

ChimeraLM detects amplification artifacts for
accurate structural variant calling in long-read
single-cell sequencing

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

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 enables unprecedented cellular heterogeneity insights but faces a fundamental challenge: Whole Genome Amplification (WGA) introduces chimeric artifacts that generate false Structural Variations (SVs), undermining biological interpretations. Current computational methods cannot distinguish amplification-induced artifacts from genuine rearrangements. Here we present ChimeraLM, a genomic language model that learns sequence-level features discriminating biological sequences from WGA artifacts. Validated on nanopore data, ChimeraLM achieves 95% recall with 70% precision and reduces chimeric reads by ~90% while preserving 72–92% of true SVs. This improves SV validation rates 8–11 fold and eliminates false-positive inversion (INV) bias, restoring SV distributions to bulk-like profiles. Attention visualization reveals ChimeraLM focuses on junction regions with single-base precision, learning interpretable features applicable across sequencing technologies. By enabling confident SV detection at single-cell resolution, ChimeraLM addresses a fundamental data quality barrier in cancer genomics, developmental biology, and precision medicine. Available at <https://github.com/ylab-hi/ChimeraLM>.

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

047 **Main**

048

049 Single-cell genomics has revolutionized our resolution of biological heterogeneity,
050 enabling the discovery of rare cell types and the reconstruction of clonal evolution in
051 cancer and development [1–3]. However, a single cell contains only 6–7 picograms of
052 DNA—approximately two genome copies—posing significant technical challenges for
053 comprehensive genomic analysis [4, 5]. Consequently, **WGA** remains an unavoidable
054 prerequisite, amplifying DNA 1,000- to 10,000-fold for high-coverage sequencing [6–
055 8]. While **WGA** provides necessary material for downstream analysis, it introduces
056 systematic errors that severely compromise genomic fidelity, particularly for **SV**
057 detection [9–11].

058 The most pernicious of these errors are chimera artifacts—artificial DNA con-
059 structs formed when highly processive polymerases, such as phi29 in **Multiple Dis-**
060 **placement Amplification (MDA)** [12], switch templates during amplification [9–11, 13].
061 These chimeras join discontinuous genomic loci into single molecules, mimicking the
062 structural signatures of biological **translocations (TRAs)** and **INVs** [10]. In long-read
063 sequencing, which is otherwise ideal for resolving complex **SVs**, chimeric reads can
064 constitute 42–76% of the **WGA** data [9], rendering standard **SV** callers unreliable [14–
065 19]. Because these tools rely on alignment heuristics and coverage deviations [14, 20],
066 they frequently misclassify artificial chimeras as genuine variants [21].

067 Distinguishing biological rearrangements from amplification artifacts remains a
068 major computational bottleneck. Current quality control methods rely on hand-
069 crafted features—such as read-pair orientation or localized coverage drops—that fail
070 to capture the sequence-intrinsic patterns of **WGA** errors [11, 13, 22]. This limitation
071 blocks the application of single-cell long-read sequencing in contexts where precision
072 is paramount, such as tracking somatic mosaicism or validating CRISPR off-target
073 effects.

074 We reasoned that **WGA** artifacts possess latent sequence motifs and structural
075 patterns distinct from genomic sequences, learnable without reliance on reference
076 alignment. Here, we present ChimeraLM, a platform-agnostic **Genomic Language**
077 **Model (GLM)** to identify and filter **WGA** artifacts with single-read resolution.

078 Leveraging advances in DNA foundation models [23–26], ChimeraLM treats arti-
079 fact detection as a sequence modeling task rather than an alignment problem. By
080 attending to long-range dependencies and contextual features within raw reads [23–
081 28], ChimeraLM achieves ~90% reduction in chimeric reads while preserving 72–92%
082 of true **SVs**. We demonstrate that this approach restores the fidelity of single-cell **SV**
083 calling, enabling robust characterization of genomic heterogeneity at the single-cell
084 level.

085

086 **Results**

087

088 **Overview of ChimeraLM workflow and model architecture**

089

090 Single-cell genomics relies on **WGA** to obtain sufficient DNA for sequencing (Fig. 1a).
091 The standard workflow includes single-cell isolation, DNA extraction, **WGA**, long-read
092 sequencing (e.g., **Oxford Nanopore Technologies (ONT)**), base calling, and alignment

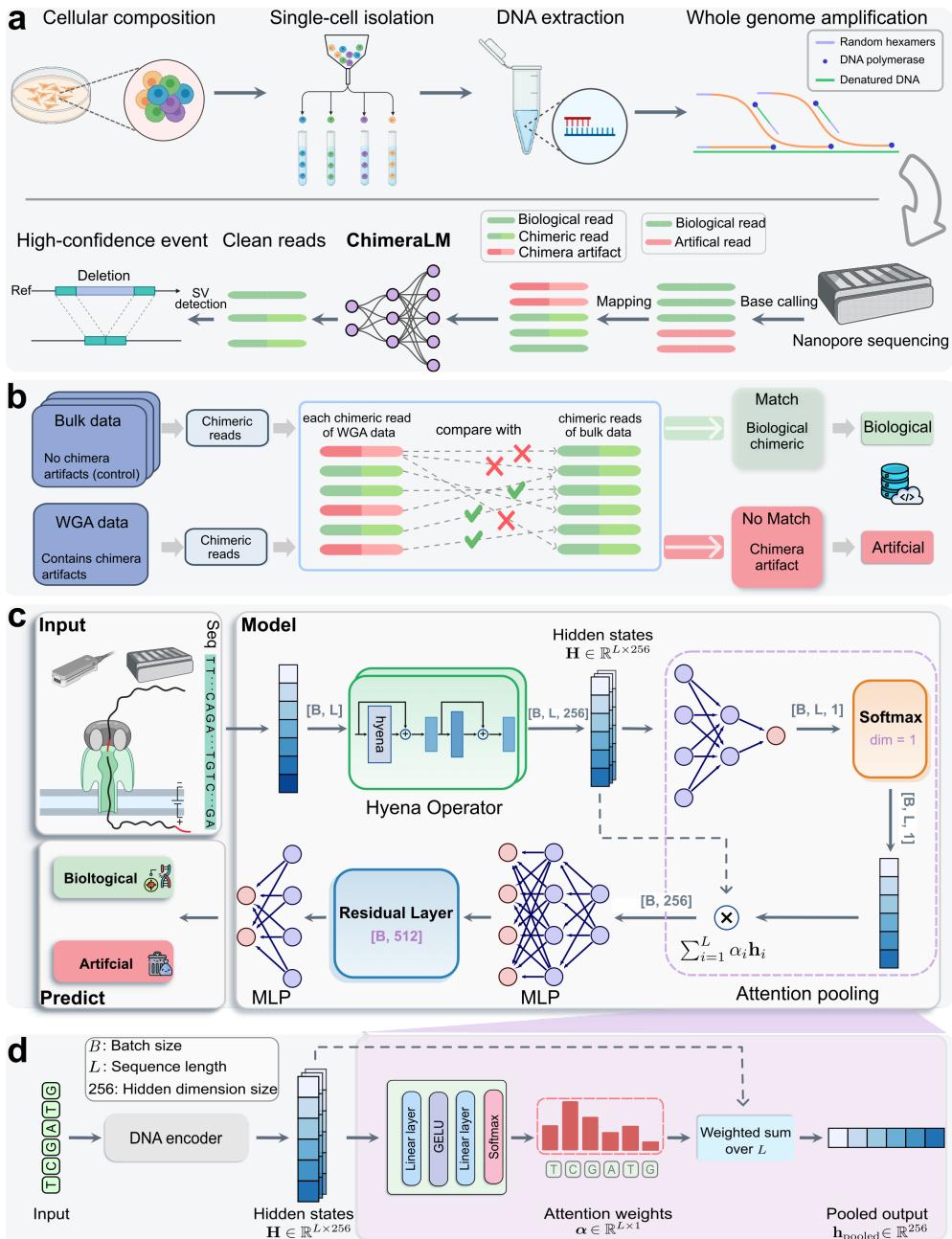


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA for genome amplification. WGA generates chimeric artifacts (red) through template switching during amplification, alongside biological reads (green). After nanopore sequencing, ChimeraLM classifies chimeric reads as biological or artificial, enabling downstream SV detection on clean reads. (b) Ground truth label generation for supervised learning. Chimeric reads from WGA data are compared against all chimeric reads from bulk sequencing data of the same cell line. Reads that match bulk data are labeled as biological (green pathway), while non-matching reads are labeled as chimera artifacts (red pathway). This provides reliable training labels. (c) ChimeraLM architecture. Input DNA sequences (batch size B , sequence length L) are tokenized and encoded into hidden states $\mathbf{H} \in \mathbb{R}^{L \times 256}$ through a DNA encoder (HyenaDNA [23]). Hyena operators capture long-range dependencies in genomic sequences. Attention pooling aggregates position-specific features using learned weights. Residual and multilayer perceptron (MLP) layers process pooled features, and a softmax layer outputs binary classification probabilities for biological versus artificial reads. (d) Attention pooling mechanism detail. The DNA encoder (HyenaDNA) transforms input sequences into hidden state $\mathbf{H} \in \mathbb{R}^{L \times 256}$. Attention weights $\alpha \in \mathbb{R}^{L \times 1}$ are computed through linear layers, GELU activation, and softmax normalization, assigning importance scores to each nucleotide position. The weighted sum $\mathbf{h}_{\text{pooled}} = \sum_{i=1}^L \alpha_i \mathbf{h}_i$ produces the pooled output $\mathbf{h}_{\text{pooled}} \in \mathbb{R}^{256}$, compressing variable-length sequences into fixed-dimensional representations. Created with BioRender.com.

093
094
095
096
097
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

139 to the reference genome. During amplification, template-switching events introduce
140 artificial chimeric reads, resulting in alignment files that contain a mixture of authentic
141 and artificial sequences. In downstream analysis, these artifacts can mimic **SV** and
142 confound variant detection. To address this challenge, we developed ChimeraLM, a
143 **GLM** designed to integrate directly into this analysis pipeline and distinguish biological
144 reads from amplification-induced artifacts.

145 ChimeraLM functions as a pre-processing filter, operating after read alignment
146 but before **SV** detection. It evaluates each chimeric read—sequences with multiple
147 alignments to distant genomic locations—and classifies it as either biological (genuine)
148 or artificial (**WGA**-induced). This binary decision enables the retention of authentic
149 genomic sequences while removing amplification artifacts prior to variant calling. The
150 resulting high-confidence biological reads are then passed to conventional **SV** detection
151 algorithms for accurate identification of genomic rearrangements.

152 A high-confidence labeled dataset was required for supervised training of the model
153 (Fig. 1b; Extended Data Fig. 1a). We constructed this dataset using sequencing data
154 from the PC3 prostate cancer cell line, which provides both **WGA**-amplified and
155 non-amplified (bulk) genomic data. The key assumption is that bulk sequencing con-
156 tains only genuine genomic sequences, whereas **WGA** data includes both genuine and
157 artificial chimeras. Chimeric reads from the PC3 **WGA** PromethION dataset were sys-
158 tematically compared against three independent bulk datasets (**ONT** PromethION,
159 **ONT** MinION, and PacBio; see Methods). **WGA** reads whose chimeric structures were
160 absent from all three bulk datasets were labeled artificial. Conversely, **WGA** reads
161 with structures validated in one or more bulk datasets were labeled biological.

162 Application of this labeling strategy to the PC3 **WGA** PromethION data
163 (Extended Data Table 1) quantified the read distribution across these categories
164 (Extended Data Fig. 1b). We identified 12,670,396 chimeric reads with zero matches
165 in the bulk reference, which were classified as artificial. Conversely, we identified a
166 total of 293,180 reads validated as biological. This biological set was composed of reads
167 matching one (Match 1: 101,094 reads), two (Match 2: 190,309 reads), or all three
168 (Match 3: 1,777 reads) of the bulk reference datasets. To construct a balanced train-
169 ing dataset, we retained all 293,180 biological reads (combining Match 1, 2, and 3)
170 and subsampled an equal number (293,180) of artificial reads from the no-match cat-
171 egory (Extended Data Fig. 1b). This set was augmented with 178,748 chimeric reads
172 subsampled from the bulk datasets as positive controls. The final dataset of 765,108
173 labeled reads was partitioned into training (70%), validation (20%), and internal test
174 (10%) sets using stratified splitting (Extended Data Fig. 1a).

175 The architecture of ChimeraLM (Fig. 1c) was specifically designed to learn
176 from this dataset by operating directly on raw DNA sequences, bypassing conven-
177 tional, feature-based classifiers. This design must address three primary technical
178 challenges: (1) efficiently processing variable-length sequences of many kilobases,
179 (2) simultaneously maintaining single-nucleotide resolution to detect the precise,
180 abrupt compositional changes that define chimeric junctions, and (3) aggregating
181 variable-length sequence representations into a consistent classification output.

182 ChimeraLM first addresses the need for high resolution by tokenizing input
183 sequences at the single-nucleotide level. This base-pair precision is required to preserve
184

the complete sequence information necessary for detecting chimeric junctions—the
breakpoints where disparate genomic regions are artificially fused and which often
exhibit abrupt compositional changes. The architecture’s core employs Hyena oper-
ators [29], selected specifically to overcome the challenge of processing long DNA
sequences. Traditional attention mechanisms scale quadratically with sequence length,
making them computationally prohibitive for long-read data. Hyena operators, by
contrast, achieve subquadratic scaling, enabling ChimeraLM to analyze full-length
reads without fragmentation and thus preserve the structural context around chimeric
junctions. To leverage existing genomic knowledge, we initialized the model with
weights from HyenaDNA [23], a genomic foundation model pre-trained on diverse
DNA sequences.

Finally, to produce a classification, the model employs an attention pooling mech-
anism to aggregate information across the entire variable-length read (Fig. 1d). This
module computes learned, position-specific weights to identify which nucleotides—such
as those at the junction boundary—are most informative for the classification deci-
sion. This weighted aggregation produces a fixed-dimensional representation, which
is then processed through MLP components with residual connections. A final soft-
max layer outputs the probability scores for the biological versus artificial classes (see
[Methods](#)). This end-to-end architecture enables ChimeraLM to learn directly from
raw sequence data, discovering complex patterns that may not be apparent through
rule-based algorithms.

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

We first evaluated ChimeraLM’s classification accuracy on the held-out test set
(derived from the PromethION training data), which comprised reads with known
biological or artificial status (Fig. 2a). The model achieved an F1 score of 0.81, reflect-
ing balanced sensitivity and specificity in artifact detection. A recall of 0.95 indicates
that 95% of true chimeric reads were correctly identified—critical for minimizing
downstream false-positive **SV** calls—while a precision of 0.70 shows that the majority
of reads flagged as chimeric were true artifacts. These results establish the model’s
reliability for identifying amplification-induced artifacts in long-read data.

We next assessed its practical effectiveness on the full PC3 **WGA** datasets, com-
paring performance on the PromethION and MinION platforms (Fig. 2b). Bulk
sequencing established a low baseline chimeric read rate (2.3% for PromethION; 2.5%
for MinION). **WGA** dramatically increased this artifact load to 46.0% (PromethION)
and 23.0% (MinION). After ChimeraLM filtering, chimeric content dropped to 4.9%
on PromethION and 1.5% on MinION—representing 10- to 15-fold reductions—while
retaining 15.8 million and 5.6 million biological reads. This restoration to near-bulk
quality demonstrates that ChimeraLM effectively separates genuine genomic reads
from **WGA**-induced artifacts.

We then benchmarked ChimeraLM against existing computational tools for
detecting amplification-induced chimeras, SACRA [22] and 3rd-ChimeraMiner [13]
(Fig. 2c,d). When applied to the same PromethION and MinION **WGA** data,

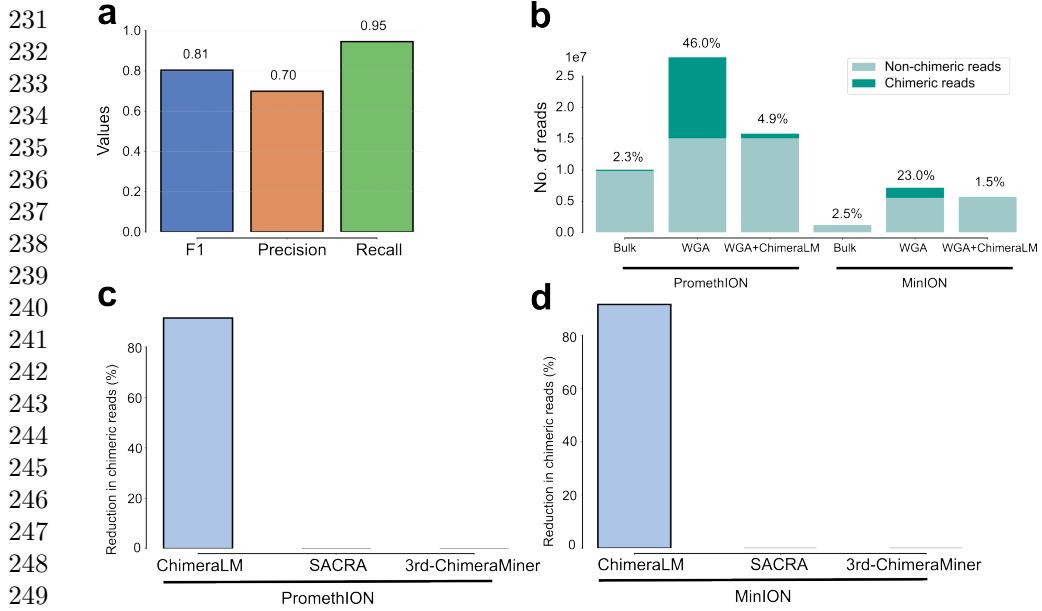


Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts.
 (a) Classification performance on held-out test data. ChimeraLM achieves high recall (0.95) in identifying chimera artifacts while maintaining acceptable precision (0.70), yielding an F1 score of 0.81 for binary classification of biological versus artificial sequences. (b) Chimeric read reduction across sequencing platforms. Stacked bars show the proportion of chimeric (dark teal) and non-chimeric (light teal) reads in bulk sequencing, WGA-amplified samples, and ChimeraLM-filtered WGA samples. Data from PC3 cell line sequenced on PromethION (left) and MinION (right) platforms demonstrate that ChimeraLM reduces chimeric read frequencies from 46.0% to 4.9% (PromethION) and from 23.0% to 1.5% (MinION), approaching bulk levels (2.3% and 2.5%, respectively). (c,d) Benchmarking against existing methods. ChimeraLM achieves approximately 90% reduction in chimeric reads on both PromethION (c) and MinION (d) platforms, whereas existing computational tools SACRA and 3rd-ChimeraMiner show no detectable reduction in chimeric content.

ChimeraLM achieved an approximately 90% reduction in chimeric reads on both platforms. In stark contrast, neither SACRA nor 3rd-ChimeraMiner showed any detectable reduction in chimeric content (0% reduction).

Together, these results demonstrate robust and platform-agnostic performance. The strong filtering on the MinION dataset (Fig. 2b) is particularly noteworthy, as this platform served as a completely independent test set—the model was trained exclusively on PromethION data yet generalized effectively to MinION. This cross-platform generalization, combined with the high recall on the internal test set (Fig. 2a) and clear superiority over existing tools (Fig. 2c,d), confirms that ChimeraLM learns universal sequence-level features of WGA-induced artifacts rather than platform-specific technical signatures. This design principle—learning from DNA sequence patterns that are invariant across sequencing technologies—suggests ChimeraLM’s applicability extends beyond nanopore platforms to other long-read and short-read sequencing technologies.

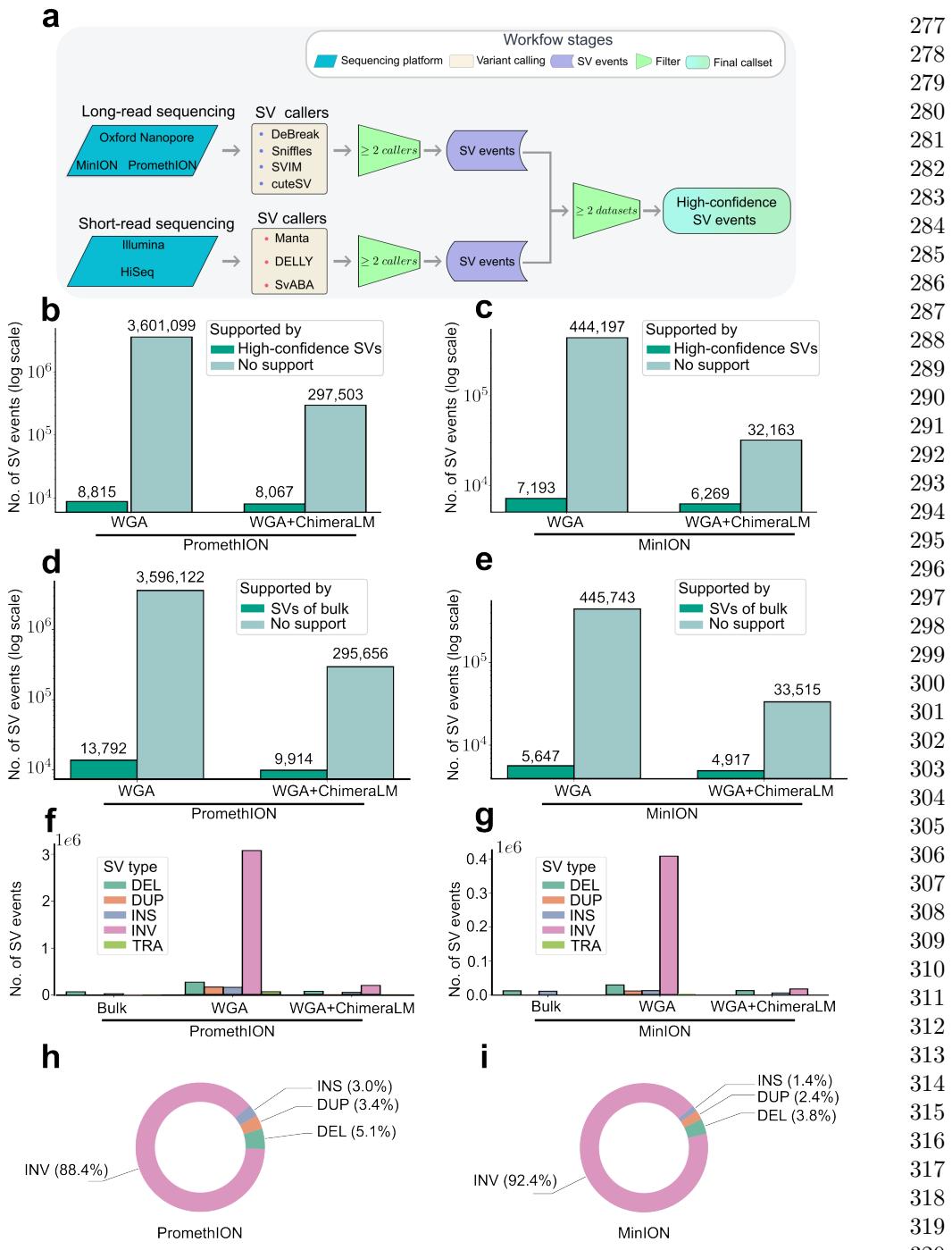


Fig. 3 ChimeraLM improves structural variant detection accuracy. (a) Construction of high-confidence SV reference dataset. PC3 bulk DNA was sequenced on multiple platforms (ONT PromethION and MinION, Illumina HiSeq) and analyzed with multiple SV calling algorithms. SV events detected by ≥ 2 callers on the same platform were retained. Events supported by both long-read and short-read platforms were designated as high-confidence gold standard SVs. (b,c) SV validation against multi-platform gold standard. Stacked bars show total SV calls (log scale, numbers above bars) classified as gold standard-supported (dark teal) or unsupported (light teal) for PromethION (b) and MinION (c). ChimeraLM substantially reduces unsupported SV calls while preserving gold standard events. (d,e) SV validation against long-read bulk sequencing (ONT PromethION and MinION). Stacked bars show SV calls classified as bulk-supported (dark teal) or unsupported (light teal) for PromethION (d) and MinION (e). Long-read bulk data from the same platform provides platform-matched validation, capturing true variants that may be specific to long-read detection. (f,g) SV type distribution across processing methods. Bar charts show the number of detected SVs by type: deletion (DEL) (green), duplication (DUP) (orange), insertion (INS) (blue), INV (pink), and TRA (light green) for PromethION (f) and MinION (g). Unfiltered WGA data shows elevated counts across all types, particularly INVs and TRAs, which are reduced to bulk-like levels after ChimeraLM filtering. (h,i) Composition of chimeric artifact-supported SVs. Pie charts show the proportion of SV types among events supported exclusively by reads classified as chimeric artifacts in unfiltered WGA data for PromethION (h) and MinION (i). These represent false-positive SV calls that would be eliminated by ChimeraLM.

277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322

323 **ChimeraLM substantially reduces false-positive structural
324 variant calls**

325 Accurate **SV** detection is essential for understanding genomic diversity and dis-
326 ease mechanisms in single cells. 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 accuracy, we compared variant calls from unfiltered
329 **WGA** data and ChimeraLM-filtered data against two independent reference standards
330 (Fig. 3).

331 We first established a high-confidence gold standard **SV** dataset by integrating
332 results from bulk PC3 DNA sequenced on multiple platforms (**ONT** PromethION,
333 **ONT** MinION, and Illumina HiSeq) and analyzed with multiple **SV** callers (Fig. 3a;
334 Extended Data Table 1). **SVs** detected by ≥ 2 callers on the same platform and sup-
335 ported by both long-read and short-read data were retained as gold-standard events,
336 ensuring high specificity across technologies.

337 Comparison against this gold standard revealed that unfiltered **WGA** data con-
338 tained extensive false-positive **SVs** (Fig. 3b,c). On PromethION, raw **WGA** data
339 produced 3.6 million **SV** calls, of which only 8,815 (0.24%) matched gold standard
340 events—indicating that over 99% were artifacts. After ChimeraLM filtering, total calls
341 dropped to 305,570 while retaining 8,067 true events, raising the validation rate to
342 2.64% (11-fold improvement) and preserving 91.5% of true variants. MinION data
343 showed similar results, with calls reduced from 451,390 to 38,432 and the validation
344 rate increasing from 1.59% to 16.3% (10-fold improvement) while retaining 87.2% of
345 true variants. These results highlight ChimeraLM’s ability to remove spurious **SV** calls
346 while maintaining biological sensitivity.

347 To complement this stringent validation, we next performed platform-matched bulk
348 validation, comparing **WGA**-derived **SV** calls against long-read bulk sequencing from
349 the same platform (Fig. 3d,e). This reference captures true **SVs** that may be missed by
350 short-read data, providing a more inclusive measure of recall. Under this benchmark,
351 ChimeraLM increased validation rates from 0.38% to 3.24% on PromethION (8.5-
352 fold improvement) and from 1.25% to 12.79% on MinION (10-fold improvement),
353 while retaining 71.9% and 87.1% of bulk-supported events, respectively. The consistent
354 improvements across independent datasets demonstrate that ChimeraLM effectively
355 suppresses **WGA**-induced artifacts without sacrificing detection of genuine **SVs**.

356 Together, these analyses demonstrate that ChimeraLM reduces false-positive **SV**
357 calls by 8–11 fold while preserving 72–92% of true variants, resulting in a substantial
358 enhancement of the signal-to-noise ratio in single-cell **SV** discovery. By restoring near-
359 bulk specificity and maintaining robust sensitivity, ChimeraLM enables more accurate
360 and interpretable downstream genomic analyses.

361
362 **ChimeraLM restores unbiased SV-type distributions and
363 characterizes artifact composition**

364 Amplification artifacts can distort the apparent spectrum of **SVs**, often inflating spe-
365 cific **SV** types. To evaluate whether ChimeraLM effectively corrects such distortions,
366
367

we compared **SV** type distributions across bulk, unfiltered **WGA**, and ChimeraLM-filtered datasets (Fig. 3f,g). Bulk sequencing showed relatively balanced proportions of **DELs**, **DUPs**, **INSSs**, **INVs**, and **TRAs**. In contrast, unfiltered **WGA** data exhibited a dramatic overrepresentation of **INVs** on both PromethION and MinION platforms, consistent with pervasive amplification artifacts. After ChimeraLM filtering, these distributions were largely restored toward bulk-like profiles: excessive **INVs** were markedly reduced while other **SV** categories remained stable. This shift reflects selective removal of artifact-supported **INVs** rather than indiscriminate loss of genuine inversion signals, demonstrating high specificity in distinguishing chimeric from biological reads.

To investigate the basis of this normalization, we analyzed **SV** calls supported exclusively by reads classified as chimeric by ChimeraLM (Fig. 3h,i). These artifact-supported events were overwhelmingly dominated by **INVs**, comprising 88.4% on PromethION and 92.4% on MinION. This pattern is consistent with template-switching junctions that produce inversion-like alignment signatures. Smaller fractions of **DELs** (5.1% and 3.8%), **DUPs** (3.4% and 2.4%), and **INSSs** (3.0% and 1.4%) were also observed, demonstrating that **WGA**-induced chimeras can mimic diverse **SV** categories rather than only **INVs**.

This characterization has important implications for single-cell genomics. Although **INVs** are the predominant artifact type, the coexistence of **DELs**, **DUPs**, and **INSSs** among chimeric events indicates that comprehensive filtering—rather than inversion-specific correction—is essential for accurate **SV** detection. Without ChimeraLM filtering, single-cell **SV** analyses would be confounded not only by false-positive **INVs** but also by other artifact-associated variants [21, 30]. By restoring biologically representative **SV** type distributions, ChimeraLM enables robust and interpretable characterization of structural variation in single cells without distortion from **WGA**-induced artifacts.

369
370
371
372
373
374
375
376
377
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

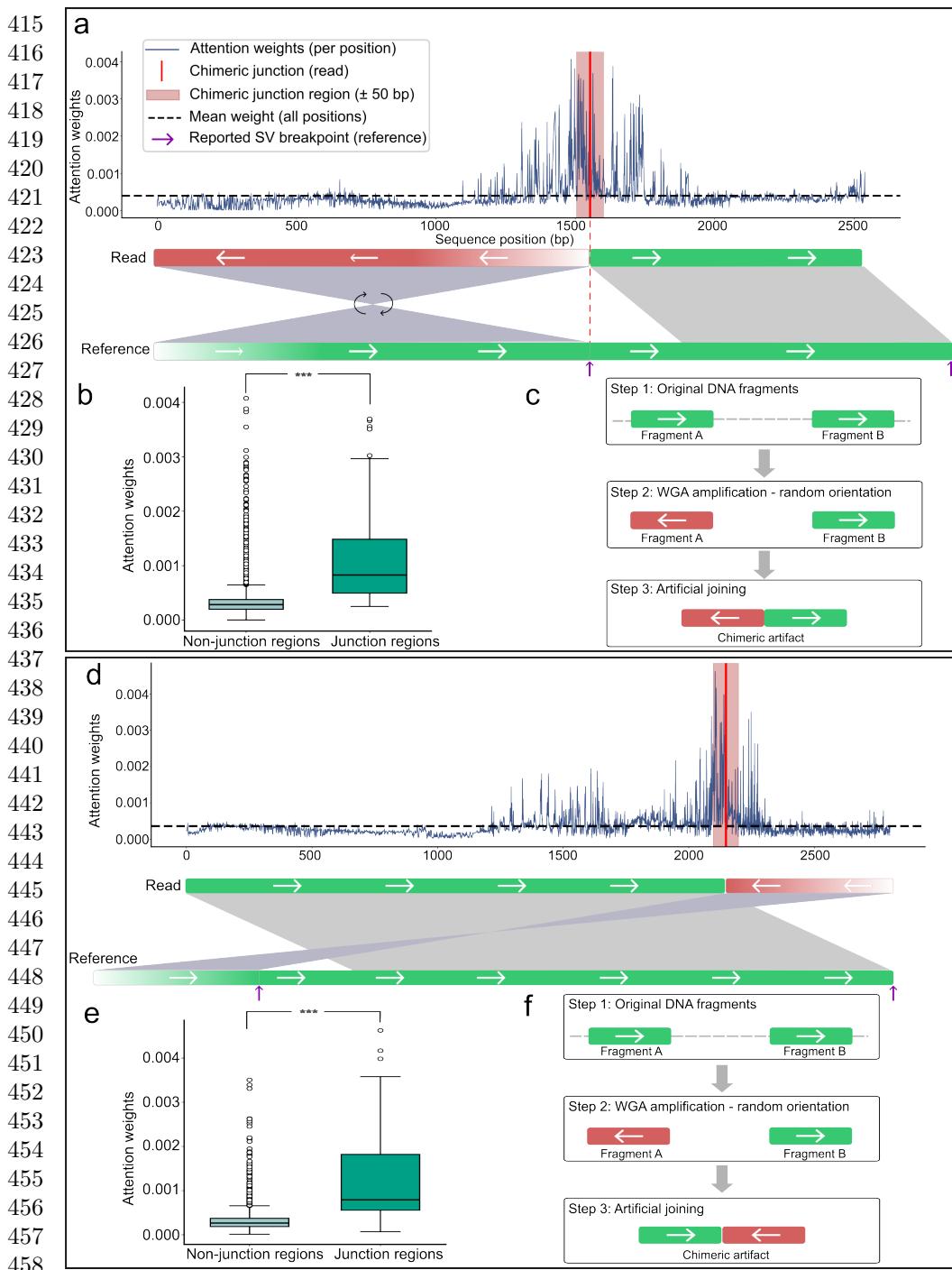


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions.

(a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per sequence position (blue line) and mean attention (dashed line). Red vertical lines mark chimeric junction positions, with pink shading indicating junction region (± 50 bp). Purple arrows show reported SV breakpoints. Lower panels illustrate read alignments: reads (top bars) show orientation transitions at junctions (green = forward, red = reverse-complemented, arrows indicate strand), while reference genome (bottom bars) maintains continuous forward orientation. Gray regions connect aligned segments. (b,e) Quantitative attention analysis. Box plots show significantly elevated attention weights in junction region versus non-junction regions for both examples ($p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$, respectively; Wilcoxon rank-sum test). (c,f) Proposed chimera formation mechanisms. Step 1: Original DNA fragments from distant genomic loci exist in forward orientation. Step 2: During WGA, one or both fragments may undergo random reverse-complementation. Step 3: Template switching joins the fragments with discordant orientations, creating chimeric artifacts. The two examples illustrate different orientation patterns (forward-to-reverse vs reverse-to-forward transitions) arising from random strand selection during amplification.

ChimeraLM provides interpretable classification through attention visualization	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
We next investigated whether ChimeraLM’s attention mechanism highlights biologically meaningful regions within sequencing reads (Fig. 4).	
For representative chimeric reads, attention weight profiles showed low baseline values across most positions but pronounced peaks at junction regions where template switching artificially joins DNA fragments from distinct genomic loci (Fig. 4a,d). These peaks coincided precisely with alignment breakpoints characterized by orientation changes between adjacent read segments—the defining signature of WGA-induced chimeric artifacts.	
Quantitative analysis confirmed that attention weights within junction regions (± 50 bp) were significantly higher than those in non-junction regions (Wilcoxon rank-sum test, $p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$) (Fig. 4b,e). Such localization indicates that ChimeraLM learns mechanistically relevant features associated with artificial junction formation rather than relying on spurious correlations.	
Schematic reconstruction of the amplification process further supports this interpretation (Fig. 4c,f). During WGA, DNA fragments from distant genomic loci may undergo random strand orientation changes before being joined by template switching. This process produces artificial junctions with discordant orientations—forward-to-reverse or reverse-to-forward—that generate inversion-like alignment signatures and are effectively recognized by the model’s attention peaks.	
Together, these analyses demonstrate that ChimeraLM’s attention mechanism can localize chimeric junctions at single-base resolution and capture the underlying orientation discontinuities that define WGA-induced artifacts.	
Discussion	488
WGA has enabled genomic analysis from single cells but introduces chimeric artifacts that compromise SV detection. ChimeraLM addresses this challenge through sequence-level classification of biological versus artificial reads, substantially improving SV calling accuracy before downstream analysis. This upstream filtering strategy—removing problematic sequences at the read level rather than correcting errors post hoc—provides a practical solution for single-cell genomics laboratories.	489
Our results demonstrate several key advantages of ChimeraLM for long-read single-cell sequencing. The method achieves approximately 90% reduction in chimeric reads across nanopore platforms while retaining 72–92% of true SVs. It reduces false-positive SV calls by 8–11 fold, enabling researchers to focus on biologically relevant variants without manually filtering thousands of artifacts. Moreover, ChimeraLM performs consistently across PromethION and MinION without platform-specific retraining, indicating that it captures generalizable sequence features of WGA-induced chimeras. These results underscore the model’s robustness across diverse datasets and sequencing conditions.	490
ChimeraLM’s effectiveness reflects the ability of deep learning models to capture complex sequence patterns that are difficult to encode in rule-based filters. Traditional quality control methods rely on predefined metrics such as mapping quality or	491

507 read depth [13, 22], which may not effectively distinguish chimeric artifacts from bio-
508 logical reads. By learning directly from sequence data, ChimeraLM discovers subtle
509 compositional and structural features that differentiate authentic genomic sequences
510 from amplification artifacts. Furthermore, the model offers interpretability through
511 attention visualization, allowing researchers to examine which sequence regions drive
512 classification. Attention weights can concentrate sharply at junctions where template
513 switching joins DNA fragments from distinct loci, matching the known mechanism of
514 chimera formation. Some reads show more diffuse attention distributions, suggesting
515 that ChimeraLM integrates multiple complementary cues—such as junction orienta-
516 tion, compositional biases, and local sequence context—to classify diverse artifact
517 types. This interpretability builds confidence in the model’s predictions and provides
518 a lens for probing the molecular processes underlying amplification-induced artifacts.

519 The improved reliability of **SV** detection has direct implications for single-cell
520 genomics. Studies of chromosomal instability, clonal evolution, and **SV** burden in
521 individual cells have long been constrained by high false-positive rates in **WGA**
522 data [21, 30]. ChimeraLM enables more confident identification of genuine **SVs**, sup-
523 porting research in cancer genomics, developmental biology, and aging where single-cell
524 resolution is essential for understanding cellular heterogeneity. Although the current
525 model processes reads independently, integrating additional contextual features—
526 such as coverage, mate-pair, or phasing information—could further enhance accuracy.
527 **Graphics Processing Unit (GPU)** resources are recommended for large-scale datasets,
528 while **Central Processing Unit (CPU)** inference remains feasible for smaller studies;
529 runtime optimization and model compression may improve accessibility for broader
530 use.

531 Future work should prioritize validation across diverse biological and technical con-
532 texts. First, testing on multiple cell types (primary, stem, or immune cells) and **WGA**
533 protocols (**Multiple Annealing and Looping-based Amplification Cycles (MALBAC)**,
534 **Linear Amplification via Transposon Insertion (LIANTI)**, **Primary Template-directed**
535 **Amplification (PTA)**) will establish biological generalizability. Second, validation on
536 additional sequencing platforms—including PacBio HiFi, Illumina linked-reads, and
537 emerging long-read technologies—will confirm the platform-agnostic design princi-
538 ple. The sequence-level approach suggests ChimeraLM should transfer effectively
539 to any platform, though platform-specific fine-tuning may optimize performance.
540 Third, the interpretability of attention-based models could be leveraged to investi-
541 giate mechanisms of chimera formation: large-scale analysis of attention patterns
542 may reveal recurrent sequence motifs or genomic contexts associated with template
543 switching, guiding the development of improved amplification protocols. More broadly,
544 ChimeraLM illustrates the potential of **GLMs** for data quality control applications [23].
545 Architectural innovations such as the **Hyena** operator for efficient long-range mod-
546 eling [29] may have utility beyond chimera detection, addressing challenges such as
547 contamination, adapter artifacts, and systematic sequencing errors across multiple
548 platforms.

549 Looking ahead, ChimeraLM’s framework could extend beyond single-cell genomics
550 to address quality control challenges in other amplification-dependent technologies,
551 including cell-free DNA analysis, ancient DNA studies, and metagenomic sequencing
552

from low-biomass samples. The model's interpretability through attention visualization also opens opportunities for mechanistic studies of polymerase fidelity and template-switching dynamics across different amplification protocols. Furthermore, integration with emerging single-cell multi-omics platforms could enable simultaneous quality control across genomic, transcriptomic, and epigenomic data layers, providing a unified framework for artifact detection in complex single-cell experiments.	553 554 555 556 557 558
ChimeraLM thus provides a practical and interpretable framework for improving long-read single-cell genomic data quality. By removing WGA-induced chimeric artifacts at the read level and revealing the mechanistic features that drive them, the method not only enhances SV detection reliability but also deepens understanding of amplification-induced bias in single-cell genomics.	559 560 561 562 563 564 565 566
Methods	567 568
Cell culture, single-clone preparation, and nanopore sequencing	569
<i>Cell culture and single-clone establishment</i>	570 571 572 573 574 575 576
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.	577 578 579 580 581 582 583 584 585 586 587
<i>DNA extraction and whole-genome amplification</i>	577
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.	578 579 580 581 582 583 584 585 586 587
<i>Nanopore library preparation and sequencing</i>	588
Sequencing libraries were prepared using the ONT Ligation Sequencing Kit V14 (SQK-LSK114) and sequenced on MinION Mk1C or PromethION P2 Solo devices with R10.4.1 flow cells according to the manufacturer's genomic DNA workflow. Because all single-cell samples originated from the same monoclonal lineage, observed differences between amplified and bulk data primarily reflect MDA-induced artifacts rather than biological variation, providing a controlled experimental setting for downstream analyses.	589 590 591 592 593 594 595 596 597 598

599 ***Basecalling and read processing***
600 Raw signal files (POD5) were basecalled using Dorado v0.5.0 with the high-accuracy
601 model `dna_r10.4.1_e8.2_400bps_hac@v4.3.0` [31]. Reads with mean quality < 10
602 or length < 500 bp were removed. Residual adapters and concatemers were trimmed
603 using Cutadapt v4.0 [32] in two-pass error-tolerant mode. Cleaned reads were aligned
604 to the GRCh38.p13 reference genome using minimap2 v2.26 (`map-ont` preset) [33].
605 Resulting BAM files were sorted and indexed with SAMtools v1.16 [34]. Read length
606 and mapping statistics were calculated using NanoPlot v1.46.1 [35]. All samples were
607 processed under identical parameters to ensure consistency across datasets.
608

609 ***Chimeric read identification***
610 Chimeric reads were identified based on the presence of supplementary alignments in
611 BAM files using the **Supplementary Alignment (SA)** tag. The **SA** tag indicates that
612 a read has additional alignments beyond the primary alignment, which is character-
613 istic of chimeric sequences that map to multiple distant genomic locations. To ensure
614 accurate identification, we applied stringent filtering criteria: reads were classified as
615 chimeric only if they (1) were not unmapped, (2) contained the **SA** tag, (3) were not
616 secondary alignments, and (4) were not supplementary alignments themselves. This
617 filtering approach ensures that only primary alignments with supplementary mapping
618 evidence are considered chimeric, avoiding double-counting of the same chimeric event
619 and excluding low-quality or ambiguous alignments. Reads without the **SA** tag (single
620 continuous alignments) were classified as non-chimeric. This approach leverages the
621 standard BAM format specification to reliably identify reads with complex alignment
622 patterns.
623

624 **Training data construction**
625

626 ***Data generation and sources***
627 To construct the training dataset, we generated **WGA** and bulk sequencing data from
628 PC3 cells. The **WGA** sample was amplified and sequenced on the PromethION P2 plat-
629 form (**ONT**), while three independent bulk datasets were produced from non-amplified
630 genomic DNA: bulk PromethION P2, bulk MinION Mk1c (**ONT**), and bulk PacBio.
631 These bulk datasets represent authentic biological sequences free from amplification-
632 induced artifacts. In contrast, **WGA** sequencing includes both genuine genomic reads
633 and artificial chimeras introduced during the amplification process. An additional
634 **WGA** dataset sequenced on the MinION Mk1c platform was reserved exclusively as
635 an independent test set for cross-platform evaluation.
636

637 ***Ground truth annotation and class definition***
638 Ground truth labels were established by systematically comparing chimeric reads from
639 the **WGA** PromethION P2 dataset against those from the three bulk datasets. For
640 each **WGA** chimeric read, all alignment segments—defined by their genomic start
641 and end coordinates—were compared to the corresponding segments of bulk chimeric
642 reads. A **WGA** read was labeled as biological if every segment matched at least one
643 bulk chimeric read within a 1 kb positional tolerance, indicating that the structural
644

configuration is also present in non-amplified DNA. Reads lacking any matching pattern across all bulk datasets were labeled as artificial chimeras, presumed to arise from the amplification process. To ensure balanced class representation, additional chimeric reads were randomly sampled from the bulk datasets and labeled as biological, as these reads originate from genuine genomic rearrangements such as true SVs. The final labeled dataset combined the annotated WGA PromethION P2 reads with the subsampled bulk chimeric reads and was subsequently partitioned into training, validation, and test sets as described below.

Dataset partitioning and cross-platform validation

The combined labeled dataset, derived from WGA PromethION P2 and bulk sequencing data, was divided into training (70%), validation (20%), and internal test (10%) sets using stratified random sampling to maintain class balance. These subsets were used respectively for model training, hyperparameter tuning, and performance evaluation on data from the same sequencing platform.

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

Model architecture

DNA encoder

ChimeraLM employs the pre-trained HyenaDNA model [23] 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 [29] 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:

$$\mathbf{e} = \text{GELU}(\text{Linear}_{256 \rightarrow 256}(\mathbf{H})) \in \mathbb{R}^{L \times 256}$$

691 $\mathbf{s} = \text{Linear}_{256 \rightarrow 1}(\mathbf{e}) \in \mathbb{R}^{L \times 1}$
 692 $\boldsymbol{\alpha} = \text{softmax}(\mathbf{s}) \in \mathbb{R}^{L \times 1}$
 693
 694 The pooled representation is the weighted sum of hidden states:
 695
 696
$$\mathbf{h}_{\text{pooled}} = \sum_{i=1}^L \alpha_i \mathbf{h}_i \in \mathbb{R}^{256}$$

 697
 698
 699
 700 This mechanism assigns learned importance weights to each sequence position,
 701 enabling the model to focus on informative regions while accommodating natural
 702 variability in read lengths.
 703
 704 ***Classification head***
 705 The pooled representation is processed through a **MLP** with residual connections. The
 706 first layer expands dimensionality:
 707
 708 $\mathbf{f}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{Linear}_{256 \rightarrow 512}(\mathbf{h}_{\text{pooled}}))) \in \mathbb{R}^{512}$
 709
 710 Subsequent residual blocks with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ compute:
 711
 712
$$\mathbf{f}_{\text{out}} = \text{Dropout}_{0.1}(\text{Linear}_{512 \rightarrow 512}(\text{GELU}(\text{Linear}_{512 \rightarrow 512}(\mathbf{f}_{\text{in}})))) + \mathbf{f}_{\text{in}}$$

 713
 714 where the skip connection enables stable gradient flow during training. The final layer
 715 produces binary classification logits:
 716
 717
$$\mathbf{z} = [z_0, z_1] = \text{Linear}_{512 \rightarrow 2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2$$

 718
 719 where z_0 and z_1 represent logits for biological and artificial chimeric classes, respec-
 720 tively. During inference, the predicted class is $\hat{y} = \text{argmax}_{i \in \{0,1\}} z_i$.
 721
 722 ***Model summary***
 723 The complete ChimeraLM pipeline processes DNA sequences through: (1) single-
 724 nucleotide tokenization, (2) HyenaDNA backbone encoding to generate contextualized
 725 representations, (3) attention pooling to aggregate position-specific features, (4) **MLP**
 726 layers with residual connections to learn classification features, and (5) binary classi-
 727 fication output. The entire model is trained end-to-end using labeled **WGA** and bulk
 728 sequencing data.
 729
 730 **Model training and optimization**
 731
 732 ***Training configuration***
 733 ChimeraLM was trained using PyTorch [36] and PyTorch Lightning [37] frameworks.
 734 Input sequences were tokenized using the tokenizer with maximum sequence length of
 735 32,768 bp. Sequences longer than this threshold were truncated; shorter sequences were
 736

padded to enable batch processing. Training employed mixed-precision computation (bf16) to accelerate training while maintaining numerical stability. 737
738
739

Optimization procedure

We used the AdamW optimizer [38] with learning rate $\eta = 1 \times 10^{-4}$ and weight decay $\lambda = 0.01$. AdamW implements adaptive learning rates with decoupled weight decay, combining the benefits of Adam optimization with proper L2 regularization. A ReduceLROnPlateau scheduler dynamically adjusted the learning rate based on validation loss, reducing it by a factor of 0.1 when no improvement occurred for 10 consecutive epochs. Early stopping with patience of 10 epochs prevented overfitting by terminating training when validation performance plateaued. A fixed random seed (12345) ensured reproducibility across training runs. 740
741
742
743
744
745
746
747
748
749
750
751

The training objective used cross-entropy loss for binary classification. For a training example with true class label $y \in \{0, 1\}$ and model logits $\mathbf{z} = [z_0, z_1]$, the loss is: 752
753

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

where z_0 and z_1 represent logits for biological and artificial chimeric classes, respectively. 754
755
756

Training implementation

Training used batch size of 16 sequences with 30 parallel data loading workers. GPU acceleration was employed for efficient processing, with training typically requiring 96–120 hours depending on dataset size. Model checkpointing saved the best-performing model based on validation metrics. Configuration management used Hydra [39] to enable reproducible experimentation. 757
758
759
760
761
762
763

Model evaluation

Performance was monitored using accuracy, precision, recall, and F1 score on the validation set after each epoch: 764
765
766
767

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

where TP (true positives) are chimeric reads correctly classified as artificial, TN (true negatives) are biological reads correctly classified as biological, FP (false positives) are biological reads misclassified as artificial, and FN (false negatives) are chimeric reads misclassified as biological. Final model selection was based on best validation performance as determined by early stopping. 768
769
770
771
772
773
774
775
776
777
778
779
780
781
782

783 **Model inference and application**

784 *Inference pipeline*

785
786 To apply ChimeraLM to new **WGA** sequencing data, the model takes a BAM file as
787 input. Chimeric reads are identified using **SA** tags and filtered to exclude unmapped,
788 secondary, or supplementary alignments. Each chimeric read sequence is tokenized
789 using the tokenizer (maximum length 32,768 bp, with truncation or padding as
790 needed). The trained model processes sequences in batches, generating two logits
791 $[z_0, z_1]$ for each read corresponding to biological and artificial chimeric classes. Clas-
792 sification is determined by $\hat{y} = \text{argmax}(z_0, z_1)$. ChimeraLM outputs a filtered BAM
793 file containing only reads classified as biological, which can be directly used for
794 downstream analyses including **SV** calling.

795
796 **Performance evaluation**

797 *Test set evaluation*

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

803
804 *SV calling*

805 **SVs** were called using multiple tools to ensure comprehensive detection. For long-
806 read data (ONT PromethION P2 and MinION Mk1c), we used Sniffles v2.5 [14, 15],
807 DeBreak v1.2 [16], SVIM v2.0.0 [17], and cuteSV v2.1.1 [18]. For short-read data of the
808 PC3 cell line, we used both the CCLE Illumina whole-genome sequencing dataset and
809 the PRJNA361315 Illumina WGS dataset, processed with Manta v1.6.0 [40], DELLY
810 v1.5.0 [41], and SvABA v1.1.0 [42]. All tools were executed with default recommended
811 parameters.

812
813 *Gold standard SV dataset construction*

814 A high-confidence gold standard **SV** dataset was generated from bulk PC3 sequencing
815 data to evaluate the impact of ChimeraLM on **SV** detection accuracy (Fig. 3a). All
816 **SV** comparison and breakpoint correction were performed using OctopusV v0.2.3 [43].
817 We used four datasets: bulk MinION Mk1c, bulk PromethION P2, the CCLE Illumina
818 WGS dataset, and the PRJNA361315 Illumina WGS dataset. Within each dataset, **SV**
819 events supported by at least two independent callers were retained. Variants supported
820 by two or more datasets were designated as gold standard **SVs** for benchmarking.

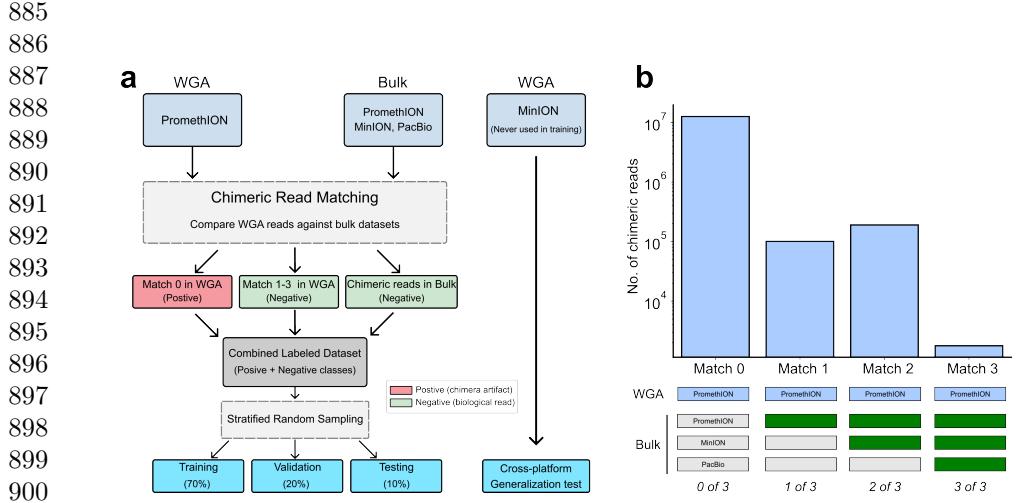
821
822 *SV benchmarking analysis*

823 To assess the impact of ChimeraLM on **SV** calling accuracy, we compared **SV** calls from
824 unfiltered **WGA** data and ChimeraLM-filtered **WGA** data against two references: (1)
825 the stringent multi-platform gold standard dataset, and (2) platform-matched long-
826 read bulk sequencing data. Benchmarking was performed using Truvari v4.2.2 [44]
827

with default parameters. SVs were considered supported if they matched reference variants within the defined breakpoint tolerance. Validation rates were calculated as the proportion of called SVs supported by the reference. This dual benchmarking strategy quantifies both improvements in detecting high-confidence multi-platform SVs and the retention of platform-specific true variants.	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
Benchmarking against existing methods	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
ChimeraLM was compared to two existing computational methods for detecting amplification-induced chimeric artifacts: SACRA [22] (GitHub commit 9a2607e) and 3rd-ChimeraMiner [13] (GitHub commit 04b5233). Both tools were applied to WGA data from PromethION P2 and MinION Mk1c platforms using default parameters as recommended in their documentation. Performance was evaluated by measuring the percentage reduction in chimeric reads relative to unprocessed WGA data. Chimeric reads were identified using WGA tag-based alignment criteria (reads with SA tags indicating split alignments), and reduction rates were calculated as the proportion of chimeric reads removed by each method.	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
Attention weight analysis	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
To investigate ChimeraLM’s interpretability, we analyzed attention weights from the pooling mechanism for representative chimeric reads. Attention weights indicate the relative importance assigned to each sequence position during classification. For selected reads, we extracted per-position attention weights and visualized them alongside read alignments to identify whether the model focuses on mechanistically relevant regions.	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
Chimeric junction positions were identified from alignment data (defined by breakpoints in SA tags). A window of ± 50 bp surrounding each junction was designated as the junction region. Attention weights within junction region were compared to non-junction regions using the Wilcoxon rank-sum test [45], with statistical significance assessed at $p < 0.001$.	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
Data visualization	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
Figures were generated using Python with Matplotlib [46] and Seaborn [47].	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
Computing resources	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
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).	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
Supplementary information.	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874
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.	860 861 862 863 864 865 866 867 868 869 870 871 872 873 874

875
876 **Extended Data Table 1** Sequencing and alignment statistics of PC3

Sample	Platform	Reads ($\times 10^6$)	Total bases (Gb)	Total aligned bases (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



902 **Extended Data Fig. 1 Training dataset construction and ground-truth labeling strategy**
903 **for PC3 cell line.** (a) Schematic workflow for generating labeled training data. **WGA** PromethION
904 data containing both biological and artificial chimeric reads is compared against three independent
905 bulk sequencing datasets from the same cell line (PromethION, MinION, and PacBio platforms).
906 Chimeric reads are classified through systematic matching: reads with no matches across all bulk
907 datasets (Match 0) are labeled as artificial chimeras (positive class, red); reads matching one or
908 more bulk datasets (Match 1–3) are labeled as biological reads (negative class, green), along with
909 chimeric reads sampled directly from bulk data. The combined labeled dataset undergoes stratified
910 random sampling to generate training (70%), validation (20%), and testing (10%) sets for model
911 development. The **WGA** MinION dataset is reserved as an independent cross-platform generalization
912 test set. (b) Distribution of chimeric read matches between **WGA** and bulk sequencing datasets. Bar
913 chart showing the number of chimeric reads (y-axis, log scale) grouped by how many bulk datasets (x-
914 axis) contained matching chimeric structures when comparing **WGA** PromethION reads against bulk
915 sequencing data. “Match 0” indicates reads with no matches in any bulk dataset (classified as artificial
916 chimeras, $\sim 10^7$ reads), whereas “Match 1–3” indicate reads with matches in one, two, or all three
917 bulk datasets (classified as biological reads, $\sim 10^5$ reads each). Color-coded boxes below bars indicate
918 which bulk platforms validated each read category: PromethION (light blue), MinION (white), and
919 PacBio (white); green boxes indicate platform-specific validation. The substantial imbalance between
920 Match 0 ($\sim 10^7$) and Match 1–3 categories ($\sim 10^5$ each) reflects the high prevalence of **WGA**-induced
artifacts, necessitating balanced subsampling for supervised learning.

921 **Declarations**

922 **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.	921 922 923 924 925 926 927 928 929 930 931 932
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=viej6cv6mgbli3n7a9a5k1bsb3	933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948
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/).	949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966
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.	943 944 945 946 947 948 949 950
Acronyms	947 948
CPU Central Processing Unit 12	949 950
DEL deletion 7, 9	951
DUP duplication 7, 9	952 953
GLM Genomic Language Model 2, 4, 12	954
GPU Graphics Processing Unit 12, 17, 19, 21	955 956
HPC High Performance Computing 19	957 958
INS insertion 7, 9	959
INV inversion 1, 2, 7, 9	960 961
LIANTI Linear Amplification via Transposon Insertion 12	962
MALBAC Multiple Annealing and Looping-based Amplification Cycles 12	963
MDA Multiple Displacement Amplification 2	964
MLP multilayer perceptron 3, 5, 16	965 966

967 **ONT** Oxford Nanopore Technologies [2, 4, 7, 8, 13, 14](#)
968
969 **PTA** Primary Template-directed Amplification [12](#)
970
971 **SA** Supplementary Alignment [14, 18, 19](#)
972 **SV** Structural Variation [1–5, 7–13, 15, 18, 19](#)
973
974 **TRA** translocation [2, 7, 9](#)
975
976 **WGA** Whole Genome Amplification [1–16, 18–20](#)

977 **References**
978

- 979 [1] Kalf-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using
980 different amplification methods and reference genomes. *Communications Biology*
981 1288 (2024).
982
983 [2] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**,
984 90–94 (2011).
985
986 [3] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons.
987 *Nature Communications* 4220 (2024).
988
989 [4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state
990 of the science. *Nature Reviews Genetics* 175–188 (2016).
991
992 [5] Chen, C. *et al.* Single-cell whole-genome analyses by linear amplification via
993 transposon insertion (LIANTI). *Science (new York, N.Y.)* **356**, 189–194 (2017).
994
995 [6] Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives.
996 *PLoS Genetics* **10**, e1004126 (2014).
997
998 [7] de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome
999 amplification methods. *PLoS ONE* e105585 (2014).
1000
1001 [8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification
1002 commercial kits using targeted sequencing. *Scientific Reports* 17171 (2021).
1003
1004 [9] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement
1005 amplification. *Computational and Structural Biotechnology Journal* 1688–1696
1006 (2023).
1007
1008 [10] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the
1009 multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).
1010
1011 [11] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displace-
1012 ment 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). 1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
- [13] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275 (2023). 1026
1027
1028
- [14] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 461–468 (2018). 1029
1030
1031
- [15] Smolka, M. *et al.* Detection of mosaic and population-level structural variants with sniffles2. *Nature Biotechnology* 1571–1580 (2024). 1032
1033
1034
- [16] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nature Communications* 283 (2023). 1035
1036
1037
1038
1039
1040
- [17] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* 2907–2915 (2019). 1041
1042
1043
- [18] Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* 189 (2020). 1044
1045
1046
1047
1048
- [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). 1049
1050
1051
- [20] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011). 1052
1053
- [21] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019). 1054
1055
1056
1057
1058
- [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).
- [23] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution*, Vol. 36, 43177–43201 (Curran Associates, Inc., 2023).
- [24] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 287–297 (2025).
- [25] Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for multi-species genomes, 1–24 (OpenReview.net, 2024).

- 1059 [26] Consens, M. E. *et al.* To transformers and beyond: Large language models for
1060 the genome (2023). [arXiv:2311.07621](https://arxiv.org/abs/2311.07621).
- 1061
- 1062 [27] Li, Y. *et al.* A genomic language model for chimera artifact detection in nanopore
1063 direct rna sequencing. *bioRxiv* (2024). URL <https://www.biorxiv.org/content/early/2024/10/25/2024.10.23.619929>.
- 1065
- 1066 [28] Routhier, E. & Mozziconacci, J. Genomics enters the deep learning era. *PeerJ*
1067 **10**, e13613 (2022).
- 1068
- 1069 [29] Poli, M. *et al.* *Hyena hierarchy: Towards larger convolutional language models*,
1070 Vol. 202, 28043–28078 (PMLR, 2023).
- 1071
- 1072 [30] Mahmoud, M. *et al.* Structural variant calling: The long and the short of it.
1073 *Genome Biology* **20**, 246 (2019).
- 1074
- 1075 [31] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023).
- 1076
- 1077 [32] Martin, M. Cutadapt removes adapter sequences from high-throughput sequenc-
1078 ing reads. *Embbnet.journal* **17**, 10–12 (2011).
- 1079
- 1080 [33] Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*
1081 3094–3100 (2018).
- 1082
- 1083 [34] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* giab008
1084 (2021).
- 1085
- 1086 [35] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of
1087 long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
- 1088
- 1089 [36] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning*
1090 library, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019).
- 1091
- 1092 [37] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub
1093 repository (2019). URL <https://github.com/Lightning-AI/lightning>.
- 1094
- 1095 [38] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).
- 1096
- 1097 [39] Yadan, O. Hydra - a framework for elegantly configuring complex applications.
1098 GitHub repository (2019). URL <https://github.com/facebookresearch/hydra>.
- 1099
- 1100 [40] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for
1101 germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).
- 1102
- 1103 [41] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end
1104 and split-read analysis. *Bioinformatics* i333–i339 (2012).

- [42] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Research* 581–591 (2018). 1105
1106
1107
- [43] 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). 1108
1109
1110
- [44] 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). 1111
1112
1113
1114
- [45] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 261–272 (2020). 1115
1116
1117
- [46] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 90–95 (2007). 1118
1119
1120
- [47] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* 3021 (2021). 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