2b). Bulk sequencing estab-
246 lished low baseline chimeric read rates of 2.3% (PromethION) and 2.5% (MinION).
247 In contrast, **WGA** increased the chimeric read fraction to 46.0% and 23.0%, respec-
248 tively. After ChimeraLM filtering, chimeric content dropped to 4.9% on PromethION
249 and 1.5% on MinION, corresponding to 10- to 15-fold reductions, while retaining 15.8
250 million and 5.6 million reads. These post-filtering rates approach the corresponding
251 bulk baselines, indicating effective removal of **WGA**-induced artifacts while preserving
252 genuine signal.

253 We compared ChimeraLM against SACRA [25], 3rd-ChimeraMiner [13], and
254 CADECT [11], existing tools for detecting amplification-induced chimeras (Fig. 2c,d).
255 ChimeraLM reduced chimeric reads by ~90% on both platforms, outperforming
256 CADECT (8.3% and 4.6% reduction on PromethION and MinION, respectively),
257 while 3rd-ChimeraMiner showed no detectable reduction. SACRA could not be
258 evaluated due to out-of-memory errors even with 500 GB RAM.

259 The MinION results further provide an independent test of model generaliza-
260 tion, as this MinION **WGA** dataset was not used during training. ChimeraLM was
261 trained exclusively on PromethION **WGA** data, yet achieved comparable chimeric read
262 reduction on MinION. This cross-platform generalization indicates that ChimeraLM
263 captures sequence-level features intrinsic to **WGA**-induced artifacts rather than
264 platform-specific signatures, supporting its potential applicability to additional long-
265 read and potentially short-read sequencing technologies.

266
267
268
269
270
271
272
273
274
275
276

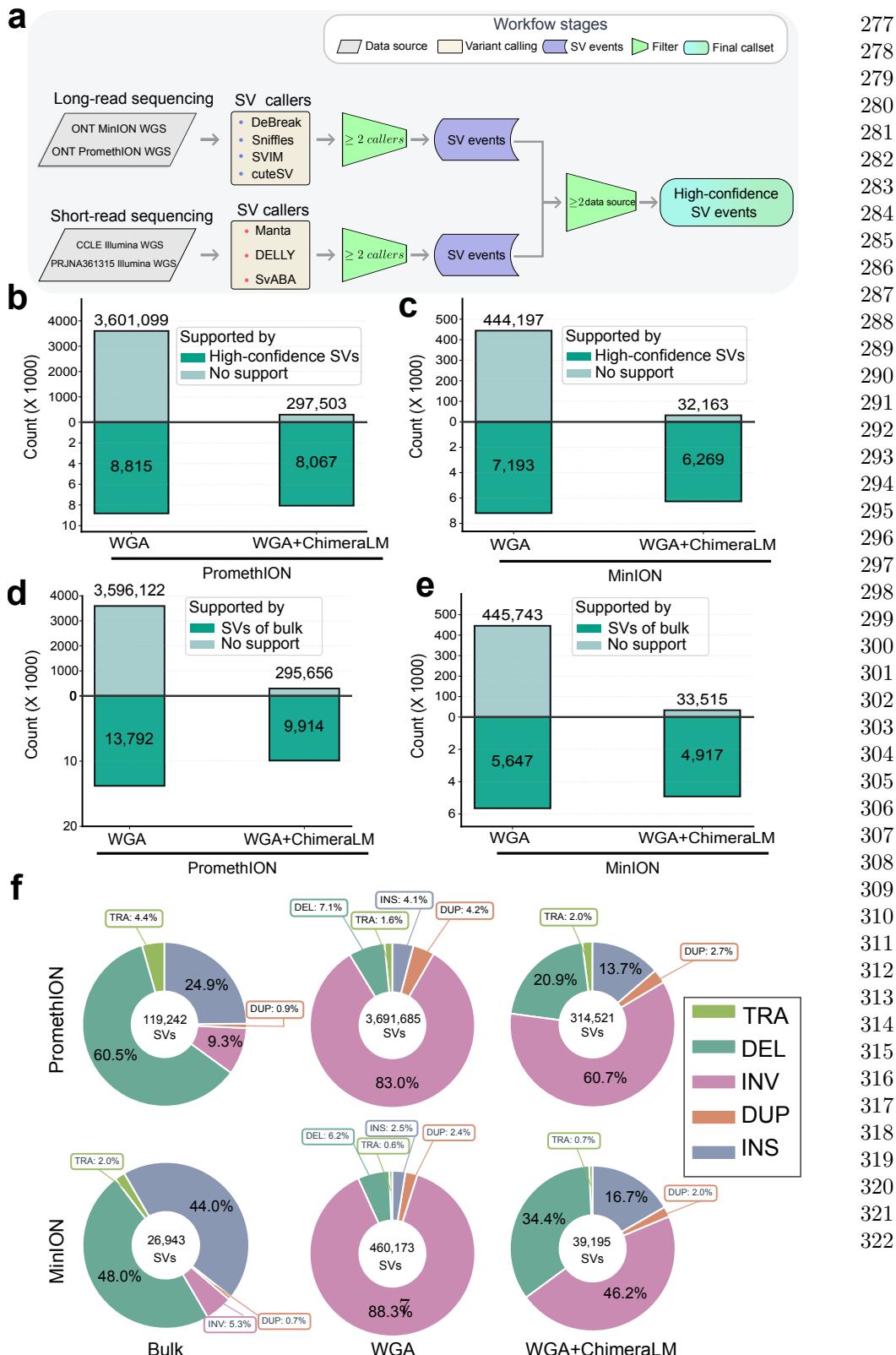


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of a high-confidence SV reference dataset from bulk PC3 sequencing. Four bulk datasets were integrated: ONT MinION Mk1C, ONT PromethION P2, the CCLE Illumina whole genome sequencing (WGS) dataset, and the PRJNA361315 Illumina WGS dataset. SVs were called independently within each dataset using multiple callers, and events supported by ≥ 2 callers per dataset were retained. SVs were then compared across datasets, and events observed in ≥ 2 of the four bulk datasets were designated as gold-standard SVs. (b,c) SV validation against the gold-standard reference for PromethION (b) and MinION (c). Bars show SV calls supported by the gold standard (dark teal) or unsupported (light teal). (d,e) SV validation against platform-matched long-read bulk sequencing for PromethION (d) and MinION (e), capturing true long-read SVs that may not be represented in the multi-platform reference. Bars show SV calls supported by the platform-matched long-read bulk data (dark teal) or unsupported (light teal). (f) Distribution of SV types for PromethION and MinION datasets across bulk samples, unfiltered WGA, and WGA with ChimeraLM filtering. Unfiltered WGA exhibits an excess of inversion (INV) events, which is substantially reduced following ChimeraLM filtering. Other SV classes, including translocation (TRA), deletion (DEL), duplication (DUP), and insertion (INS), are also shown.

323 **ChimeraLM reduces unsupported structural variant calls**

324 Accurate **SV** detection from single cells is essential for characterizing genomic diversity
325 and disease mechanisms. However, **WGA**-induced chimeric artifacts can be misinter-
326 preted as genuine **SVs**, potentially leading to incorrect biological inferences. To assess
327 the impact of ChimeraLM on **SV** calling, we compared **SV** callsets generated from
328 unfiltered **WGA** reads with those obtained after applying ChimeraLM filtering (**WGA**
329 + ChimeraLM). Both callsets were evaluated against two complementary bulk-derived
330 references: (i) a stringent gold-standard **SV** set derived from bulk PC3 DNA through
331 cross-dataset consensus, and (ii) platform-matched long-read bulk **SV** callsets used as
332 platform-specific references.

333 We first constructed a high-confidence gold-standard **SV** set from bulk PC3
334 DNA by integrating four independent sequencing datasets: **ONT** PromethION, **ONT**
335 **MinION**, and two Illumina short-read **WGS** datasets from the Cancer Cell Line
336 Encyclopedia (CCLE Illumina **WGS** [34]) and from a previously published study
337 (PRJNA361315 Illumina **WGS** [35]) (Fig. 3a; Extended Data Table 1). **SVs** were called
338 separately within each dataset using multiple **SV** callers. Events supported by at least
339 two callers within a dataset were retained, and only **SVs** observed in at least two of
340 the four datasets were designated as gold-standard events.
341

342 Relative to this stringent gold standard, unfiltered **WGA** produced a large number
343 of unsupported **SVs**. On PromethION, **WGA** yielded 3,601,099 **SV** calls, of which
344 only 8,815 (0.24%) overlapped gold-standard events. After ChimeraLM filtering, total
345 calls decreased to 305,570 (a 91.5% reduction) while retaining 8,067 gold-standard
346 events (91.5% retention), increasing the validation rate to 2.64% (11-fold) (Fig. 3b).
347 On MinION, total calls decreased from 451,390 to 38,432 (a 91.5% reduction), while
348 gold-standard-supported events decreased from 7,193 to 6,269, corresponding to 87.2%
349 retention. The validation rate increased from 1.59% to 16.3% (10-fold) (Fig. 3c).

350 Because the gold standard is intentionally stringent and may exclude true **SVs**
351 detectable only in long-read data, we next performed platform-matched validation
352 using long-read bulk sequencing from the same platform (Fig. 3d,e). This analysis
353 provides a platform-specific estimate of recall and reduces bias introduced by the
354 strict gold-standard definition. ChimeraLM increased validation rates from 0.38% to
355 3.24% on PromethION (8.5-fold) and from 1.25% to 12.79% on MinION (10-fold),
356 while retaining 71.9% and 87.1% of bulk-supported events, respectively. Together,
357 these results show that ChimeraLM removes the vast majority of unsupported **SV**
358 calls while preserving the majority of bulk-supported variants across platforms.
359

360 **ChimeraLM mitigates WGA-induced SV type biases**

361 Amplification artifacts can profoundly distort the apparent spectrum of **SVs**, com-
362 plicating biological interpretation. To quantify these effects, we compared **SV**-type
363 compositions across bulk, unfiltered **WGA**, and ChimeraLM-filtered datasets gener-
364 ated on both **ONT** PromethION and MinION platforms (Fig. 3f). Bulk datasets
365 displayed **SV** profiles dominated by deletions and insertions, whereas **WGA** datasets
366 exhibited a striking inflation of inversions, reaching 83.0% on PromethION and
367 88.3% on MinION. This inversion-heavy signature mirrors previous observations from
368

single-cell long-read sequencing using PacBio and droplet-based MDA [36], suggesting that many of these events arise from amplification-induced chimeras rather than genuine genomic rearrangements.	369
	370
	371
To directly assess which SV types are preferentially associated with amplification artifacts, we examined SV calls supported exclusively by reads classified as WGA-derived chimeras. These artifact-supported events were overwhelmingly enriched for inversions, comprising 88.4% on PromethION and 92.4% on MinION (Extended Data Fig. 2). Smaller fractions of deletions, duplications, and insertions were also present, indicating that WGA artifacts inflate not only inversion calls but generate false positives across multiple SV categories.	372
	373
	374
	375
	376
	377
	378
Applying ChimeraLM substantially reduced these biases. The inversion fraction decreased from 83.0% to 60.7% on PromethION and from 88.3% to 46.2% on MinION, while the relative abundance of deletions and insertions increased toward bulk-like levels (Fig. 3f). By selectively removing artifact-supported events, ChimeraLM suppresses the dominant inversion inflation and reduces spurious calls across SV types, shifting SV-type distributions toward bulk profiles and improving the accuracy and interpretability of single-cell SV analyses [36].	379
	380
	381
	382
	383
	384
	385
	386
	387
	388
	389
	390
	391
	392
	393
	394
	395
	396
	397
	398
	399
	400
	401
	402
	403
	404
	405
	406
	407
	408
	409
	410
	411
	412
	413
	414

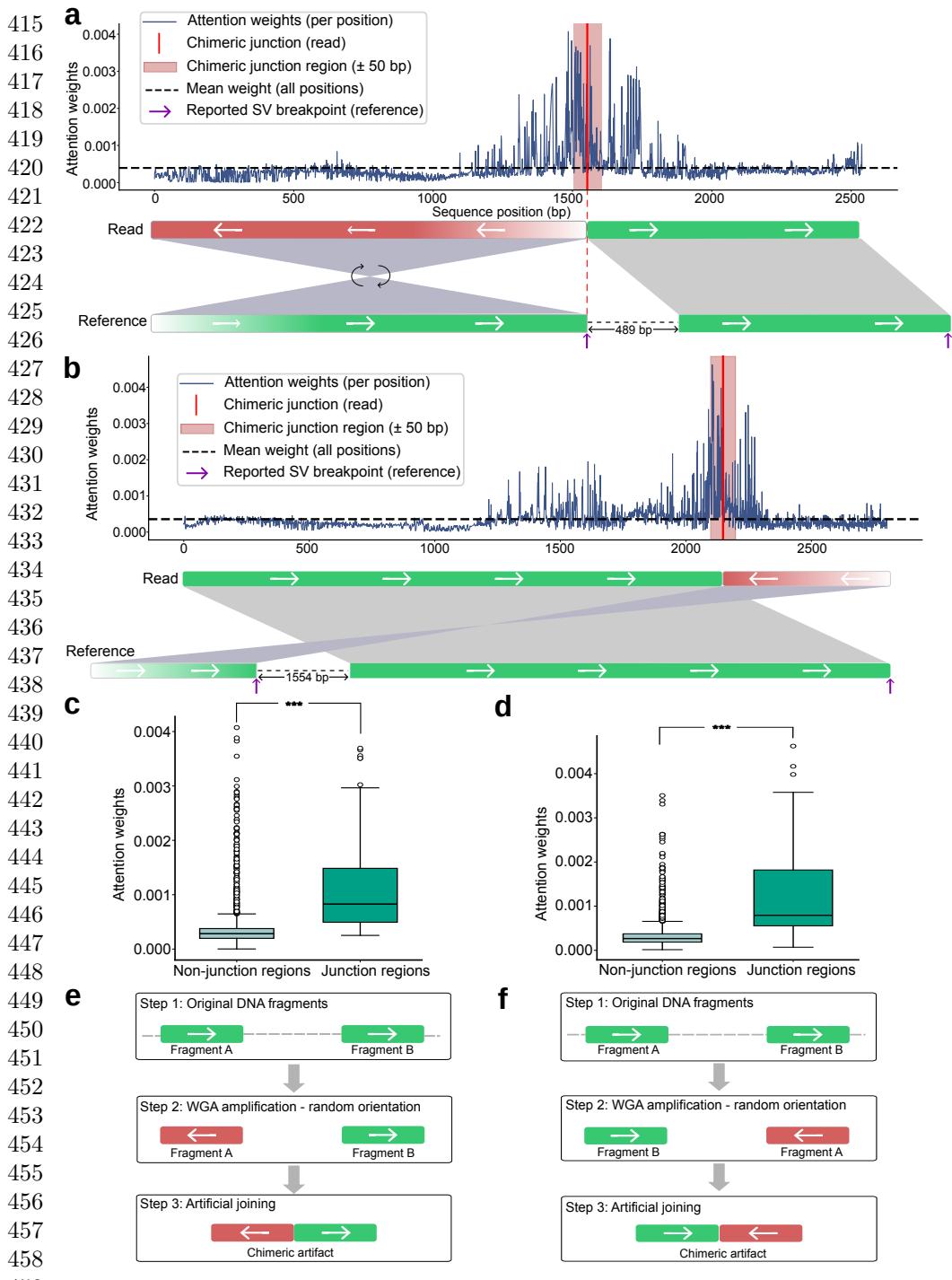


Fig. 4 ChimeraLM attention weights are enriched at chimeric junction regions.

(a,b) Attention weight profiles for two representative chimeric reads exhibiting distinct junction configurations. Upper panels show per-position attention weights (blue) with the mean attention across the read indicated by a dashed line. Red vertical lines mark inferred chimeric junction positions, and pink shading denotes the junction-centered region (± 50 bp). Lower panels display read-level alignments, highlighting orientation transitions at the junctions (green, forward orientation; red, reverse-complemented orientation). (c,d) Quantitative comparison of attention weights between junction and non-junction regions. Junction-centered windows show significantly elevated attention weights relative to non-junction regions ($P = 5.3 \times 10^{-14}$ and $P = 6.8 \times 10^{-15}$; Wilcoxon rank-sum test). (e,f) Schematic illustration of WGA-induced chimera formation. During amplification, DNA fragments originating from distant genomic loci can be amplified in either orientation, joining them into a single molecule with discordant orientations, producing inversion-like alignment signatures. The two examples illustrate forward-to-reverse and reverse-to-forward orientation transitions.

Attention visualization reveals interpretable classification features	461
	462
	463
	464
	465
	466
	467
	468
	469
	470
	471
	472
	473
	474
	475
	476
	477
	478
	479
	480
	481
	482
	483
	484
	485
	486
	487
	488
	489
	490
	491
	492
	493
	494
	495
	496
	497
	498
	499
	500
	501
	502
	503
	504
	505
	506
WGA enables genomic analysis from single cells and other low-input samples, but it also introduces chimeric artifacts that compromise SV detection. ChimeraLM addresses this challenge by classifying chimeric reads as genuine events or WGA -induced artifacts directly from read information, and removing artifacts before variant calling, rather than attempting to correct artifact-driven calls post hoc. Across ONT platforms, ChimeraLM improved data quality at the read and variant levels: it reduced chimeric reads by ~90% while retaining 72–92% of bulk-supported SVs , and it increased the ratio of supported SV calls by 8.5–11.0-fold. Notably, performance generalized from PromethION (used for training) to MinION without platform-specific retraining, suggesting that ChimeraLM captures properties shared by WGA -induced artifacts rather than instrument-specific signatures.	486
In comparison, existing methods showed limited effectiveness on our ONT WGA datasets. Two tools originally developed and primarily evaluated on PacBio data (SACRA and 3rd-ChimeraMiner) [13, 25] either failed to complete under our evaluation setting (SACRA, > 500 GB RAM) or showed no detectable reduction (3rd-ChimeraMiner), highlighting poor cross-platform generalization. CADECT [11], which was designed for ONT data, achieved only 8.3% and 4.6% reduction on PromethION and MinION, respectively. CADECT detects concatemers via sliding-window self-alignment, an effective strategy for repeat-like structures with internal sequence similarity, but it is not designed to capture the broader diversity of WGA -induced chimeras. Together, these comparisons suggest that rule-based or subtype-specific	487

507 heuristics do not comprehensively address amplification-induced chimeras and moti-
508 vate learning-based models that can extract discriminative sequence signatures
509 without imposing predefined structural assumptions.

510 ChimeraLM also illustrates the value of deep learning for quality-control problems
511 where conventional alignment- and coverage-derived criteria provide limited resolu-
512 tion. [11, 13, 25, 28]. By learning directly from sequence, ChimeraLM discovers subtle
513 compositional and structural features that separate genuine events from amplification
514 artifacts. The model also offers interpretability through attention visualization: atten-
515 tention weights concentrate at junction regions where template switching joins discordant
516 loci, validating the biological relevance of the learned features. These methodological
517 advances have direct implications for single-cell genomics, where high false-positive
518 rates in WGA data have constrained robust characterization of chromosomal instabil-
519 ity, clonal evolution, and SV burden [20, 22, 37]. By improving the signal-to-noise ratio
520 and clarifying SV-type spectra that are otherwise distorted by amplification artifacts,
521 ChimeraLM enables more confident identification of genuine SVs, supporting studies
522 of cancer evolution, developmental biology, and somatic mosaicism where single-cell
523 resolution is essential [26, 27].

524 Several limitations warrant consideration. First, the current model processes reads
525 independently; integrating contextual features such as coverage or phasing informa-
526 tion may further enhance accuracy. Second, regarding computational resources, while
527 central processing unit (CPU) inference is feasible, graphics processing unit (GPU)
528 acceleration is recommended for processing large-scale datasets. Finally, future work
529 should extend validation to diverse cell types, sequencing platforms (e.g., PacBio
530 HiFi), and alternative WGA protocols, including multiple annealing and looping-based
531 amplification cycles (MALBAC) [38], linear amplification via transposon insertion
532 (LIANTI) [5], primary template-directed amplification (PTA) [19], and droplet-based
533 MDA (dMDA) [39]. While the sequence-based approach suggests broad applicability,
534 systematic validation across amplification chemistries is needed to assess generalization
535 limits and optimize performance for specific protocols.

536 More broadly, ChimeraLM illustrates the potential of GLMs for data qual-
537 ity control. As long-context architectures continue to advance [29], extending the
538 model’s context window to handle megabase-scale inputs could enable artifact detec-
539 tion in more complex genomic structures. This framework could extend to other
540 amplification-dependent technologies, such as cell-free DNA analysis, ancient DNA
541 studies, and low-biomass metagenomics. Furthermore, attention-based interpretability
542 opens opportunities for studying template-switching dynamics, potentially guiding the
543 development of improved amplification protocols. In summary, ChimeraLM provides
544 a practical and interpretable framework for enhancing long-read single-cell genomic
545 fidelity, ensuring that downstream biological insights are derived from genuine SVs
546 rather than technical artifacts.

547
548
549
550
551
552

Methods	553
Cell culture, single-clone preparation, and nanopore sequencing	554
<i>Cell culture and single-clone establishment</i>	555
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.	556
<i>DNA extraction and whole-genome amplification</i>	557
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.	558
<i>Nanopore library preparation and sequencing</i>	559
Libraries were prepared using the ONT Ligation Sequencing Kit V14 (SQK-LSK114) and sequenced on MinION Mk1C or PromethION P2 Solo devices with R10.4.1 flow cells following the manufacturer's genomic DNA workflow. Because all single-cell samples originated from the same monoclonal lineage, differences between amplified and bulk datasets primarily reflect MDA-induced artifacts rather than biological variation.	560
<i>Basecalling and read processing</i>	561
POD5 files were basecalled using Dorado v0.5.0 with the high-accuracy model dna_r10.4.1_e8.2_400bps_hac@v4.3.0 [40]. Reads with mean quality < 10 or length < 500 bp were removed. Adapters and concatemers were trimmed using Cutadapt v4.0 [41] in a two-pass, error-tolerant procedure. Filtered reads were aligned to the GRCh38.p13 reference genome using minimap2 v2.26 (map-ont preset) [42]. BAM files were sorted and indexed using SAMtools v1.16 [43]. Read-length and mapping statistics were computed using NanoPlot v1.46.1 [44]. All samples were processed using identical parameters.	562
<i>Chimeric read identification</i>	563
Chimeric reads were identified from BAM files using supplementary alignment (SA) tags. Reads were classified as chimeric if they (i) were mapped, (ii) contained an SA tag, (iii) were primary alignments (not secondary), and (iv) were not supplementary alignments themselves. This definition counts each chimeric read once using its primary	564
	565
	566
	567
	568
	569
	570
	571
	572
	573
	574
	575
	576
	577
	578
	579
	580
	581
	582
	583
	584
	585
	586
	587
	588
	589
	590
	591
	592
	593
	594
	595
	596
	597
	598

599 alignment while excluding secondary/supplementary records, thereby avoiding double-
600 counting and reducing ambiguity from low-confidence alignments. Reads lacking **SA**
601 tags were classified as non-chimeric.

602

603 **Training data construction**

604

605 *Data generation and sources*

606 To construct the training dataset, we generated **WGA** and bulk sequencing data from
607 PC3 cells. The **WGA** sample was amplified and sequenced on the PromethION P2 plat-
608 form (**ONT**), while three independent bulk datasets were produced from non-amplified
609 genomic DNA: bulk PromethION P2, bulk MinION Mk1c (**ONT**), and bulk **PacBio**.
610 These bulk datasets represent authentic biological sequences free from amplification-
611 induced artifacts. In contrast, **WGA** sequencing includes both genuine genomic reads
612 and artificial chimeras introduced during the amplification process.

613

614 *Ground truth annotation and class definition*

615 Ground truth labels were established by systematically comparing chimeric reads from
616 the **WGA** PromethION P2 dataset against those from the three bulk datasets. For each
617 **WGA** chimeric read, all alignment segments—defined by their genomic start and end
618 coordinates—were compared to the corresponding segments of bulk chimeric reads.
619 A **WGA** read was labeled as biological if every segment matched at least one bulk
620 chimeric read within a 1 kb positional tolerance, indicating that the structural con-
621 figuration is also present in non-amplified DNA. Reads lacking any matching pattern
622 across all bulk datasets were labeled as artificial chimeras, presumed to arise from the
623 amplification process. Additional chimeric reads were randomly sampled from the bulk
624 datasets and labeled as biological, as these reads originate from genuine genomic rear-
625 rangements such as true **SVs**. The final labeled dataset combined the annotated **WGA**
626 PromethION P2 reads with the subsampled bulk chimeric reads and was subsequently
627 partitioned into training, validation, and test sets as described below.

628

629 *Dataset partitioning and cross-platform validation*

630 The combined labeled dataset, derived from **WGA** PromethION P2 and bulk sequenc-
631 ing data, was divided into training (70%), validation (20%), and test (10%) sets using
632 stratified random sampling. These subsets were used respectively for model training,
633 hyperparameter tuning, and performance evaluation on data from the same sequencing
634 platform.

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

640

641

642

643

644

Model architecture	645
DNA encoder	646

ChimeraLM employs the pre-trained HyenaDNA model [29] as its DNA encoder. This model was pre-trained on large-scale genomic data and provides robust sequence representations. DNA sequences are tokenized at single-nucleotide resolution, with each base (A, C, G, T, N) mapped to a unique integer token (7, 8, 9, 10, 11, respectively). Special tokens include [CLS]=0, [PAD]=4, and others for sequence processing. Input sequences are truncated at 32,768 bp or padded to enable batch processing.

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

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

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L)$ represents position-wise hidden states with dimension 256. The Hyena operators [33] efficiently capture both local sequence motifs and long-range dependencies essential for distinguishing biological sequences from chimeric artifacts.

Attention pooling

To aggregate variable-length sequence representations into fixed-size vectors, ChimeraLM implements attention-based pooling. For hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$, attention weights are computed through a two-layer network:

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

The pooled representation is the weighted sum of hidden states:

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

This mechanism assigns learned importance weights to each sequence position, enabling the model to focus on informative regions while accommodating natural variability in read lengths.

Classification head

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

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

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

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

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

693

$$694 \quad \mathbf{z} = [z_0, z_1] = \text{Linear}_{512 \rightarrow 2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2$$

695

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

698

699 **Model summary**

700 The complete ChimeraLM pipeline processes DNA sequences through: (1) single-
701 nucleotide tokenization, (2) HyenaDNA backbone encoding to generate contextualized
702 representations, (3) attention pooling to aggregate position-specific features, (4) [MLP](#)
703 layers with residual connections to learn classification features, and (5) binary clas-
704 sification output. The entire model with ~ 4.2 M trainable parameters is trained
705 end-to-end using labeled data.

706

707 **Model training and optimization**

708

709 **Training configuration**

710 ChimeraLM was trained using PyTorch [45] and PyTorch Lightning [46] frameworks.
711 Input sequences were tokenized using the tokenizer with maximum sequence length of
712 32,768 bp. Sequences longer than this threshold were truncated; shorter sequences were
713 padded to enable batch processing. Training employed mixed-precision computation
714 (bf16) to accelerate training while maintaining numerical stability.

715

716 **Optimization procedure**

717 We used the AdamW optimizer [47] with learning rate $\eta = 1 \times 10^{-4}$ and weight
718 decay $\lambda = 0.01$. AdamW implements adaptive learning rates with decoupled weight
719 decay, combining the benefits of Adam optimization with proper L2 regularization.
720 A ReduceLROnPlateau scheduler dynamically adjusted the learning rate based on
721 validation loss, reducing it by a factor of 0.1 when no improvement occurred for 10
722 consecutive epochs. Early stopping with patience of 10 epochs prevented overfitting
723 by terminating training when validation performance plateaued. A fixed random seed
724 (12345) ensured reproducibility across training runs.

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

727

$$728 \quad \mathcal{L}(\mathbf{z}, y) = -\log \left(\frac{\exp(z_y)}{\exp(z_0) + \exp(z_1)} \right) = -z_y + \log(\exp(z_0) + \exp(z_1))$$

729

730 where z_0 and z_1 represent logits for biological and artificial chimeric classes, respec-
731 tively.

732

733 **Training implementation**

734 Training used batch size of 16 sequences with 30 parallel data loading workers. [GPU](#)
735 acceleration was employed for efficient processing, with training typically requiring

55 hours. Model checkpointing saved the best-performing model based on validation metrics. Configuration management used Hydra [48] to enable reproducible experimentation.

Model evaluation

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

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

where TP (true positives) are chimeric reads correctly classified as artificial, TN (true negatives) are biological reads correctly classified as biological, FP (false positives) are biological reads misclassified as artificial, and FN (false negatives) are artificial reads misclassified as biological. Final model selection was based on best validation performance as determined by early stopping.

Model inference and application

Inference pipeline

To apply ChimeraLM to new WGA sequencing data, the model takes a BAM file as input. Chimeric reads are identified using SA tags and filtered to exclude unmapped, secondary, or supplementary alignments. Each chimeric read sequence is tokenized using the tokenizer (maximum length 32,768 bp, with truncation or padding as needed). The trained model processes sequences in batches, generating two logits $[z_0, z_1]$ for each read corresponding to biological and artificial chimeric classes. Classification is determined by $\hat{y} = \text{argmax}(z_0, z_1)$. ChimeraLM outputs a filtered BAM file containing only reads classified as biological, which can be directly used for downstream analyses including SV calling.

Performance evaluation

Test set evaluation

Final model performance was evaluated on the held-out test set and the independent MinION Mk1c dataset. Metrics (precision, recall, and F1 score) were computed as described in the training section, where true positives represent chimeric reads correctly classified as artificial and true negatives represent biological reads correctly classified as biological.

SV calling

SVs were called using multiple tools to ensure comprehensive detection. For long-read data (ONT PromethION P2 and MinION Mk1c), we used Sniffles v2.5 [14, 15], DeBreak v1.2 [16], SVIM v2.0.0 [17], and cuteSV v2.1.1 [18]. For short-read data of the PC3 cell line, we used both the CCLE Illumina WGS dataset and the PRJNA361315

783 Illumina **WGS** dataset, processed with Manta v1.6.0 [49], DELLY v1.5.0 [50], and
784 SvABA v1.1.0 [51]. All tools were executed with default recommended parameters.

785

786 ***Gold standard SV dataset construction***

787 To evaluate the impact of ChimeraLM on **SV** detection accuracy, we generated a high-
788 confidence gold-standard **SV** set from bulk PC3 sequencing data. All **SV** comparisons
789 and breakpoint corrections were performed using OctopuSV v0.2.3 [52]. Four bulk
790 datasets were integrated: **ONT** MinION Mk1c, **ONT** PromethION P2, the CCLE Illumina
791 **WGS** dataset, and the PRJNA361315 Illumina **WGS** dataset. **SVs** were called
792 independently within each dataset, and events supported by at least two **SV** callers
793 were retained. The remaining calls were then compared across datasets, and **SVs**
794 observed in at least two of the four datasets were designated as gold-standard events
795 for benchmarking.

796

797 ***SV benchmarking analysis***

798 To assess the impact of ChimeraLM on **SV** calling accuracy, we compared **SV** calls from
799 unfiltered **WGA** data and ChimeraLM-filtered **WGA** data against two references: (1)
800 the stringent multi-platform gold standard dataset, and (2) platform-matched long-
801 read bulk sequencing data. Benchmarking was performed using Truvari v4.2.2 [53]
802 with default parameters. **SVs** were considered supported if they matched reference
803 variants within the defined breakpoint tolerance. Validation rates were calculated as
804 the proportion of called **SVs** supported by the reference. This dual benchmarking
805 strategy quantifies both improvements in detecting high-confidence multi-platform
806 **SVs** and the retention of platform-specific true variants.

807

808 **Benchmarking against existing methods**

809 ChimeraLM was compared to existing computational methods for detecting
810 amplification-induced chimeric artifacts: SACRA [25] (GitHub commit 9a2607e), 3rd-
811 ChimeraMiner [13] (GitHub commit 04b5233), and CADECT v1.2 [11]. Both tools
812 were applied to **WGA** data from PromethION P2 and MinION Mk1c platforms using
813 default parameters as recommended in their documentation. Performance was evalua-
814 ted by measuring the percentage reduction in chimeric reads relative to unprocessed
815 **WGA** data. Chimeric reads were identified using **WGA** tag-based alignment criteria
816 (reads with **SA** tags indicating split alignments), and reduction rates were calculated
817 as the proportion of chimeric reads removed by each method.

818

819 **Attention weight analysis**

820

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

827

828

Chimeric junction positions were identified from alignment data (defined by break-points in SA tags). A region of ± 50 bp surrounding each junction was designated as the junction region. Attention weights within junction region were compared to non-junction regions using the Wilcoxon rank-sum test [54], with statistical significance assessed at $p < 0.001$. 829
830
831
832
833
834

Data visualization

Figures were generated using Python with Matplotlib [55] and Seaborn [56]. 835
836
837

Computing resources

Computations were performed on a high performance computing (HPC) server with 64-core Intel Xeon Gold 6338 CPU, 256 GB RAM, and two NVIDIA A100 GPUs (80 GB memory each). 838
839
840
841
842
843
844
845

Extended Data Table 1 Sequencing and alignment statistics of PC3

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

Supplementary information.

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

**901 Extended Data Fig. 1 Training dataset construction and bulk-supported labeling strat-
902 egy.** (a) Workflow for generating labeled training data. **WGA** PromethION data is compared against
903 three independent bulk sequencing datasets (PromethION, MinION, and PacBio). Reads with no
904 bulk matches (Match 0) are labeled artificial; reads matching one or more bulk datasets (Match 1–3)
905 are labeled biological, along with chimeric reads sampled directly from bulk data. The labeled dataset
906 is split into training (70%), validation (20%), and test (10%) sets. The **WGA** MinION dataset is
907 reserved for independent cross-platform evaluation. (b) Distribution of chimeric read matches. Bar
908 chart shows the number of **WGA** PromethION chimeric reads (log scale) by bulk dataset matches.
909 Match 0 reads ($\sim 10^7$) lacking bulk validation are classified as artificial; Match 1–3 reads with bulk sup-
910 port are classified as biological. The substantial imbalance reflects high prevalence of **WGA**-induced
911 artifacts.

912

913

914

915

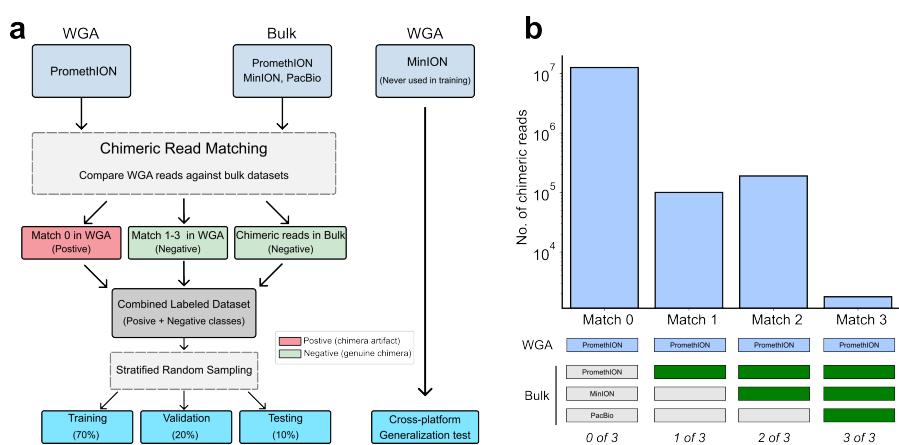
916

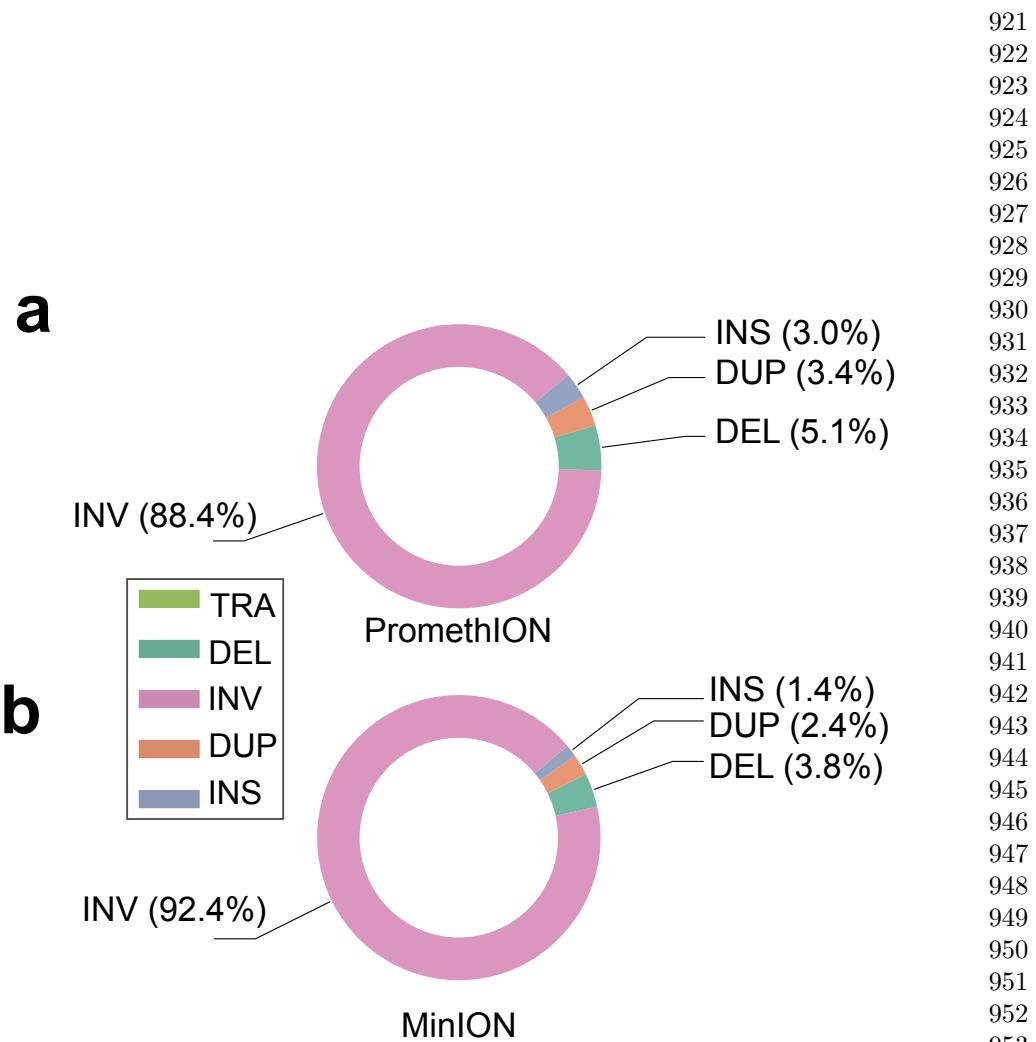
917

918

919

920





Extended Data Fig. 2 Composition of artifact-supported SVs for PromethION (a) and MinION (b). Donut charts summarize the distribution of SV types including translocation (TRA), deletion (DEL), inversion (INV), duplication (DUP), and insertion (INS) among events supported exclusively by chimera artifacts, representing artificial SVs preferentially removed by ChimeraLM.

921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966

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

970

971 **Declarations**

972

973 **Author Contributions.** YL, QG and RY designed the study. YL and QG per-
974 formed the analysis. QG performed the experiments. YL and QG designed and
975 implemented the model. YL built the command-line tool and documentation. YL, QG
976 and RY wrote the manuscript. RY supervised this work.

977

978 **Data Availability.** The raw sequencing data generated in this study have been
979 deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession
980 PRJNA1354861. The dataset includes Oxford Nanopore long-read whole-genome
981 sequencing of PC3 prostate cancer cells and MDA-amplified single-cell derivatives. The
982 individual SRA accessions are as follows: PC3 bulk (MinION Mk1C), SRR35904028;
983 PC3 bulk (PromethION P2), SRR35904029; PC3 10-cell WGA (MinION Mk1C),
984 SRR35904026; PC3 10-cell WGA (PromethION P2), SRR35904027. We can access the
985 data at the following link: <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1354861?reviewer=viej6cv6mgbl3n7a9a5k1bsb3>

986

987 **Code Availability.** ChimeraLM, implemented in Python, is open source and
988 available on GitHub (<https://github.com/ylab-hi/ChimeraLM>) under the Apache
989 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
990 macOS to ensure easy cross-platform installation. An interactive demo is available on
991 Hugging Face (<https://huggingface.co/spaces/yangliz5/ChimeraLM>), allowing users
992 to test ChimeraLM's functionality without local installation. For large-scale analy-
993 ses, we recommend using ChimeraLM on systems with GPU acceleration. Detailed
994 system requirements and optimization guidelines are available in the repository's
995 documentation (<https://ylab-hi.github.io/ChimeraLM/>).

996

997 **Conflict of interest.** The authors declare no competing interests.

998

999

1000 **Acronyms**

1001

1002 **CPU** central processing unit [12](#)

1003

1004 **DEL** deletion [7, 21](#)

1005 **dMDA** droplet-based MDA [12](#)

1006

1007 **DUP** duplication [7, 21](#)

1008

1009 **GLM** genomic language model [1, 2, 12](#)

1010 **GPU** graphics processing unit [12, 16, 19, 22](#)

1011

1012 **HPC** high performance computing [19](#)

1013

1014 **INS** insertion [7, 21](#)

INV inversion	7, 21	1013
LIANTI linear amplification via transposon insertion	12	1014
MALBAC multiple annealing and looping-based amplification cycles	12	1015
MDA multiple displacement amplification	2, 9	1016
MLP multilayer perceptron	4, 5, 15, 16	1017
ONT Oxford Nanopore Technologies	3, 7, 8, 11, 13, 14, 17, 18	1018
PacBio Pacific Biosciences	3, 14	1019
PTA primary template-directed amplification	12	1020
SA supplementary alignment	13, 14, 17–19	1021
SV structural variation	1–4, 6–9, 11, 12, 14, 17, 18, 21	1022
TRA translocation	7, 21	1023
WGA whole genome amplification	1–12, 14, 17–20	1024
WGS whole genome sequencing	7, 8, 17, 18	1025
References		
[1]	Kalef-Ezra, E. <i>et al.</i> Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. <i>Communications Biology</i> 1288 (2024).	1026
[2]	Navin, N. <i>et al.</i> Tumour evolution inferred by single-cell sequencing. <i>Nature</i> 472 , 90–94 (2011).	1027
[3]	Sun, C. <i>et al.</i> Mapping recurrent mosaic copy number variation in human neurons. <i>Nature Communications</i> 4220 (2024).	1028
[4]	Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. <i>Nature Reviews Genetics</i> 175–188 (2016).	1029
[5]	Chen, C. <i>et al.</i> Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). <i>Science (new York, N.Y.)</i> 356 , 189–194 (2017).	1030
[6]	Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives. <i>PLOS Genetics</i> 10 , e1004126 (2014).	1031
[7]	de Bourcy, C. F. A. <i>et al.</i> A quantitative comparison of single-cell whole genome amplification methods. <i>PLoS ONE</i> e105585 (2014).	1032
[8]	Biezuner, T. <i>et al.</i> Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. <i>Scientific Reports</i> 17171 (2021).	1033

- 1059 [9] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement
1060 amplification. *Computational and Structural Biotechnology Journal* 1688–1696
1061 (2023).
- 1062
- 1063 [10] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the
1064 multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).
- 1065
- 1066 [11] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displace-
1067 ment amplification with whole-genome oxford nanopore sequencing. *Molecular*
1068 *Ecology Resources* e14094 (2025).
- 1069
- 1070 [12] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple
1071 displacement amplification. *Proceedings of the National Academy of Sciences* **99**,
1072 5261–5266 (2002).
- 1073
- 1074 [13] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric
1075 sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275
1076 (2023).
- 1077
- 1078 [14] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using
1079 single-molecule sequencing. *Nature Methods* 461–468 (2018).
- 1080
- 1081 [15] Smolka, M. *et al.* Detection of mosaic and population-level structural variants
1082 with sniffles2. *Nature Biotechnology* 1571–1580 (2024).
- 1083
- 1084 [16] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using
1085 long sequencing reads with DeBreak. *Nature Communications* 283 (2023).
- 1086
- 1087 [17] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped
1088 long reads. *Bioinformatics* 2907–2915 (2019).
- 1089
- 1090 [18] Jiang, T. *et al.* Long-read-based human genomic structural variation detection
1091 with cuteSV. *Genome Biology* 189 (2020).
- 1092
- 1093 [19] Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with
1094 primary template-directed amplification. *Proceedings of the National Academy of*
1095 *Sciences* **118**, e2024176118 (2021).
- 1096
- 1097 [20] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection
1098 algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).
- 1099
- 1100 [21] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and
1101 genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011).
- 1102
- 1103 [22] Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nature*
1104 *Communications* **14**, 5164 (2023).

- [23] Gupta, P., O'Neill, H., Wolvetang, E. J., Chatterjee, A. & Gupta, I. Advances in single-cell long-read sequencing technologies. *NAR Genomics and Bioinformatics* **6**, lqae047 (2024).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [24] Wen, L. & Tang, F. Single-cell omics sequencing technologies: The long-read generation. *Trends in Genetics* **0** (2025).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [25] 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).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [26] Ha, Y.-J. *et al.* Comprehensive benchmarking and guidelines of mosaic variant calling strategies. *Nature Methods* **20**, 2058–2067 (2023).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [27] Höijer, I. *et al.* Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. *Genome Biology* **21**, 290 (2020).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [28] Li, Y. *et al.* A genomic language model for chimera artifact detection in nanopore direct rna sequencing. *bioRxiv* (2024). URL <https://www.biorxiv.org/content/early/2024/10/25/2024.10.23.619929>.
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [29] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution*, Vol. 36, 43177–43201 (Curran Associates, Inc., 2023).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [30] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 287–297 (2025).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [31] Zhou, Z. *et al.* *DNABERT-2: Efficient foundation model and benchmark for multi-species genomes*, 1–24 (OpenReview.net, 2024).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [32] Consens, M. E. *et al.* To transformers and beyond: Large language models for the genome (2023). [arXiv:2311.07621](https://arxiv.org/abs/2311.07621).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [33] Poli, M. *et al.* *Hyena hierarchy: Towards larger convolutional language models*, Vol. 202, 28043–28078 (PMLR, 2023).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [34] Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [35] Seim, I., Jeffery, P. L., Thomas, P. B., Nelson, C. C. & Chopin, L. K. Whole-Genome Sequence of the Metastatic PC3 and LNCaP Human Prostate Cancer Cell Lines. *G3 (Bethesda, Md.)* **7**, 1731–1741 (2017).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [36] Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nature Communications* **14**, 5164 (2023).
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150

- 1151 [37] Mahmoud, M. *et al.* Structural variant calling: The long and the short of it.
1152 *Genome Biology* **20**, 246 (2019).
- 1153
1154 [38] Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-
1155 nucleotide and copy-number variations of a single human cell. *Science* 1622–1626
1156 (2012).
- 1157
1158 [39] Dippenaar, A. *et al.* Droplet based whole genome amplification for sequencing
1159 minute amounts of purified mycobacterium tuberculosis DNA. *Scientific Reports*
1160 **14**, 9931 (2024).
- 1161 [40] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023).
- 1162
1163 [41] Martin, M. Cutadapt removes adapter sequences from high-throughput sequenc-
1164 ing reads. *Embojournal* **17**, 10–12 (2011).
- 1165
1166 [42] Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*
1167 3094–3100 (2018).
- 1168
1169 [43] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* giab008
1170 (2021).
- 1171
1172 [44] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of
1173 long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
- 1174
1175 [45] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning*
1176 library, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019).
- 1177
1178 [46] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub
1179 repository (2019). URL <https://github.com/Lightning-AI/lightning>.
- 1180 [47] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).
- 1181
1182 [48] Yadan, O. Hydra - a framework for elegantly configuring complex applications.
1183 GitHub repository (2019). URL <https://github.com/facebookresearch/hydra>.
- 1184
1185 [49] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for
1186 germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).
- 1187
1188 [50] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end
1189 and split-read analysis. *Bioinformatics* i333–i339 (2012).
- 1190
1191 [51] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and
1192 indels by local assembly. *Genome Research* 581–591 (2018).
- 1193
1194 [52] Guo, Q., Li, Y., Wang, T.-Y., Ramakrishnan, A. & Yang, R. OctopusV and
1195 TentacleSV: A one-stop toolkit for multi-sample, cross-platform structural variant
1196 comparison and analysis. *Bioinformatics* btaf599 (2025).

- [53] English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined structural variant comparison preserves allelic diversity. *Genome Biology* **23**, 271 (2022). 1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
- [54] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 261–272 (2020). 1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
- [55] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 90–95 (2007). 1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
- [56] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* 3021 (2021). 1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242