

ChimeraLM filters amplification artifacts for
accurate structural variant calling in long-read
single-cell 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 genomics provides unprecedented insights into cellular heterogeneity, but Whole Genome Amplification (WGA)—required to obtain sufficient DNA—introduces chimeric artifacts that generate false-positive Structural Variations (SVs) and undermine biological interpretations. Current computational methods cannot reliably distinguish amplification-induced artifacts from genuine rearrangements. Here we present ChimeraLM, a genomic language model that learns sequence-level features to discriminate biological sequences from WGA artifacts. Validated on nanopore long-read data, ChimeraLM achieves 95% recall with 70% precision, reduces chimeric reads by ~90%, and preserves 72–92% of true SVs. This improves SV validation rates 8.5–11.0 fold and eliminates artificial inversion (INV) bias, restoring SV type distributions to bulk-like profiles. Attention visualization reveals that ChimeraLM focuses on chimeric junction regions, learning mechanistically interpretable features that generalize across sequencing platforms. By enabling reliable SV detection at single-cell resolution, ChimeraLM addresses a critical data quality barrier in cancer genomics, developmental biology, and somatic mosaicism studies. ChimeraLM is available at <https://github.com/ylab-hi/ChimeraLM>.

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

Main

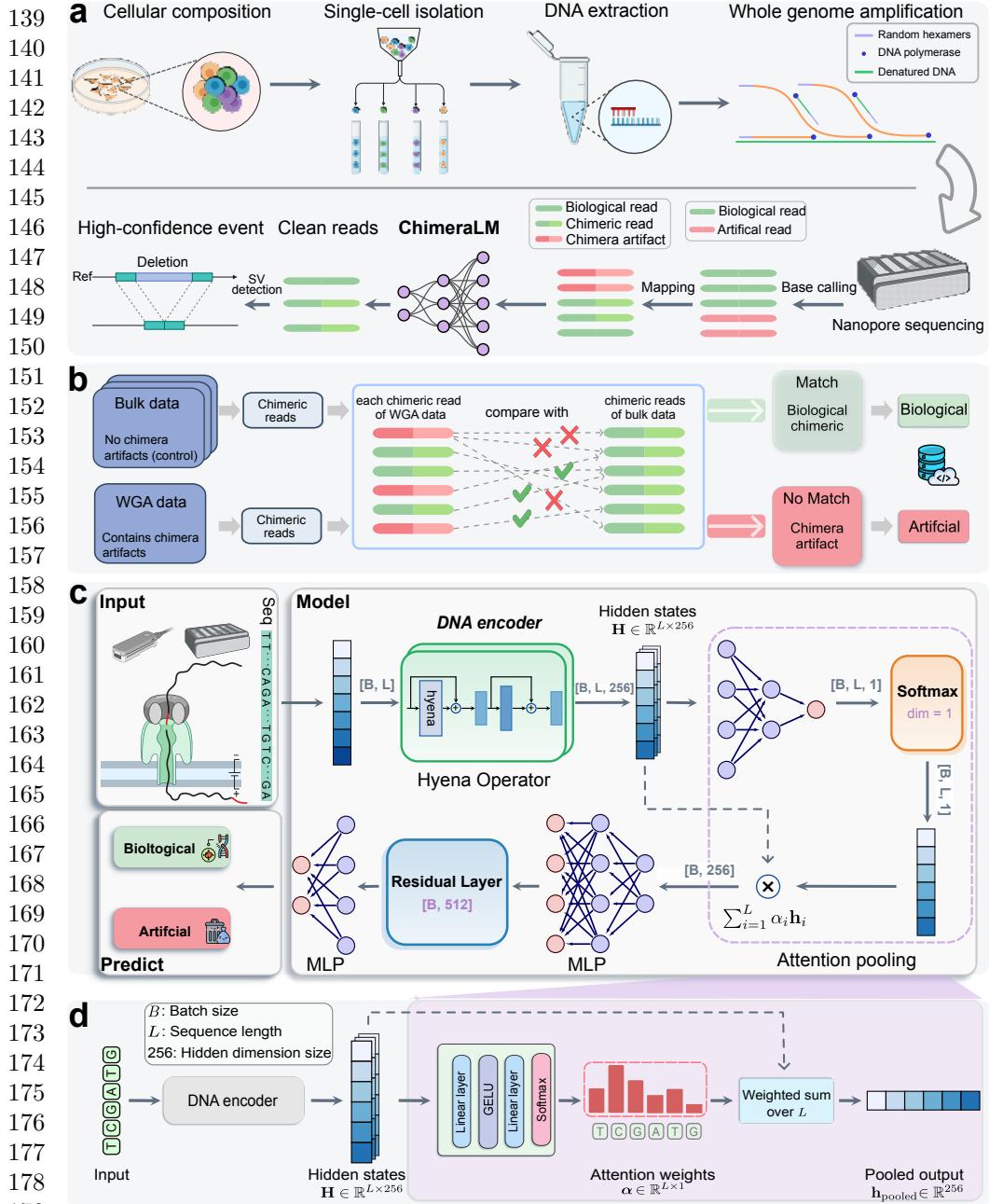
Single-cell and ultra-low-input genomics have transformed our ability to resolve biological heterogeneity, enabling the discovery of rare cell states and the reconstruction of clonal evolution in cancer and development [1–3]. However, the limited DNA input (on the order of picograms per cell) makes comprehensive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

062 A prominent source of error is amplification-induced chimera formation, in
 063 which highly processive polymerases—such as phi29 used in [Multiple Displacement](#)
 064 [Amplification \(MDA\)](#)—switch templates and join discontinuous genomic loci into
 065 a single molecule [9–13]. In long-read sequencing, which is otherwise well suited
 066 for resolving complex [SVs](#), chimeric reads can constitute a substantial fraction of
 067 [WGA](#) data [9], generating alignment patterns that resemble genuine [translocations](#)
 068 ([TRAs](#)) and inversions ([inversions \(INVs\)](#)) [10]. As a result, [SV](#) callers that rely
 069 on alignment-based signals (for example, split-read and supplementary alignments)
 070 and coverage-derived evidence frequently misinterpret amplification artifacts as true
 071 rearrangements, inflating false positives and distorting [SV](#) spectra [14–21].

Distinguishing biological rearrangements from amplification artifacts remains a major computational bottleneck. Existing quality-control approaches typically rely on handcrafted rules or alignment-derived features, such as read orientation signatures or local coverage anomalies [11, 13, 22]. These heuristics are often sensitive to platform- and protocol-specific variation and fail to capture sequence-intrinsic patterns near chimera junctions as well as long-range context within individual reads. This limitation has constrained the practical use of low-input long-read sequencing in applications where precision is essential, including somatic mosaicism profiling and validation of CRISPR off-target effects.

To address this challenge, we present ChimeraLM, an interpretable [Genomic Language Model \(GLM\)](#) for single-read identification and filtering of [WGA](#)-induced artifacts. Unlike traditional approaches that depend on alignment heuristics, ChimeraLM formulates artifact detection as a sequence modeling task, learning discriminative features directly from raw DNA sequences [23]. Leveraging advances in DNA foundation models [24–27], it captures latent motifs and structural dependencies that generalize across nanopore platforms. Across nanopore [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 restoring bulk-like [SV](#)-type distributions. By enabling reliable [SV](#) discovery in long-read sequencing, ChimeraLM removes a critical data-quality barrier for single-cell genomics.

Results	093
Overview of ChimeraLM workflow and model architecture	094
Single-cell long-read genomics requires WGA to obtain sufficient DNA for sequencing (Fig. 1a). However, WGA introduces artifacts that generate chimeric reads composed of segments from distant genomic loci, which can confound SV detection. ChimeraLM integrates into this workflow as a post-alignment filtering module (Fig. 1a). It evaluates each chimeric read prior to variant calling, predicting whether the sequence reflects a genuine genomic rearrangement or a WGA -induced artifact. This binary classification enables the selective removal of false positives while preserving authentic biological signal.	095
To train and evaluate the model, we constructed a high-confidence labeled dataset using two WGA datasets generated from the same monoclonal PC3 lineage on independent Oxford Nanopore Technologies (ONT) platforms. The ONT PromethION WGA dataset was designated for model training, whereas the ONT MinION WGA dataset was reserved exclusively for the generalization test. For supervised labeling, chimeric reads from the PromethION WGA dataset were compared against unamplified bulk DNA sequenced on three long-read platforms (ONT PromethION, ONT MinION, and Pacific Biosciences (PacBio)) (see Methods ; Fig. 1b; Extended Data Fig. 1a)). WGA reads were labeled biological if their chimeric structures were supported by any bulk dataset, and artificial if they were absent from all bulk references.	096
This labeling procedure yielded two groups of WGA chimeric reads (Extended Data Fig. 1b). Most reads—12,670,396 in total—showed no supporting alignments in any bulk dataset and were classified as artificial. The remaining 293,180 reads had at least one matching breakpoint in bulk sequencing, providing evidence that they represent biological rearrangements rather than amplification artifacts. To construct the supervised training dataset, we retained all 293,180 biological reads and selected an equal number of artificial reads through random subsampling. We further expanded the biological class by adding 178,748 chimeric reads sampled directly from bulk sequencing data, which represent genuine structural rearrangements independent of WGA -induced artifacts. This augmentation ensures that the classifier is exposed to a broader spectrum of true biological chimeric structures. The final labeled dataset (765,108 reads) was split into training (70%), validation (20%), and test (10%) sets using stratified sampling.	097
Effective classification of WGA chimeric reads requires a model that can process long, variable-length DNA sequences while retaining single-nucleotide resolution (Fig. 1c). ChimeraLM addresses these challenges through a sequence encoder based on HyenaDNA [24], a genomic foundation model pre-trained on diverse DNA sequences. Input reads are tokenized at single-nucleotide resolution and processed by Hyena operators [28], which are designed to model long-range dependencies on long reads while still allowing full-length analysis without splitting the sequence. The encoder outputs a matrix of hidden states across the read. To obtain a fixed-length representation, an attention-pooling module (Fig. 1d) assigns learned, position-specific weights and computes a weighted sum over the sequence. The pooled vector is then processed by MLP	098
	099
	100
	101
	102
	103
	104
	105
	106
	107
	108
	109
	110
	111
	112
	113
	114
	115
	116
	117
	118
	119
	120
	121
	122
	123
	124
	125
	126
	127
	128
	129
	130
	131
	132
	133
	134
	135
	136
	137
	138



180 **Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts.** (a) Single-cell
181 genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction
182 and **WGA**. During amplification, chimeric artifacts (red) are generated alongside biological
183 reads (green). After base calling and mapping, ChimeraLM classifies chimeric reads as biological
184 or artificial, enabling downstream **SV** detection on filtered data. (b) Ground truth label genera-
tion. Chimeric reads from **WGA** data are compared against chimeric reads from bulk sequencing of
the same cell line. Reads matching bulk data are labeled biological (green); non-matching reads are
labeled artificial (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 [24]). Hyena operators capture long-range dependencies. Atten-
tion pooling aggregates position-specific features, and multilayer perceptron (MLP) layers with residual
connections process pooled representations for binary classification. (d) Attention pooling mech-
anism. 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.

blocks with residual connections, and a final softmax layer outputs the probability that each read is biological or artificial.

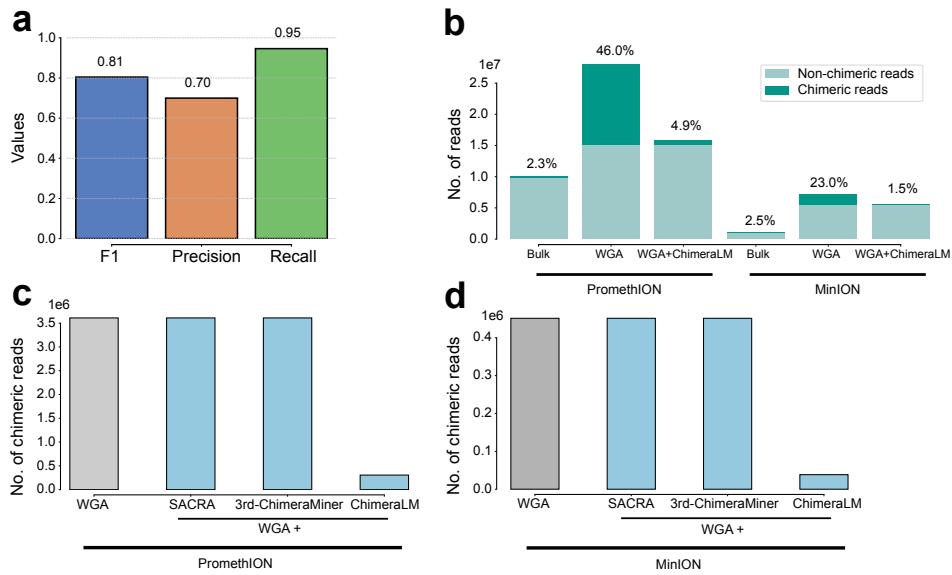


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts. (a) Classification performance on held-out test data. ChimeraLM achieves recall of 0.95, precision of 0.70, and F1 score of 0.81. (b) Chimeric read reduction across sequencing platforms. Stacked bars show proportions of 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 bar represents the total count of chimeric reads. ChimeraLM achieves approximately 90% reduction in chimeric reads on both platforms; SACRA and 3rd-ChimeraMiner show no detectable reduction.

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

We first evaluated ChimeraLM on the held-out test set derived from the labeled dataset (Fig. 2a; see Methods). This test set comprises chimeric reads with known biological or artificial status based on the ground-truth labeling procedure described above. On this benchmark, ChimeraLM achieved an F1 score of 0.81, with a recall of 0.95 and a precision of 0.70. The high recall indicates that 95% of artificial chimeric reads were correctly identified and removed, which is critical for minimizing downstream false-positive SV calls, while the precision confirms that most flagged reads correspond to true artifacts rather than biological rearrangements.

We next asked whether ChimeraLM filtering could restore chimeric read rates in full PC3 WGA datasets to bulk baselines on both PromethION and MinION platforms (Fig. 2b). Bulk sequencing established low baseline chimeric read rates of 2.3%

231 (PromethION) and 2.5% (MinION). In contrast, **WGA** increased the chimeric fraction to 232 46.0% and 23.0%, respectively. After ChimeraLM filtering, chimeric content 233 dropped to 4.9% on PromethION and 1.5% on MinION, corresponding to 10- to 15-fold 234 reductions, while retaining 15.8 million and 5.6 million biological reads. These post- 235 filtering rates approach bulk baselines, indicating effective removal of **WGA**-induced 236 artifacts while preserving authentic biological signal.

237 We benchmarked ChimeraLM against SACRA [22] and 3rd-ChimeraMiner [13], two 238 existing tools for detecting amplification-induced chimeras (Fig. 2c,d). ChimeraLM 239 achieved approximately 90% reduction in chimeric reads on both platforms, whereas 240 neither SACRA nor 3rd-ChimeraMiner produced a detectable reduction (0%).

241 The MinION results are particularly informative because this platform was never 242 used during model training. ChimeraLM was trained exclusively on PromethION 243 **WGA** data, yet achieved comparable chimeric read reduction on MinION. This cross- 244 platform generalization indicates that ChimeraLM captures sequence-level features 245 intrinsic to **WGA**-induced artifacts rather than platform-specific signatures, sup- 246 porting its potential applicability to additional long-read and short-read sequencing 247 technologies.

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

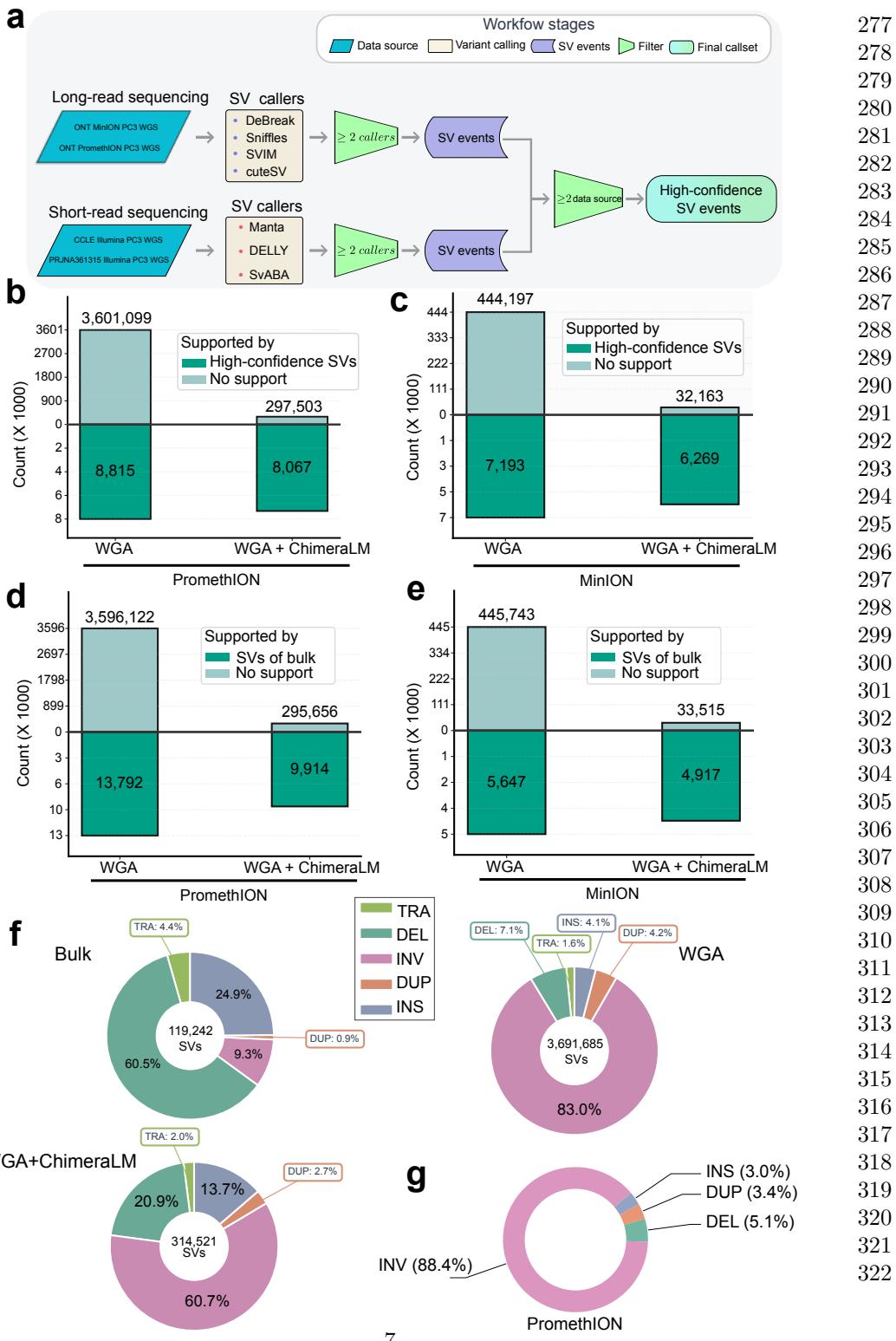


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of a high-confidence SV reference dataset from bulk PC3 sequencing. Four bulk datasets were integrated: ONT MinION Mk1C, ONT PromethION P2, the CCLE Illumina Whole Genome Sequencing (WGS) dataset, and the PRJNA361315 Illumina 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 SV type distributions for PromethION across bulk, unfiltered WGA, and WGA+ChimeraLM. Unfiltered WGA shows an excess of INVs, which is reduced after ChimeraLM filtering. g Computation of artifact-supported SVs for PromethION. Donut charts summarize SV types among events supported exclusively by chimeric reads, representing artificial SVs preferentially removed by ChimeraLM.

323 **ChimeraLM substantially reduces unsupported structural
324 variant calls**

325 Accurate **SV** detection from single cells is essential for understanding genomic diver-
326 sity and disease mechanisms. However, **WGA**-induced chimeric artifacts can be
327 misidentified as genuine **SVs**, leading to incorrect biological conclusions. To quantify
328 ChimeraLM’s impact on **SV** calling, we compared **SV** callsets generated from unfiltered
329 **WGA** reads with those generated after ChimeraLM filtering (**WGA** + ChimeraLM).
330 We evaluated both callsets against two complementary references (Fig. 3): (i) a
331 stringent gold-standard **SV** set derived from bulk PC3 DNA by cross-dataset con-
332 sensus (Fig. 3a), and (ii) platform-matched long-read bulk **SV** callsets used as a
333 platform-specific reference for recall (Fig. 3d,e).

334 We first constructed a high-confidence gold-standard **SV** set from bulk PC3 DNA
335 using four independent sequencing datasets: **ONT** PromethION, **ONT** MinION, and
336 two Illumina whole-genome datasets (the CCLE PC3 **WGS** dataset and PRJNA361315
337 PC3 **WGS** dataset) (Fig. 3a; Extended Data Table 1). **SVs** were called separately
338 within each dataset using multiple **SV** callers. Events supported by at least two callers
339 within a dataset were retained, and only **SVs** observed in at least two of the four
340 datasets were kept as gold-standard events.

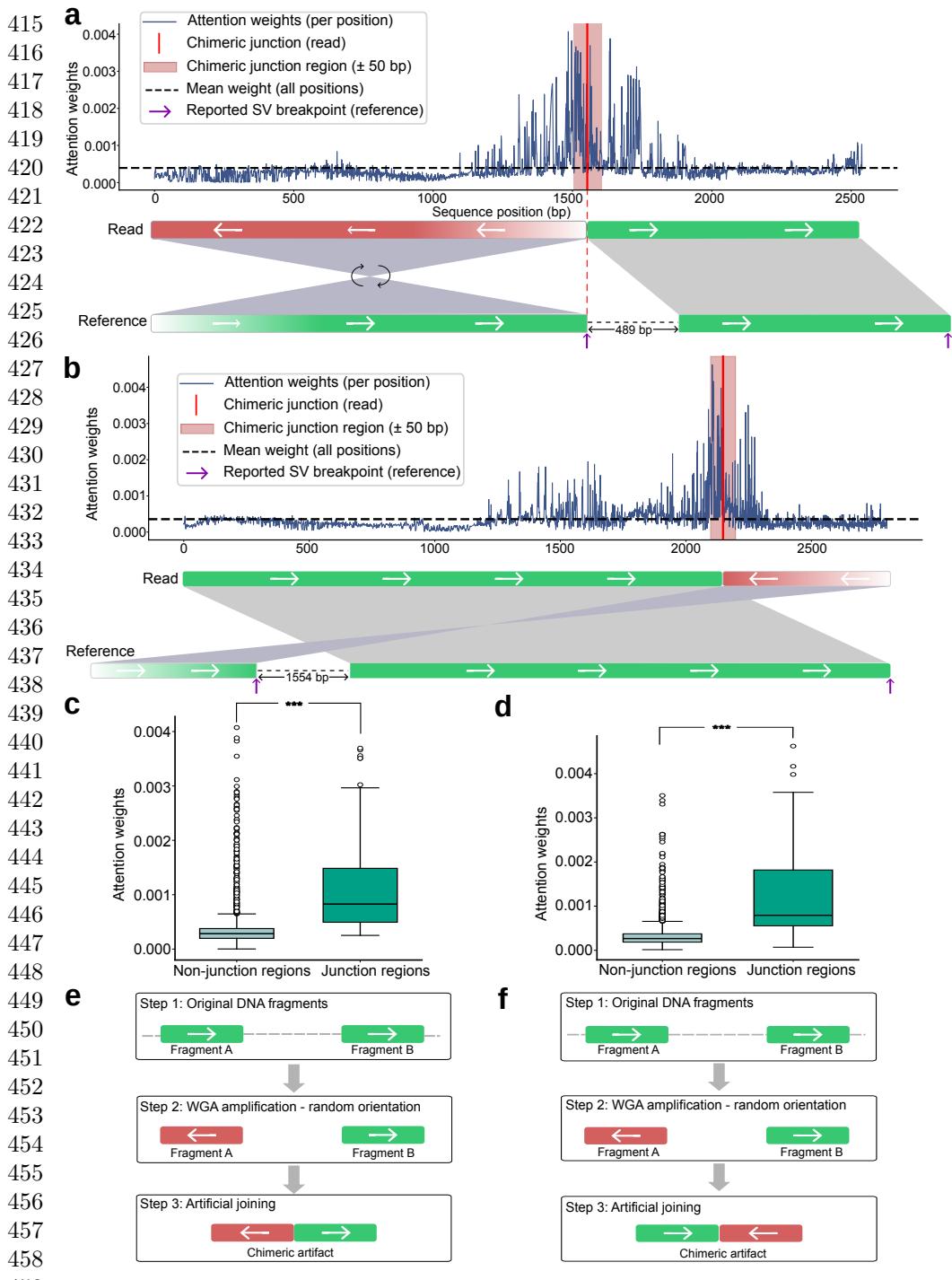
341 Relative to this stringent gold standard, unfiltered **WGA** produced extensive
342 unsupported **SVs** (Fig. 3b,c). On PromethION, **WGA** yielded 3,601,099 **SV** calls, of
343 which only 8,815 (0.24%) overlapped gold-standard events. After ChimeraLM filter-
344 ing, total calls dropped to 305,570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

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

358
359 **ChimeraLM restores bulk-like SV-type distributions**

360 Amplification artifacts can distort the apparent spectrum of **SVs**. We therefore com-
361 pared **SV** type distributions across bulk, unfiltered **WGA**, and ChimeraLM-filtered
362 datasets on both nanopore platforms (Fig. 3f; Extend Data Fig. 2). Bulk sequenc-
363 ing showed a balanced mixture of **deletions (DELs)**, **duplications (DUPs)**, **insertions**
364 (**INSs**), **INVs**, and **TRAs**. In contrast, unfiltered **WGA** callsets were dominated by
365 **INVs** on both platforms. After ChimeraLM filtering, excessive **INVs** were markedly
366
367
368

reduced, and the overall SV type profile shifted toward the bulk distribution, while the relative proportions of other SV classes remained largely stable.	369 370
To identify which SV types were primarily driven by WGA -induced artifacts, we examined SV calls supported exclusively by reads classified as chimera artifacts (Fig. 3g; Extend Data Fig. 3). These artifact-supported events were overwhelmingly INVs , accounting for 88.4% on PromethION and 92.4% on MinION. The remaining calls included smaller fractions of DELs (5.1% and 3.8%), DUPs (3.4% and 2.4%), and INSs (3.0% and 1.4%), indicating that WGA -induced chimeras can generate false positives across multiple SV categories.	371 372 373 374 375 376 377
Together, these results show that WGA artifacts preferentially inflate INVs but are not limited to a single SV class. By selectively removing artifact-supported events and restoring SV type distributions toward bulk-like patterns, ChimeraLM improves the robustness and interpretability of single-cell SV analyses.	378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414



460 **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 INV-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
Discussion	483
WGA enables genomic analysis from single cells but introduces chimeric artifacts that compromise SV detection. ChimeraLM addresses this challenge by classifying chimeric reads as biological or artificial from sequence information and filtering WGA -induced artifacts before variant calling, rather than attempting to correct artifact-driven calls post hoc. Across nanopore platforms, ChimeraLM yielded consistent improvements at both read and variant levels. It reduced chimeric reads by ~90% while retaining 72–92% of bulk-supported SVs , and it lowered unsupported SV calls by 8.5–11.0 fold. Performance generalized from PromethION (used for training) to MinION without platform-specific retraining, indicating that ChimeraLM captures properties shared by WGA -induced artifacts rather than instrument-specific signatures. In contrast, SACRA and 3rd-ChimeraMiner failed to reduce chimeric content on our long-read WGA datasets, underscoring the limitations of heuristic strategies and indicating the need for models that learn discriminative features directly from sequence data.	485
The efficacy of ChimeraLM highlights the utility of deep learning in quality control tasks where conventional metrics (e.g., mapping quality, read depth) provide limited resolution [13, 22, 23]. By learning directly from sequence data, ChimeraLM discovers subtle compositional and structural features that differentiate authentic sequences from amplification artifacts. The model also offers interpretability through attention visualization: attention weights concentrate at junction regions where template switching joins discordant loci, validating the biological relevance of the learned features. These methodological advances have direct implications for single-cell genomics,	499

507 where high false-positive rates in **WGA** data have constrained robust characterization
508 of chromosomal instability, clonal evolution, and **SV** burden [20, 29]. By improving
509 the signal-to-noise ratio and clarifying **SV**-type spectra that are otherwise distorted
510 by amplification artifacts, ChimeraLM enables more confident identification of gen-
511 uine **SVs**, supporting studies of cancer evolution, developmental biology, and somatic
512 mosaicism where single-cell resolution is essential.

513 Several limitations warrant consideration. First, the current model processes reads
514 independently; integrating contextual features such as coverage or phasing informa-
515 tion may further enhance accuracy. Second, regarding computational resources, while
516 **Central Processing Unit (CPU)** inference is feasible, **Graphics Processing Unit (GPU)**
517 acceleration is recommended for processing large-scale datasets. Finally, future work
518 should extend validation to diverse cell types, sequencing platforms (e.g., PacBio
519 HiFi), and alternative **WGA** protocols—including **Multiple Annealing and Looping-**
520 **based Amplification Cycles (MALBAC)** [30], **Linear Amplification via Transposon**
521 **Insertion (LIANTI)** [5], **Primary Template-directed Amplification (PTA)** [19], and
522 **droplet-based MDA (dMDA)** [31]. Although the sequence-level approach implies effec-
523 tive transferability, such broad validation is essential to optimize performance across
524 specific amplification chemistries.

525 Broadly, ChimeraLM illustrates the potential of **GLMs** for genomic data quality
526 control. With emerging long-context architectures [24], the model’s context window
527 could be extended to 1M tokens, enabling analysis of increasingly complex genomic
528 structures. This framework could extend to other amplification-dependent technolo-
529 gies, such as cell-free DNA analysis, ancient DNA studies, and metagenomics from
530 low-biomass samples. Furthermore, attention-based interpretability opens opportuni-
531 ties for studying template-switching dynamics, potentially guiding the development of
532 improved amplification protocols. In summary, ChimeraLM provides a practical and
533 interpretable framework for enhancing long-read single-cell genomic fidelity, ensuring
534 that downstream biological insights are derived from genuine **SVs** rather than technical
535 artifacts.

536

537 **Methods**

538

539 **Cell culture, single-clone preparation, and nanopore sequencing**

540

541 *Cell culture and single-clone establishment*

542 PC3 prostate cancer cells (ATCC® CRL-1435™) were cultured in RPMI-1640 medium
543 supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin at 37 °C
544 with 5% CO₂. To minimize biological heterogeneity, a monoclonal population was
545 established by serial dilution in 96-well plates, ensuring that each culture originated
546 from a single cell. Mycoplasma contamination was routinely tested and confirmed
547 negative prior to DNA extraction.

548

549

550

551

552

DNA extraction and whole-genome amplification	553
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.	554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598
Nanopore library preparation and sequencing	564
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.	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
Basecalling and read processing	571
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 [32]. Reads with mean quality < 10 or length < 500 bp were removed. Adapters and concatemers were trimmed using Cutadapt v4.0 [33] in a two-pass, error-tolerant procedure. Filtered reads were aligned to the GRCh38.p13 reference genome using minimap2 v2.26 (map-ont preset) [34]. BAM files were sorted and indexed using SAMtools v1.16 [35]. Read-length and mapping statistics were computed using NanoPlot v1.46.1 [36]. All samples were processed using identical parameters.	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
Chimeric read identification	581
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 alignment while excluding secondary/supplementary records, thereby avoiding double-counting and reducing ambiguity from low-confidence alignments. Reads lacking SA tags were classified as non-chimeric.	582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598
Training data construction	591
Data generation and sources	592
To construct the training dataset, we generated WGA and bulk sequencing data from PC3 cells. The WGA sample was amplified and sequenced on the PromethION P2 platform (ONT), while three independent bulk datasets were produced from non-amplified genomic DNA: bulk PromethION P2, bulk MinION Mk1c (ONT), and bulk PacBio.	593 594 595 596 597 598

599 These bulk datasets represent authentic biological sequences free from amplification-
600 induced artifacts. In contrast, **WGA** sequencing includes both genuine genomic reads
601 and artificial chimeras introduced during the amplification process.

602

603 ***Ground truth annotation and class definition***

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

618 ***Dataset partitioning and cross-platform validation***

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

624 To evaluate cross-platform generalization, the complete **WGA** MinION Mk1c
625 dataset was reserved. This dataset, generated on a different nanopore platform,
626 was never used during model training or internal testing. This two-level evaluation
627 design allowed us to test whether ChimeraLM captures general sequence features of
628 amplification-induced chimeras rather than platform-specific artifacts.
629

630 **Model architecture**

632 ***DNA encoder***

633 ChimeraLM employs the pre-trained HyenaDNA model [24] as its DNA encoder. This
634 model was pre-trained on large-scale genomic data and provides robust sequence rep-
635 resentations. DNA sequences are tokenized at single-nucleotide resolution, with each
636 base (A, C, G, T, N) mapped to a unique integer token (7, 8, 9, 10, 11, respectively).
637 Special tokens include [CLS]=0, [PAD]=4, and others for sequence processing. Input
638 sequences are truncated at 32,768 bp or padded to enable batch processing.
639

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

$$642 \quad \mathbf{H} = \text{HyenaDNA}(\mathbf{x}) \in \mathbb{R}^{L \times 256}$$

643

644

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L)$ represents position-wise hidden states with dimension 256. 645
The Hyena operators [28] efficiently capture both local sequence motifs and long-range 646
dependencies essential for distinguishing biological sequences from chimeric artifacts. 647

Attention pooling

To aggregate variable-length sequence representations into fixed-size vectors, 650
ChimeraLM implements attention-based pooling. For hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$, 651
attention weights are computed through a two-layer network: 652

$$\begin{aligned}\mathbf{e} &= \text{GELU}(\text{Linear}_{256 \rightarrow 256}(\mathbf{H})) \in \mathbb{R}^{L \times 256} & 654 \\ \mathbf{s} &= \text{Linear}_{256 \rightarrow 1}(\mathbf{e}) \in \mathbb{R}^{L \times 1} & 655 \\ \boldsymbol{\alpha} &= \text{softmax}(\mathbf{s}) \in \mathbb{R}^{L \times 1} & 656\end{aligned}$$

The pooled representation is the weighted sum of hidden states: 657

$$\mathbf{h}_{\text{pooled}} = \sum_{i=1}^L \alpha_i \mathbf{h}_i \in \mathbb{R}^{256} \quad 658$$

This mechanism assigns learned importance weights to each sequence position, 660
enabling the model to focus on informative regions while accommodating natural 661
variability in read lengths. 662

Classification head

The pooled representation is processed through a [MLP](#) with residual connections. The 663
first layer expands dimensionality: 664

$$\mathbf{f}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{Linear}_{256 \rightarrow 512}(\mathbf{h}_{\text{pooled}}))) \in \mathbb{R}^{512} \quad 665$$

Subsequent residual blocks with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ compute: 666

$$\mathbf{f}_{\text{out}} = \text{Dropout}_{0.1}(\text{Linear}_{512 \rightarrow 512}(\text{GELU}(\text{Linear}_{512 \rightarrow 512}(\mathbf{f}_{\text{in}})))) + \mathbf{f}_{\text{in}} \quad 667$$

where the skip connection enables stable gradient flow during training. The final layer 668
produces binary classification logits: 669

$$\mathbf{z} = [z_0, z_1] = \text{Linear}_{512 \rightarrow 2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2 \quad 670$$

where z_0 and z_1 represent logits for biological and artificial chimeric classes, respectively. During inference, the predicted class is $\hat{y} = \text{argmax}_{i \in \{0,1\}} z_i$. 671

Model summary

The complete ChimeraLM pipeline processes DNA sequences through: (1) single-nucleotide tokenization, (2) HyenaDNA backbone encoding to generate contextualized 688
representations, (3) attention pooling to aggregate position-specific features, (4) 689

691 MLP layers with residual connections to learn classification features, and (5) binary
692 classification output. The entire model is trained end-to-end using labeled data.

693

694 Model training and optimization

695

696 Training configuration

697 ChimeraLM was trained using PyTorch [37] and PyTorch Lightning [38] frameworks.
698 Input sequences were tokenized using the tokenizer with maximum sequence length of
699 32,768 bp. Sequences longer than this threshold were truncated; shorter sequences were
700 padded to enable batch processing. Training employed mixed-precision computation
701 (bf16) to accelerate training while maintaining numerical stability.

702

703 Optimization procedure

704 We used the AdamW optimizer [39] with learning rate $\eta = 1 \times 10^{-4}$ and weight
705 decay $\lambda = 0.01$. AdamW implements adaptive learning rates with decoupled weight
706 decay, combining the benefits of Adam optimization with proper L2 regularization.
707 A ReduceLROnPlateau scheduler dynamically adjusted the learning rate based on
708 validation loss, reducing it by a factor of 0.1 when no improvement occurred for 10
709 consecutive epochs. Early stopping with patience of 10 epochs prevented overfitting
710 by terminating training when validation performance plateaued. A fixed random seed
711 (12345) ensured reproducibility across training runs.

712 The training objective used cross-entropy loss for binary classification. For a
713 training example with class label $y \in \{0, 1\}$ and model logits $\mathbf{z} = [z_0, z_1]$, the loss is:
714

$$715 \quad \mathcal{L}(\mathbf{z}, y) = -\log \left(\frac{\exp(z_y)}{\exp(z_0) + \exp(z_1)} \right) = -z_y + \log(\exp(z_0) + \exp(z_1))$$

717

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

720

721 Training implementation

722 Training used batch size of 16 sequences with 30 parallel data loading workers. GPU
723 acceleration was employed for efficient processing, with training typically requiring
724 55 hours. Model checkpointing saved the best-performing model based on valida-
725 tion metrics. Configuration management used Hydra [40] to enable reproducible
726 experimentation.

727

728 Model evaluation

729

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

$$732 \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

733

$$734 \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

735

736

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.	737
	738
	739
	740
	741
	742
Model inference and application	743
<i>Inference pipeline</i>	744
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.	745
	746
	747
	748
	749
	750
	751
	752
	753
	754
	755
	756
Performance evaluation	757
<i>Test set evaluation</i>	758
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.	759
	760
	761
	762
	763
	764
	765
<i>SV calling</i>	766
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 Illumina WGS dataset, processed with Manta v1.6.0 [41], DELLY v1.5.0 [42], and SvABA v1.1.0 [43]. All tools were executed with default recommended parameters.	767
	768
	769
	770
	771
	772
<i>Gold standard SV dataset construction</i>	773
To evaluate the impact of ChimeraLM on SV detection accuracy, we generated a high-confidence gold-standard SV set from bulk PC3 sequencing data. All SV comparisons and breakpoint corrections were performed using OctopusV v0.2.3 [44]. Four bulk datasets were integrated: ONT MinION Mk1c, ONT PromethION P2, the CCLE Illumina WGS dataset, and the PRJNA361315 Illumina WGS dataset. SVs were called independently within each dataset, and events supported by at least two SV callers were retained. The remaining calls were then compared across datasets, and SVs observed in at least two of the four datasets were designated as gold-standard events for benchmarking.	774
	775
	776
	777
	778
	779
	780
	781
	782

783 ***SV* benchmarking analysis**

784 To assess the impact of ChimeraLM on **SV** calling accuracy, we compared **SV** calls from
785 unfiltered **WGA** data and ChimeraLM-filtered **WGA** data against two references: (1)
786 the stringent multi-platform gold standard dataset, and (2) platform-matched long-
787 read bulk sequencing data. Benchmarking was performed using Truvari v4.2.2 [45]
788 with default parameters. **SVs** were considered supported if they matched reference
789 variants within the defined breakpoint tolerance. Validation rates were calculated as
790 the proportion of called **SVs** supported by the reference. This dual benchmarking
791 strategy quantifies both improvements in detecting high-confidence multi-platform
792 **SVs** and the retention of platform-specific true variants.
793

794 **Benchmarking against existing methods**

795 ChimeraLM was compared to two existing computational methods for detecting
796 amplification-induced chimeric artifacts: SACRA [22] (GitHub commit 9a2607e) and
797 3rd-ChimeraMiner [13] (GitHub commit 04b5233). Both tools were applied to **WGA**
798 data from PromethION P2 and MinION Mk1c platforms using default parameters as
799 recommended in their documentation. Performance was evaluated by measuring the
800 percentage reduction in chimeric reads relative to unprocessed **WGA** data. Chimeric
801 reads were identified using **WGA** tag-based alignment criteria (reads with **SA** tags
802 indicating split alignments), and reduction rates were calculated as the proportion of
803 chimeric reads removed by each method.

804 **Attention weight analysis**

805 To investigate ChimeraLM’s interpretability, we analyzed attention weights from
806 the pooling mechanism for representative chimeric reads. Attention weights indicate
807 the relative importance assigned to each sequence position during classification. For
808 selected reads, we extracted per-position attention weights and visualized them along-
809 side read alignments to identify whether the model focuses on mechanistically relevant
810 regions.

811 Chimeric junction positions were identified from alignment data (defined by break-
812 points in **SA** tags). A region of ± 50 bp surrounding each junction was designated as
813 the junction region. Attention weights within junction region were compared to non-
814 junction regions using the Wilcoxon rank-sum test [46], with statistical significance
815 assessed at $p < 0.001$.
816

817 **Data visualization**

818 Figures were generated using Python with Matplotlib [47] and Seaborn [48].

819 **Computing resources**

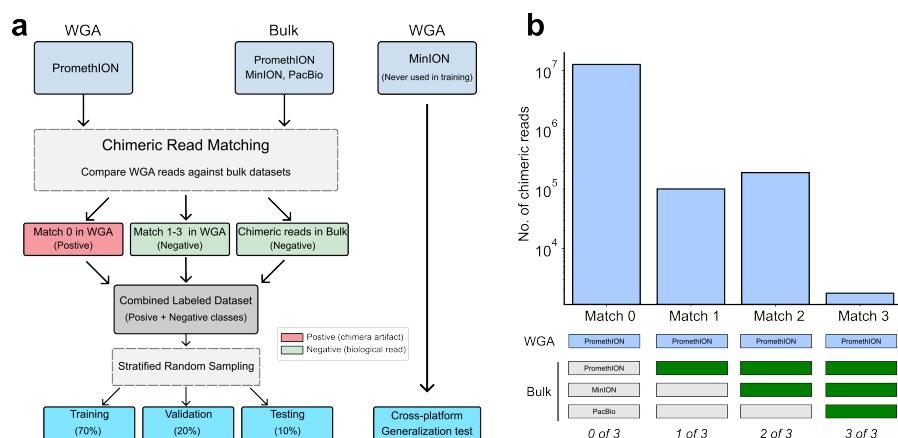
820 Computations were performed on a **High Performance Computing (HPC)** server with
821 64-core Intel Xeon Gold 6338 CPU, 256 GB RAM, and two NVIDIA A100 **GPUs** (80
822 GB memory each).
823

824 **Supplementary information.**

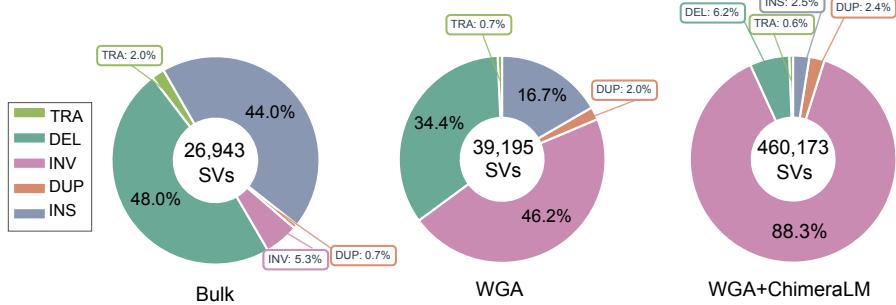
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874

Extended Data Table 1 Sequencing and alignment statistics of PC3

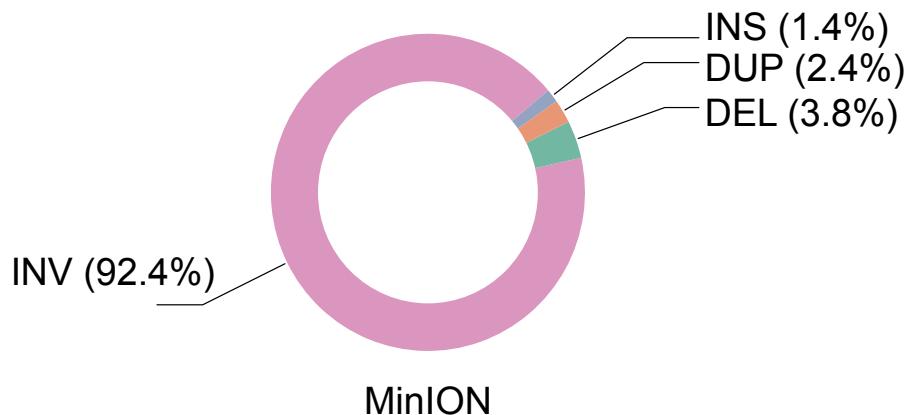
Sample	Platform	Reads ($\times 10^6$)	Total bases (Gb)	Total bases aligned (Gb)	Fraction aligned	Mean length (bp)	Mean quality (Q)	Average identity (%)
WGA	MinION	9.11	14.6	10.4	0.7	1,603	14.3	97.6
WGA	PromethION	44.69	128.2	69.2	0.5	2,869	14.5	96.1
Bulk	MinION	0.97	8.1	7.1	0.9	8,310	17.2	97.3
Bulk	PromethION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



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



Extended Data Fig. 2 **SV** type distributions for MinION across bulk, unfiltered **WGA**, and **WGA**+**ChimeraLM**. Unfiltered **WGA** shows an excess of **INV**s, which is reduced after **ChimeraLM** filtering.



Extended Data Fig. 3 Composition of artifact-supported **SVs** for MinION. Donut charts summarize **SV** types among events supported exclusively by chimeric reads, representing artificial **SVs** preferentially removed by **ChimeraLM**.

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.	921 922 923 924 925 926
Declarations		927 928 929 930
Author Contributions.	YL, QG and RY designed the study. YL and QG performed the analysis. QG performed the experiments. YL and QG designed and implemented the model. YL built the command-line tool and documentation. YL, QG and RY wrote the manuscript. RY supervised this work.	931 932 933 934 935 936 937 938 939 940
Data Availability.	The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1354861. The dataset includes Oxford Nanopore long-read whole-genome sequencing of PC3 prostate cancer cells and MDA-amplified single-cell derivatives. The individual SRA accessions are as follows: PC3 bulk (MinION Mk1C), SRR35904028; PC3 bulk (PromethION P2), SRR35904029; PC3 10-cell WGA (MinION Mk1C), SRR35904026; PC3 10-cell WGA (PromethION P2), SRR35904027. We can access the data at the following link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA1354861?reviewer=viej6cv6mgbl3n7a9a5k1bsb3	941 942 943 944 945 946 947 948 949 950
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/ChimeraLM), allowing users to test DeepChopper'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 (https://ylab-hi.github.io/ChimeraLM/).	951 952 953 954 955 956 957 958
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.	959 960 961 962
Acronyms		963 964 965 966
CPU	Central Processing Unit	12
DEL	deletion	8, 9
dMDA	droplet-based MDA	12
DUP	duplication	8, 9
GLM	Genomic Language Model	2, 12
GPU	Graphics Processing Unit	12, 16, 18, 21
HPC	High Performance Computing	18

967 **INS** insertion 8, 9
968 **INV** inversion 1, 2, 7–10, 20
969
970 **LIANTI** Linear Amplification via Transposon Insertion 12
971
972 **MALBAC** Multiple Annealing and Looping-based Amplification Cycles 12
973 **MDA** Multiple Displacement Amplification 2
974 **MLP** multilayer perceptron 3, 4, 15
975
976 **ONT** Oxford Nanopore Technologies 3, 7, 8, 13, 17
977
978 **PacBio** Pacific Biosciences 3, 13
979 **PTA** Primary Template-directed Amplification 12
980
981 **SA** Supplementary Alignment 13, 17, 18
982 **SV** Structural Variation 1–5, 7–9, 11, 12, 14, 17, 18, 20
983
984 **TRA** translocation 2, 8
985
986 **WGA** Whole Genome Amplification 1–14, 17–20
987 **WGS** Whole Genome Sequencing 7, 8, 17
988
989

References

990 [1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using
991 different amplification methods and reference genomes. *Communications Biology*
992 1288 (2024).
993
994 [2] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**,
995 90–94 (2011).
996
997 [3] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons.
998 *Nature Communications* 4220 (2024).
999
1000 [4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state
1001 of the science. *Nature Reviews Genetics* 175–188 (2016).
1002
1003 [5] Chen, C. *et al.* Single-cell whole-genome analyses by linear amplification via
1004 transposon insertion (LIANTI). *Science (new York, N.Y.)* **356**, 189–194 (2017).
1005
1006 [6] Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives.
1007 *PLOS Genetics* **10**, e1004126 (2014).
1008
1009 [7] de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome
1010 amplification methods. *PLoS ONE* e105585 (2014).
1011
1012 [8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification
1013 commercial kits using targeted sequencing. *Scientific Reports* 17171 (2021).

- [9] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal* 1688–1696 (2023). 1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
- [10] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).
- [11] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displacement amplification with whole-genome oxford nanopore sequencing. *Molecular Ecology Resources* e14094 (2025).
- [12] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* **99**, 5261–5266 (2002).
- [13] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275 (2023).
- [14] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 461–468 (2018).
- [15] Smolka, M. *et al.* Detection of mosaic and population-level structural variants with sniffles2. *Nature Biotechnology* 1571–1580 (2024).
- [16] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nature Communications* 283 (2023).
- [17] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* 2907–2915 (2019).
- [18] Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* 189 (2020).
- [19] Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences* **118**, e2024176118 (2021).
- [20] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).
- [21] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011).
- [22] 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).

- 1059 [23] Li, Y. *et al.* A genomic language model for chimera artifact detection in nanopore
1060 direct rna sequencing. *bioRxiv* (2024). URL <https://www.biorxiv.org/content/early/2024/10/25/2024.10.23.619929>.
- 1062
- 1063 [24] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single*
1064 *nucleotide resolution*, Vol. 36, 43177–43201 (Curran Associates, Inc., 2023).
- 1065
- 1066 [25] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust
1067 foundation models for human genomics. *Nature Methods* 287–297 (2025).
- 1068
- 1069 [26] Zhou, Z. *et al.* *DNABERT-2: Efficient foundation model and benchmark for*
1070 *multi-species genomes*, 1–24 (OpenReview.net, 2024).
- 1071
- 1072 [27] Consens, M. E. *et al.* To transformers and beyond: Large language models for
1073 the genome (2023). [arXiv:2311.07621](https://arxiv.org/abs/2311.07621).
- 1074
- 1075 [28] Poli, M. *et al.* *Hyena hierarchy: Towards larger convolutional language models*,
1076 Vol. 202, 28043–28078 (PMLR, 2023).
- 1077
- 1078 [29] Mahmoud, M. *et al.* Structural variant calling: The long and the short of it.
1079 *Genome Biology* **20**, 246 (2019).
- 1080
- 1081 [30] Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-
1082 nucleotide and copy-number variations of a single human cell. *Science* 1622–1626
(2012).
- 1083
- 1084 [31] Dippenaar, A. *et al.* Droplet based whole genome amplification for sequencing
1085 minute amounts of purified mycobacterium tuberculosis DNA. *Scientific Reports*
1086 **14**, 9931 (2024).
- 1087
- 1088 [32] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023).
- 1089
- 1090 [33] Martin, M. Cutadapt removes adapter sequences from high-throughput sequenc-
1091 ing reads. *Embojournal* **17**, 10–12 (2011).
- 1092
- 1093 [34] Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*
1094 3094–3100 (2018).
- 1095
- 1096 [35] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* giab008
1097 (2021).
- 1098
- 1099 [36] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of
1100 long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
- 1101
- 1102 [37] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning*
1103 *library*, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019).
- 1104

- [38] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub repository (2019). URL <https://github.com/Lightning-AI/lightning>.
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
- [39] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).
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
- [40] Yadan, O. Hydra - a framework for elegantly configuring complex applications. GitHub repository (2019). URL <https://github.com/facebookresearch/hydra>.
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
- [41] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).
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
- [42] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* i333–i339 (2012).
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
- [43] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Research* 581–591 (2018).
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
- [44] Guo, Q., Li, Y., Wang, T.-Y., Ramakrishnan, A. & Yang, R. OctopusV and TentacleSV: A one-stop toolkit for multi-sample, cross-platform structural variant comparison and analysis. *Bioinformatics* btaf599 (2025).
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
- [45] 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).
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
- [46] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 261–272 (2020).
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [47] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 90–95 (2007).
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
- [48] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* 3021 (2021).
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150