

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046

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 genomic analysis relies on Whole Genome Amplification (WGA) to generate sufficient DNA for sequencing, but this process introduces chimeric artifacts that manifest as false-positive Structural Variations (SVs) and compromise downstream interpretation. Here we present ChimeraLM, a genomic language model that identifies and removes WGA-induced chimeric reads from long-read sequencing data. ChimeraLM uses a model architecture based on Hyena operators to analyze DNA sequences at single-nucleotide resolution, learning generalizable sequence features that distinguish genuine biological sequences from amplification-induced artifacts. When applied to nanopore sequencing data from WGA-amplified cells, ChimeraLM reduces chimeric read content by approximately 90% while retaining 87-92% of true SVs. This filtering improves SV validation rates 10-16 fold and normalizes SV type distributions toward bulk sequencing profiles, eliminating the characteristic false-positive inversion (INV) bias in unprocessed WGA data. Attention weight analysis reveals that ChimeraLM can focus on chimeric junction regions, learning biologically interpretable sequence features. ChimeraLM addresses a fundamental bottleneck in single-cell genomics,

047 enabling more confident detection of chromosomal instability and **SV** in appli-
048 cations across cancer biology, developmental biology, and neuroscience. The
049 software is available at <https://github.com/ylab-hi/ChimeraLM>.

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

052

053

054

055 Main

056

057 Single-cell genomics has revolutionized our understanding of cellular heterogeneity
058 by enabling characterization of individual cells rather than bulk populations [1–4].
059 This approach has proven instrumental in uncovering rare cell types [4], tracking
060 developmental trajectories [3], and elucidating tumor evolution through clonal archi-
061 tecture analysis. However, the limited DNA content in a single cell—typically only
062 6–7 picograms containing approximately two copies of the 3-billion-base-pair human
063 genome—poses significant technical challenges for comprehensive genomic analysis [5–
064 7]. To overcome this limitation, **WGA** has become essential for single-cell genomic
065 studies [4, 7–10]. Various **WGA** techniques have been developed, each with distinct
066 amplification mechanisms and characteristic error profiles. **Multiple Displacement**
067 **Amplification (MDA)**, introduced by Dean et al. [10], utilizes the highly processive
068 phi29 DNA polymerase to achieve isothermal amplification with products exceeding
069 10 kb, though it suffers from pronounced amplification bias and chimera forma-
070 tion [11, 12]. **Degenerate Oligonucleotide-Primed PCR (DOP-PCR)**, pioneered by
071 Telenius et al. [13], employs thermocycling with degenerate primers to achieve more
072 uniform coverage but generates shorter amplicons. **Multiple Annealing and Looping-**
073 **based Amplification Cycles (MALBAC)** combines quasi-linear preamplification with
074 exponential amplification to reduce bias [8], while **Linear Amplification via Transposon**
075 **Insertion (LIANTI)** uses transposon insertion to create defined amplification origins,
076 significantly improving uniformity and reducing artifacts [7]. More recently, **Primary**
077 **Template-directed Amplification (PTA)** [14] and **droplet-based MDA (dMDA)** [15, 16]
078 have emerged as promising alternatives that modify reaction conditions to suppress
079 chimera formation, though these methods require specialized equipment and protocols
080 that have limited their widespread adoption. These amplification methods can increase
081 DNA content by several orders of magnitude (typically 1,000- to 10,000-fold), gener-
082 ating sufficient material for high-coverage sequencing necessary for reliable variant
083 calling, copy number analysis, and **SV** detection [4, 17–21].

084

085 Accurate single-cell genomics is particularly critical for multiple applications where
086 false-positive **SVs** can lead to incorrect biological conclusions. In cancer research, dis-
087 tinguishing genuine clonal evolution patterns from amplification artifacts is essential
088 for understanding tumor heterogeneity and therapeutic resistance [3]. In develop-
089 mental biology, accurate detection of somatic mosaicism enables the reconstruc-
090 tion of lineage relationships and identification of pathogenic mutations in rare cell pop-
091 ulations. For CRISPR-based genome editing, single-cell analysis with reliable **SV**
092 detection is crucial for comprehensive assessment of off-target effects and ensuring

genomic stability [14]. However, false-positive SVs introduced during amplification can confound these analyses, leading to misinterpretation of genomic rearrangements and their biological significance [4, 22].	093
Despite its critical role, WGA introduces systematic artifacts that significantly impact downstream analyses [7, 11, 12, 22, 23]. Among the most problematic are chimeric sequences—artificial DNA constructs formed through template switching during amplification. During MDA , the highly processive phi29 polymerase can dissociate from one genomic template and reinitiate	096
o detect SVs from long-read data, including Sniffles2 [24, 25], DeBreak [26], SVIM [27], and cuteSV [28]. These methods typically employ read alignment analysis, split-read detection, and local assembly strategies to identify SV signatures [29]. However, distinguishing genuine biological SVs from WGA -induced chimeric artifacts remains challenging [23, 30–32].	097
Current computational approaches for identifying WGA -induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [23, 30]. However, these heuristic methods often fail to distinguish genuine SVs from amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns, occur in repetitive genomic regions, or involve multiple genomic loci [31, 32]. This lack of robust, automated artifact detection has limited the reliability of SV analysis in single-cell studies and hindered the full realization of single-cell genomics’ potential for studying somatic mosaicism, tumor evolution, and rare cell populations.	098
The emergence of deep learning, particularly language models based on transformer architectures, has demonstrated remarkable success in genomics applications [33–36]. Recent genomic language models have shown the ability to learn complex sequence patterns and contextual relationships in DNA sequences, enabling improved performance in tasks such as regulatory element prediction, variant effect prediction, and functional annotation [36–38]. These models treat DNA sequences analogously to natural language, learning representations that capture both local motifs and long-range dependencies [33]. By training on large-scale genomic datasets, such models can internalize patterns of genuine biological sequences, including characteristic features of repetitive elements, chromatin structure, and sequence composition biases.	099
Here, we developed ChimeraLM, a genomic language model specifically designed to detect chimeric artifacts introduced by WGA . By leveraging deep learning to capture sequence patterns, structural features, and contextual information in genomic reads [33–36, 38], ChimeraLM effectively distinguishes genuine biological sequences from WGA -induced chimeric artifacts. We demonstrate that ChimeraLM achieves superior performance compared to existing methods and substantially improves the reliability of SV detection in single-cell genomic studies, thereby enabling accurate SV analysis at single-cell resolution.	100
Results	101
Overview of ChimeraLM workflow and model architecture	102
Single-cell genomics relies on WGA to obtain sufficient DNA for sequencing (Fig. 1a). The standard workflow includes single-cell isolation, DNA extraction, WGA , long-read	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

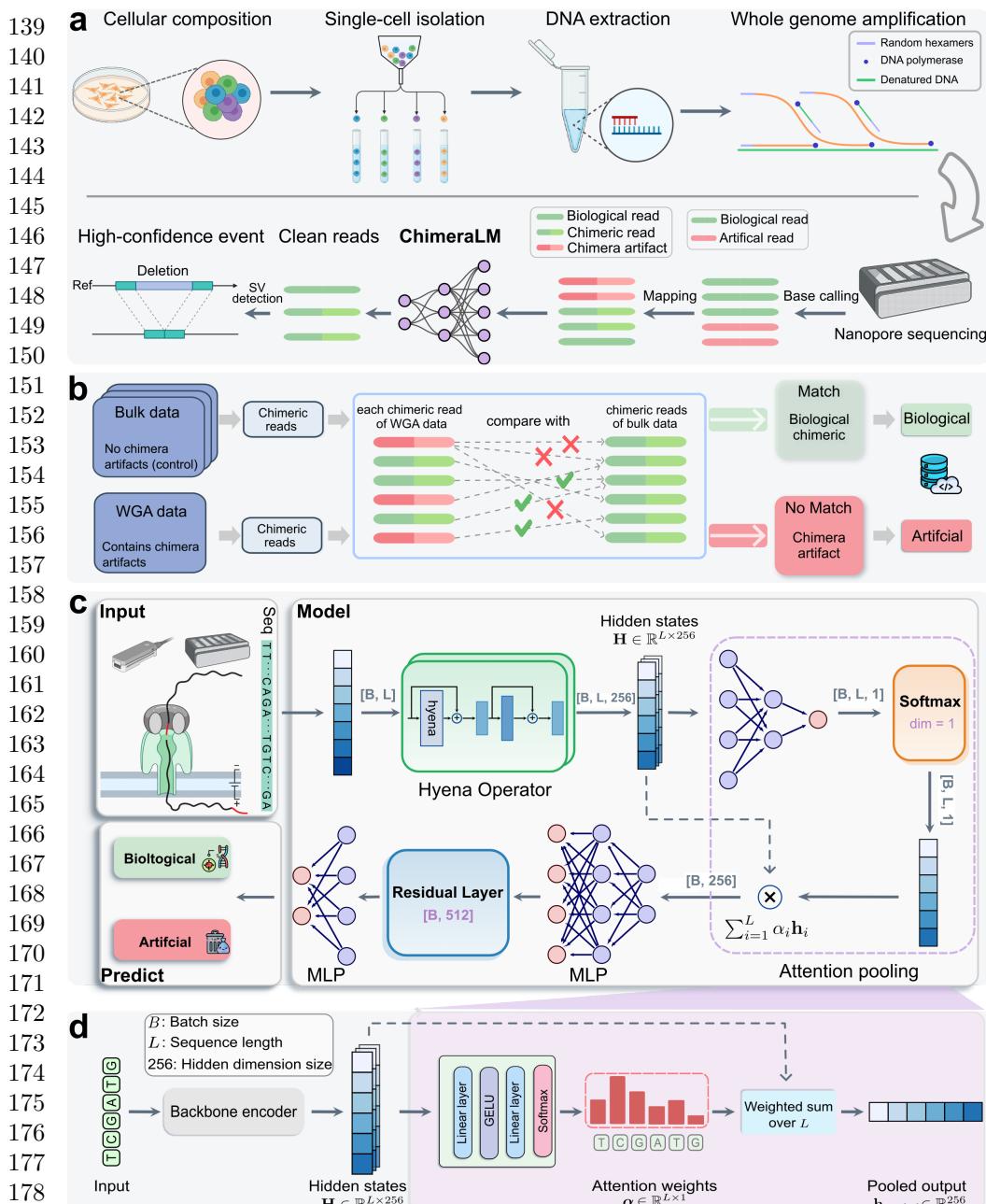


Fig. 1 ChimeraLM workflow and architecture for detecting WGA artifacts in single-cell sequencing. (a) Single-cell genomic workflow and ChimeraLM integration. Single cells are isolated, followed by DNA extraction and WGA for genome amplification. WGA generates chimeric artifacts (red) through template 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 backbone encoder (HyenaDNA [35]). 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 backbone 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.

sequencing (e.g., Oxford Nanopore Technologies (ONT)), base calling, and alignment to the reference genome. During amplification, template-switching events introduce artificial chimeric reads, resulting in alignment files that contain a mixture of authentic and artifactual sequences. In downstream analysis, these artifacts can mimic SV and confound variant detection. To address this challenge, we developed ChimeraLM, a Genomic Language Model (GLM) designed to integrate directly into this analysis pipeline and distinguish biological reads from amplification-induced artifacts.

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

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

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

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

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

231 the complete sequence information necessary for detecting chimeric junctions—the
232 breakpoints where disparate genomic regions are artificially fused and which often
233 exhibit abrupt compositional changes.

234 The architecture’s core employs Hyena operators [39], selected specifically to
235 overcome the challenge of processing long DNA sequences. Traditional attention
236 mechanisms scale quadratically with sequence length, making them computationally
237 prohibitive for long-read data. Hyena operators, by contrast, achieve subquadratic
238 scaling, enabling ChimeraLM to analyze full-length reads without fragmentation and
239 thus preserve the structural context around chimeric junctions. To leverage existing
240 genomic knowledge, we initialized the model with weights from HyenaDNA [35], a
241 genomic foundation model pre-trained on diverse DNA sequences.

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

252

253 **ChimeraLM achieves high accuracy and reduces artifacts to 254 near-bulk levels across platforms**

255

256 We first evaluated ChimeraLM’s classification accuracy on the held-out test set
257 (derived from the PromethION training data), which comprised reads with known bio-
258 logical or artificial status (Fig. 2a). The model achieved an F1 score of 0.81, reflecting
259 balanced sensitivity and specificity in artifact detection. A recall of 0.95 indicates that
260 95% of true chimeric reads were correctly identified—critical for minimizing down-
261 stream false-positive structural variant calls—while a precision of 0.70 shows that the
262 majority of reads flagged as chimeric were true artifacts. These results establish the
263 model’s reliability for identifying amplification-induced artifacts in long-read data.

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

273 We then benchmarked ChimeraLM against existing computational tools for
274 detecting amplification-induced chimeras, SACRA [30] and 3rd-ChimeraMiner [23]
275 (Fig. 2c,d). When applied to the same PromethION and MinION WGA data,
276

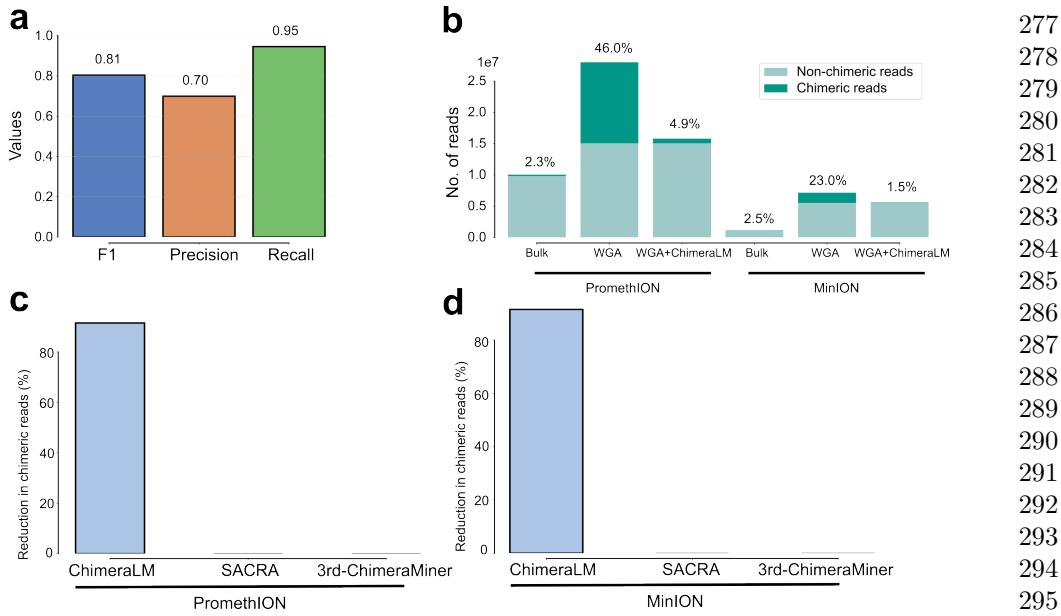


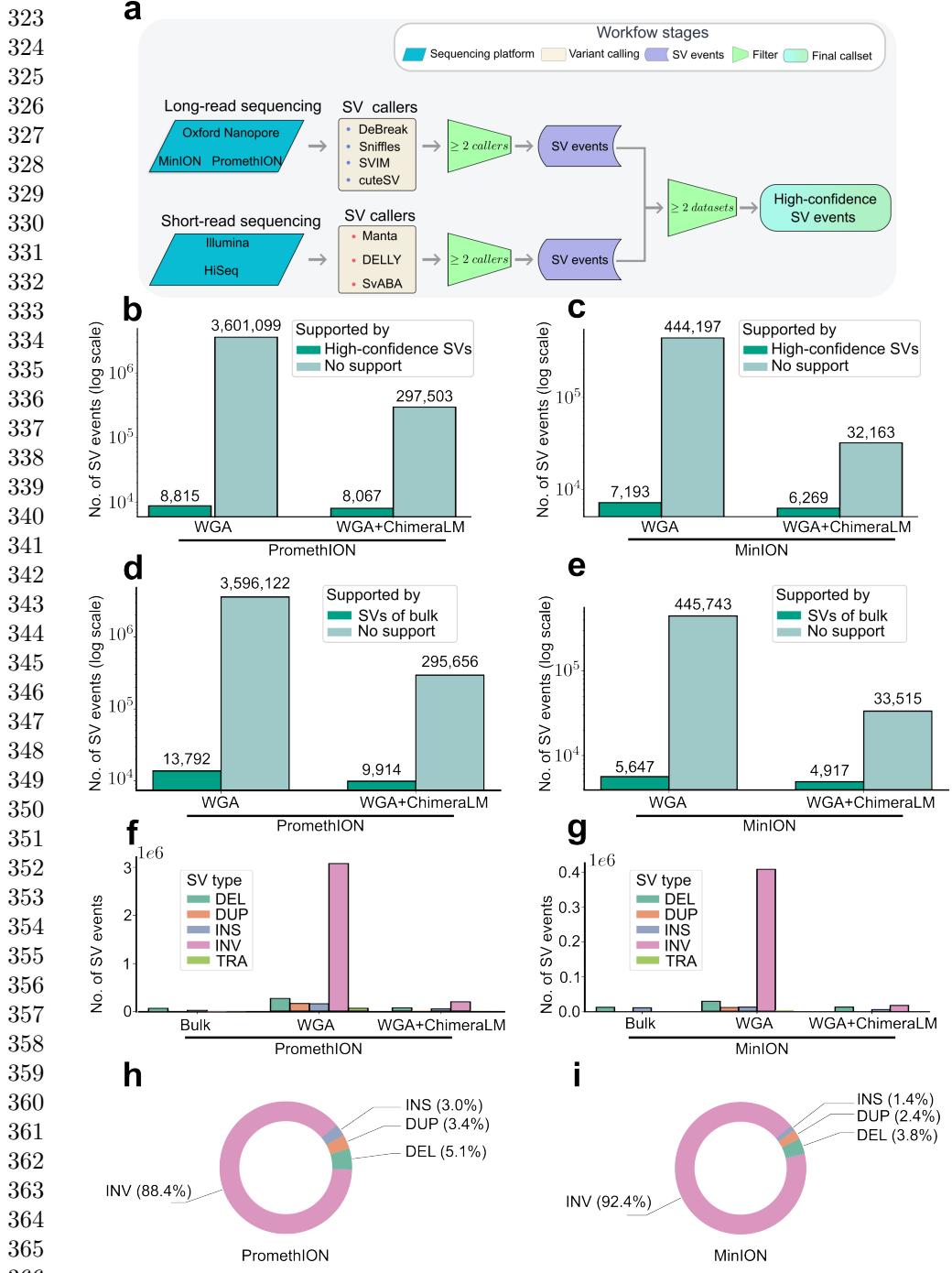
Fig. 2 ChimeraLM accurately identifies and removes WGA-induced chimeric artifacts.

(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 a robust and generalizable performance. The strong filtering on the MinION dataset (Fig. 2b) is particularly noteworthy, as this dataset served as a completely independent test set; the model was trained exclusively on PromethION data. This cross-platform generalization, combined with the high recall on the internal test set (Fig. 2a) and the clear superiority over existing tools (Fig. 2c,d), indicates that ChimeraLM learns fundamental, generalizable sequence features of WGA-induced artifacts rather than platform-specific signatures.

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



367 **Fig. 3 ChimeraLM improves structural variant detection accuracy.** (a) Construction of
368 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 translocation
(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.

ChimeraLM substantially reduces false-positive structural variant calls	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
ChimeraLM restores unbiased SV-type distributions and characterizes artifact composition	414
Amplification artifacts can distort the apparent spectrum of SVs , often inflating specific SV types. To evaluate whether ChimeraLM effectively corrects such distortions,	414

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

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

432 This characterization has important implications for single-cell genomics. Although
433 **INVs** are the predominant artifact type, the coexistence of **DELs**, **DUPs**, and **INSS**
434 among chimeric events indicates that comprehensive filtering—rather than inversion-
435 specific correction—is essential for accurate **SV** detection. Without ChimeraLM
436 filtering, single-cell **SV** analyses would be confounded not only by false-positive **INVs**
437 but also by other artifact-associated variants [31, 32]. By restoring biologically repre-
438 sentative **SV** type distributions, ChimeraLM enables robust and interpretable charac-
439 terization of structural variation in single cells without distortion from **WGA**-induced
440 artifacts.

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

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

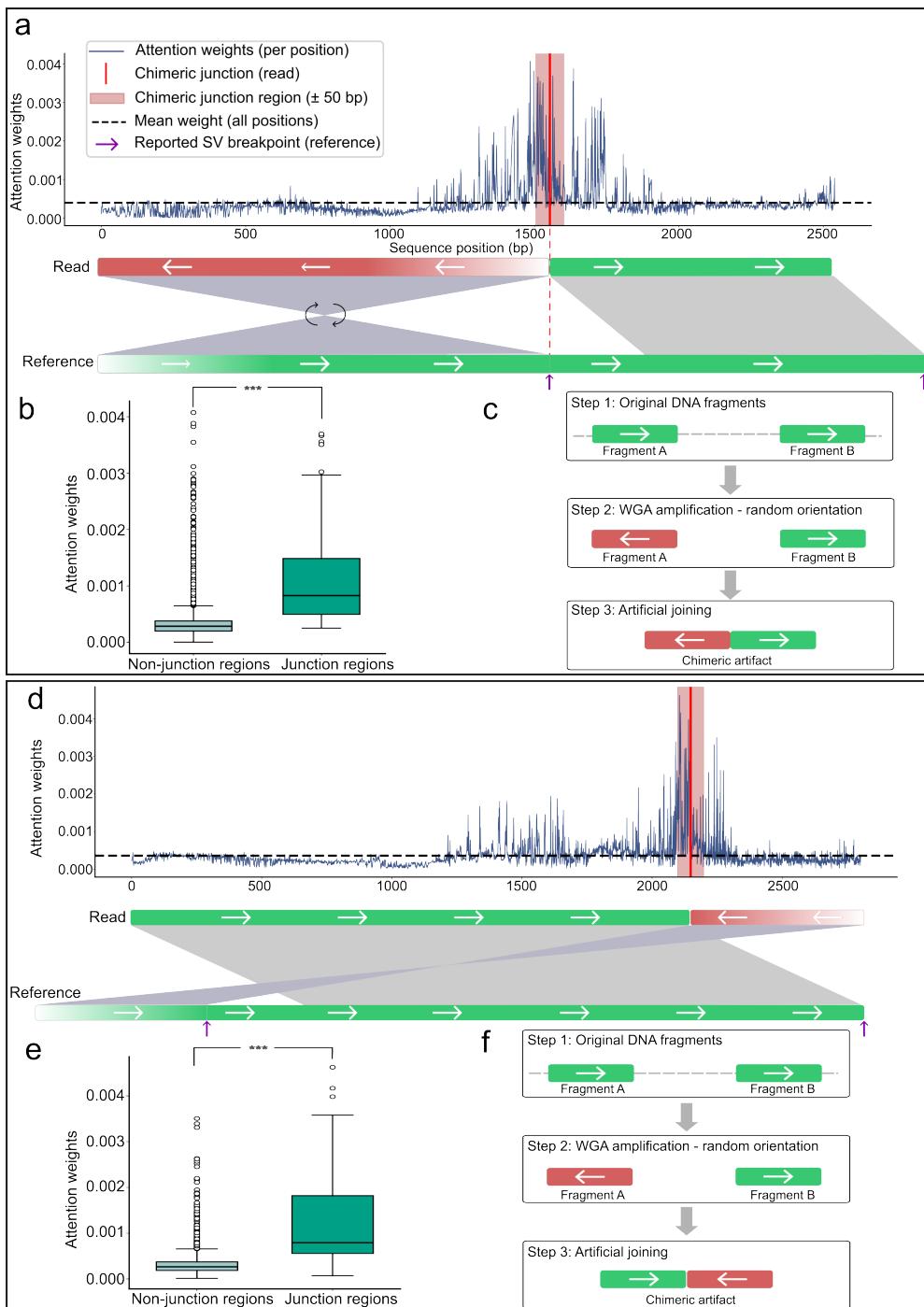


Fig. 4 ChimeraLM attention weights can localize to chimeric junction regions. (a,d) Attention weight profiles for two representative chimeric reads. Upper panels show attention weights per sequence position (blue line) and mean attention (dashed line). Red vertical lines mark chimeric junction positions, 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.

507 **ChimeraLM provides interpretable classification through
508 attention visualization**

509 We next investigated whether ChimeraLM’s attention mechanism highlights biologi-
510 cally meaningful regions within sequencing reads (Fig. 4).

512 For representative chimeric reads, attention weight profiles showed low baseline
513 values across most positions but pronounced peaks at junction regions where tem-
514 plate switching artificially joins DNA fragments from distinct genomic loci (Fig. 4a,d).
515 These peaks coincided precisely with alignment breakpoints characterized by orienta-
516 tion changes between adjacent read segments—the defining signature of **WGA**-induced
517 chimeric artifacts.

518 Quantitative analysis confirmed that attention weights within junction regions
519 (± 50 bp) were significantly higher than those in non-junction regions (Wilcoxon rank-
520 sum test, $p = 5.3 \times 10^{-14}$ and $p = 6.8 \times 10^{-15}$) (Fig. 4b,e). Such localization indicates
521 that ChimeraLM learns mechanistically relevant features associated with artificial
522 junction formation rather than relying on spurious correlations.

523 Schematic reconstruction of the amplification process further supports this inter-
524 pretation (Fig. 4c,f). During **WGA**, DNA fragments from distant genomic loci may
525 undergo random strand orientation changes before being joined by template switching.
526 This process produces artificial junctions with discordant orientations—forward-to-
527 reverse or reverse-to-forward—that generate inversion-like alignment signatures and
528 are effectively recognized by the model’s attention peaks.

529 Together, these analyses demonstrate that ChimeraLM’s attention mechanism
530 can localize chimeric junctions at single-base resolution and capture the underlying
531 orientation discontinuities that define **WGA**-induced artifacts.

532 **533 Discussion**

534 **WGA** has enabled genomic analysis from single cells but introduces chimeric arti-
535 facts that compromise **SV** detection. ChimeraLM addresses this challenge through
536 sequence-level classification of biological versus artificial reads, substantially improving
537 **SV** calling accuracy before downstream analysis. This upstream filtering strategy—
538 removing problematic sequences at the read level rather than correcting errors post
539 hoc—provides a practical solution for single-cell genomics laboratories.

540 Our results demonstrate several key advantages of ChimeraLM for long-read single-
541 cell sequencing. The method achieves approximately 90% reduction in chimeric reads
542 across nanopore platforms while retaining 87–92% of true **SVs**. It reduces false-positive
543 **SV** calls by 8–16 fold, enabling researchers to focus on biologically relevant variants
544 without manually filtering thousands of artifacts. Moreover, ChimeraLM performs
545 consistently across PromethION and MinION without platform-specific retraining,
546 indicating that it captures generalizable sequence features of **WGA**-induced chimeras.
547 These results underscore the model’s robustness across diverse datasets and sequencing
548 conditions.

549 ChimeraLM’s effectiveness reflects the ability of deep learning models to capture
550 complex sequence patterns that are difficult to encode in rule-based filters. Traditional
551 quality control methods rely on predefined metrics such as mapping quality or read

depth [23, 30], which may not effectively distinguish chimeric artifacts from biological reads. By learning directly from data, ChimeraLM discovers subtle compositional and structural features that differentiate authentic genomic sequences from amplification artifacts. Furthermore, the model offers interpretability through attention visualization, allowing researchers to examine which sequence regions drive classification. Attention weights can concentrate sharply at junctions where template switching joins DNA fragments from distinct loci, matching the known mechanism of chimera formation. Some reads show more diffuse attention distributions, suggesting that ChimeraLM integrates multiple complementary cues—such as junction orientation, compositional biases, and local sequence context—to classify diverse artifact types. This interpretability builds confidence in the model’s predictions and provides a lens for probing the molecular processes underlying amplification-induced artifacts.

The improved reliability of **SV** detection has direct implications for single-cell genomics. Studies of chromosomal instability, clonal evolution, and **SV** burden in individual cells have long been constrained by high false-positive rates in **WGA** data [31, 32]. ChimeraLM enables more confident identification of genuine **SVs**, supporting research in cancer genomics, developmental biology, and aging where single-cell resolution is essential for understanding cellular heterogeneity. Although the current model processes reads independently, integrating additional contextual features—such as coverage, mate-pair, or phasing information—could further enhance accuracy. **Graphics Processing Unit (GPU)** resources are recommended for large-scale datasets, while **Central Processing Unit (CPU)** inference remains feasible for smaller studies; runtime optimization and model compression may improve accessibility for broader use.

Future work should prioritize validation across diverse biological contexts in long-read single-cell sequencing. Testing on multiple cell types (primary, stem, or immune cells) and **WGA** protocols will establish generalizability. The interpretability of attention-based models could also be leveraged to investigate mechanisms of chimera formation: large-scale analysis of attention patterns may reveal recurrent sequence motifs or genomic contexts associated with template switching, guiding the development of improved amplification protocols. More broadly, ChimeraLM illustrates the potential of **GLMs** for data quality control applications [35]. Architectural innovations such as the **Hyena** operator for efficient long-range modeling [39] may have utility beyond chimera detection, addressing challenges such as contamination, adapter artifacts, and systematic sequencing errors.

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.

599 **Methods**

600

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

602 *Cell culture and single-clone establishment*

603

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

610

611 *DNA extraction and whole-genome amplification*

612 From the monoclonal population, two types of DNA samples were prepared: a
613 bulk (non-amplified) control and ten single-cell MDA-amplified genomes. Bulk high-
614 molecular-weight DNA was extracted using the Monarch® HMW DNA Extraction
615 Kit for Cells & Blood (New England Biolabs). Individual cells were isolated using
616 1CellDish-60 mm (iBiochips) and amplified using the REPLI-g Advanced DNA Sin-
617 gle Cell Kit (Qiagen) following the manufacturer's protocol. DNA concentration and
618 fragment integrity were assessed with a Qubit 4 fluorometer and Agilent TapeStation
619 (DNA 1000/5000 ScreenTape). Only samples meeting quality standards were used for
620 library construction.

621

622 *Nanopore library preparation and sequencing*

623 Sequencing libraries were prepared using the ONT Ligation Sequencing Kit V14 (SQK-
624 LSK114) and sequenced on MinION Mk1C or PromethION P2 Solo devices with
625 R10.4.1 flow cells according to the manufacturer's genomic DNA workflow. Because
626 all single-cell samples originated from the same monoclonal lineage, observed differ-
627 ences between amplified and bulk data primarily reflect MDA-induced artifacts rather
628 than biological variation, providing a controlled experimental setting for downstream
629 analyses.

630

631 *Basecalling and read processing*

632 Raw signal files (POD5) were basecalled using Dorado v0.5.0 with the high-accuracy
633 model dna_r10.4.1_e8.2_400bps_hac@v4.3.0 [40]. Reads with mean quality < 10
634 or length < 500 bp were removed. Residual adapters and concatemers were trimmed
635 using Cutadapt v4.0 [41] in two-pass error-tolerant mode. Cleaned reads were aligned
636 to the GRCh38.p13 reference genome using minimap2 v2.26 (`map-ont` preset) [42].
637 Resulting BAM files were sorted and indexed with SAMtools v1.16 [43]. Read length
638 and mapping statistics were calculated using NanoPlot v1.46.1 [44]. All samples were
639 processed under identical parameters to ensure consistency across datasets.

640

641 *Chimeric read identification*

642

643 Chimeric reads were identified based on the presence of supplementary alignments in
644 BAM files using the [Supplementary Alignment \(SA\)](#) tag. The SA tag indicates that

a read has additional alignments beyond the primary alignment, which is characteristic of chimeric sequences that map to multiple distant genomic locations. To ensure accurate identification, we applied stringent filtering criteria: reads were classified as chimeric only if they (1) were not unmapped, (2) contained the **SA** tag, (3) were not secondary alignments, and (4) were not supplementary alignments themselves. This filtering approach ensures that only primary alignments with supplementary mapping evidence are considered chimeric, avoiding double-counting of the same chimeric event and excluding low-quality or ambiguous alignments. Reads without the **SA** tag (single continuous alignments) were classified as non-chimeric. This approach leverages the standard BAM format specification to reliably identify reads with complex alignment patterns.

Training data construction

Data generation and sources

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. These bulk datasets represent authentic biological sequences free from amplification-induced artifacts. In contrast, **WGA** sequencing includes both genuine genomic reads and artificial chimeras introduced during the amplification process. An additional **WGA** dataset sequenced on the MinION Mk1c platform was reserved exclusively as an independent test set for cross-platform evaluation.

Ground truth annotation and class definition

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

691 were used respectively for model training, hyperparameter tuning, and performance
692 evaluation on data from the same sequencing platform.

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

699

700 Model architecture

701 *Backbone encoder*

703 ChimeraLM employs the pre-trained HyenaDNA model [35] as its backbone encoder.
704 This model was pre-trained on large-scale genomic data and provides robust sequence
705 representations. DNA sequences are tokenized at single-nucleotide resolution, with
706 each base (A, C, G, T, N) mapped to a unique integer token (7, 8, 9, 10, 11, respec-
707 tively). Special tokens include [CLS]=0, [PAD]=4, and others for sequence processing.
708 Input sequences are truncated at 32,768 bp or padded to enable batch processing.

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

$$711 \quad \mathbf{H} = \text{HyenaDNA}(\mathbf{x}) \in \mathbb{R}^{L \times 256}$$

713

714 where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L)$ represents position-wise hidden states with dimension 256.
715 The Hyena operators [39] efficiently capture both local sequence motifs and long-range
716 dependencies essential for distinguishing biological sequences from chimeric artifacts.

717

718 *Attention pooling*

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

$$722 \quad \mathbf{e} = \text{GELU}(\text{Linear}_{256 \rightarrow 256}(\mathbf{H})) \in \mathbb{R}^{L \times 256}$$

$$723 \quad \mathbf{s} = \text{Linear}_{256 \rightarrow 1}(\mathbf{e}) \in \mathbb{R}^{L \times 1}$$

$$724 \quad \boldsymbol{\alpha} = \text{softmax}(\mathbf{s}) \in \mathbb{R}^{L \times 1}$$

727

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

729

$$730 \quad \mathbf{h}_{\text{pooled}} = \sum_{i=1}^L \alpha_i \mathbf{h}_i \in \mathbb{R}^{256}$$

732

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

736

Classification head	737
The pooled representation is processed through a MLP with residual connections. The first layer expands dimensionality:	738
	739
	740
$\mathbf{f}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{Linear}_{256 \rightarrow 512}(\mathbf{h}_{\text{pooled}}))) \in \mathbb{R}^{512}$	741
Subsequent residual blocks with input $\mathbf{f}_{\text{in}} \in \mathbb{R}^{512}$ compute:	742
	743
$\mathbf{f}_{\text{out}} = \text{Dropout}_{0.1}(\text{Linear}_{512 \rightarrow 512}(\text{GELU}(\text{Linear}_{512 \rightarrow 512}(\mathbf{f}_{\text{in}})))) + \mathbf{f}_{\text{in}}$	744
where the skip connection enables stable gradient flow during training. The final layer produces binary classification logits:	745
	746
$\mathbf{z} = [z_0, z_1] = \text{Linear}_{512 \rightarrow 2}(\mathbf{f}_{\text{final}}) \in \mathbb{R}^2$	747
where z_0 and z_1 represent logits for biological and artificial chimeric classes, respectively. During inference, the predicted class is $\hat{y} = \text{argmax}_{i \in \{0,1\}} z_i$.	748
	749
Model summary	750
The complete ChimeraLM pipeline processes DNA sequences through: (1) single-nucleotide tokenization, (2) HyenaDNA backbone encoding to generate contextualized representations, (3) attention pooling to aggregate position-specific features, (4) MLP layers with residual connections to learn classification features, and (5) binary classification output. The entire model is trained end-to-end using labeled WGA and bulk sequencing data.	751
	752
	753
	754
	755
	756
	757
	758
	759
	760
	761
	762
	763
	764
	765
	766
	767
	768
	769
	770
	771
	772
	773
	774
	775
	776
	777
	778
	779
	780
	781
	782
Model training and optimization	763
Training configuration	764
ChimeraLM was trained using PyTorch [45] and PyTorch Lightning [46] frameworks. Input sequences were tokenized using the tokenizer with maximum sequence length of 32,768 bp. Sequences longer than this threshold were truncated; shorter sequences were padded to enable batch processing. Training employed mixed-precision computation (bf16) to accelerate training while maintaining numerical stability.	765
	766
	767
	768
	769
	770
	771
	772
	773
	774
	775
	776
	777
	778
	779
	780
	781
	782
Optimization procedure	772
We used the AdamW optimizer [47] with learning rate 1×10^{-4} and weight decay 0.01. 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.	773
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 $z = [z_0, z_1]$, the loss	774

783 is:

$$784 \quad \mathcal{L} = -\log \left(\frac{\exp(z_y)}{\exp(z_0) + \exp(z_1)} \right)$$

785

786 where z_0 and z_1 represent logits for biological and artificial chimeric classes.

787

788 ***Training implementation***

789 Training used batch size of 16 sequences with 30 parallel data loading workers. **GPU**
790 acceleration was employed for efficient processing, with training typically requiring 96-
791 120 hours depending on dataset size. Model checkpointing saved the best-performing
792 model based on validation metrics. Configuration management used Hydra [48] to
793 enable reproducible experimentation.

794

795 ***Model evaluation***

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

798

$$799 \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

800

$$801 \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

802

803

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

809

810 **Model inference and application**

811

812 ***Inference pipeline***

813 To apply ChimeraLM to new **WGA** sequencing data, the model takes a BAM file as
814 input. Chimeric reads are identified using **SA** tags and filtered to exclude unmapped,
815 secondary, or supplementary alignments. Each chimeric read sequence is tokenized
816 using the tokenizer (maximum length 32,768 bp, with truncation or padding as
817 needed). The trained model processes sequences in batches, generating two logits
818 $[z_0, z_1]$ for each read corresponding to biological and artificial chimeric classes. Clas-
819 sification is determined by $\hat{y} = \text{argmax}(z_0, z_1)$. ChimeraLM outputs a filtered BAM
820 file containing only reads classified as biological, which can be directly used for
821 downstream analyses including **SV** calling.

822

823 **Performance evaluation**

824

825 ***Test set evaluation***

826 Final model performance was evaluated on the held-out test set and the independent
827 MinION Mk1c dataset. Metrics (precision, recall, F1 score, accuracy) were computed
828 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. 829

SV calling

SVs were called using multiple tools to ensure comprehensive detection. For long-read data (ONT PromethION P2 and MinION Mk1c), we used Sniffles v2.5 [24, 25], DeBreak v1.2 [26], SVIM v2.0.0 [27], and cuteSV v2.1.1 [28]. For short-read data of the PC3 cell line, we used both the CCLE Illumina whole-genome sequencing dataset and the PRJNA361315 Illumina WGS dataset, processed with Manta v1.6.0 [49], DELLY v1.5.0 [50], and SvABA v1.1.0 [51]. All tools were executed with default recommended parameters.

Gold standard SV dataset construction

A high-confidence gold standard **SV** dataset was generated from bulk PC3 sequencing data to evaluate the impact of ChimeraLM on **SV** detection accuracy (Fig. 3a). All **SV** comparison and breakpoint correction were performed using OctopuSV v0.2.3 [52]. We used four datasets: bulk MinION Mk1c, bulk PromethION P2, the CCLE Illumina WGS dataset, and the PRJNA361315 Illumina WGS dataset. Within each dataset, **SV** events supported by at least two independent callers were retained. Variants supported by two or more datasets were designated as gold standard **SVs** for benchmarking.

SV benchmarking analysis

To assess the impact of ChimeraLM on **SV** calling accuracy, we compared **SV** calls from unfiltered **WGA** data and ChimeraLM-filtered **WGA** data against two references: (1) the stringent multi-platform gold standard dataset, and (2) platform-matched long-read bulk sequencing data. Benchmarking was performed using Truvari v4.2.2 [53] 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.

Benchmarking against existing methods

ChimeraLM was compared to two existing computational methods for detecting amplification-induced chimeric artifacts: SACRA [30] (GitHub commit 9a2607e) and 3rd-ChimeraMiner [23] (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.

875 **Attention weight analysis**

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

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

888
889 **Data visualization**

890 Figures were generated using Python with Matplotlib [55] and Seaborn [56].

891
892 **Computing resources**

893 Computations were performed on a [High Performance Computing \(HPC\)](#) server with
894 64-core Intel Xeon Gold 6338 CPU, 256 GB RAM, and two NVIDIA A100 [GPUs](#) (80
895 GB memory each).

896
897

898

899

900 **Extended Data Table 1** Sequencing and alignment statistics of PC3

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

906
907
908

909
910
911 **Supplementary information.**

912

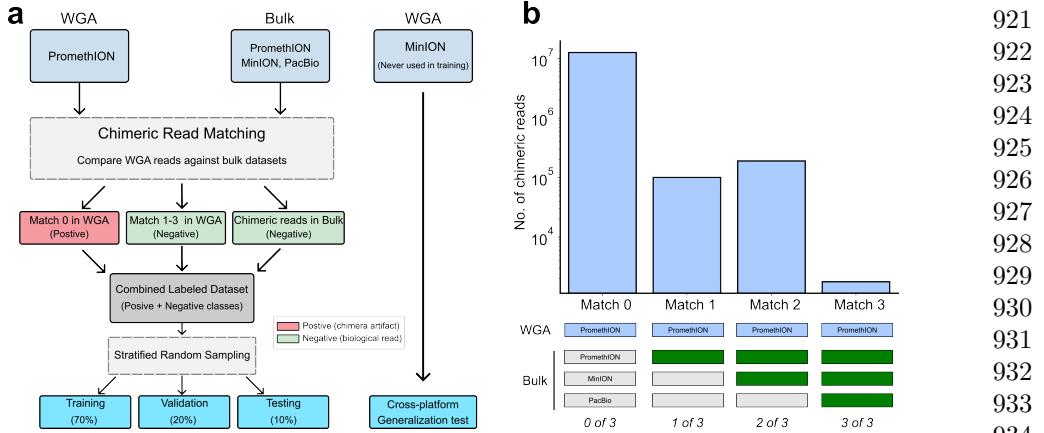
913 **Acknowledgements.** We thank Tingyou Wang for guidance on figure preparation.
914 This project was supported in part by NIH grants R35GM142441 and R01CA259388
915 awarded to RY.

916

917 **Declarations**

918

919 **Author Contributions.** YL, QG and RY designed the study. YL and QG per-
920 formed the analysis. QG performed the experiments. YL and QG designed and



Extended Data Fig. 1 Distribution of chimeric read matches between WGA and bulk sequencing datasets. Bar chart showing the number of chimeric reads (y-axis, log scale) grouped by how many bulk datasets (x-axis) contained matching chimeric structures when comparing WGA PromethION reads against bulk sequencing data. “Match 0” indicates reads with no matches in any bulk dataset (classified as artificial chimeras, $\sim 10^7$ reads), whereas “Match 1–3” indicate reads with matches in one, two, or all three bulk datasets (classified as biological reads, $\sim 10^5$ reads each). Symbols below the bars represent the number of bulk datasets with a match (0 of 3 – 3 of 3). This matching scheme forms the basis for ground-truth labeling in supervised training.

implemented the model. YL built the command-line tool and documentation. YL, QG and RY wrote the manuscript. RY supervised this work.

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>

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/>).

967 **Conflict of interest.** RY has served as an advisor/consultant for Tempus AI, Inc.
968 This relationship is unrelated to and did not influence the research presented in this
969 study.

970

971 **Acronyms**

972

973 **CPU** Central Processing Unit [13](#)

974

975 **DEL** deletion [8, 10](#)

976 **dMDA** droplet-based MDA [2](#)

977 **DOP-PCR** Degenerate Oligonucleotide-Primed PCR [2](#)

978 **DUP** duplication [8, 10](#)

979

980 **GLM** Genomic Language Model [5, 13](#)

981 **GPU** Graphics Processing Unit [13, 18, 20, 21](#)

982

983 **HPC** High Performance Computing [20](#)

984

985 **INS** insertion [8, 10](#)

986 **INV** inversion [1, 8, 10](#)

987

988 **LIANTI** Linear Amplification via Transposon Insertion [2](#)

989

990 **MALBAC** Multiple Annealing and Looping-based Amplification Cycles [2](#)

991 **MDA** Multiple Displacement Amplification [2, 3](#)

992 **MLP** multilayer perceptron [4, 6, 17](#)

993

994 **ONT** Oxford Nanopore Technologies [5, 8, 9, 14, 15](#)

995

996 **PTA** Primary Template-directed Amplification [2](#)

997

998 **SA** Supplementary Alignment [14, 15, 18–20](#)

999 **SV** Structural Variation [1–5, 8–13, 15, 18, 19](#)

1000

1001 **WGA** Whole Genome Amplification [1–13, 15–21](#)

1002

1003 **References**

1004

1005 [1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using
1006 different amplification methods and reference genomes. *Communications Biology*
1007 1288 (2024).

1008

1009 [2] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons.
1010 *Nature Communications* 4220 (2024).

1011

1012

- [3] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011). 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
- [4] Macaulay, I. C. & Voet, T. Single cell genomics: Advances and future perspectives. *PLOS Genetics* **10**, e1004126 (2014).
- [5] Leung, M. L. *et al.* Highly multiplexed targeted dna sequencing from single nuclei. *Nature Protocols* 214–235 (2016).
- [6] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* 175–188 (2016).
- [7] Chen, C. *et al.* Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science (new York, N.Y.)* **356**, 189–194 (2017).
- [8] Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 1622–1626 (2012).
- [9] Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annual Review of Genomics and Human Genetics* 79–102 (2015).
- [10] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* **99**, 5261–5266 (2002).
- [11] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnology* **7**, 19 (2007).
- [12] Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216 (2006).
- [13] Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
- [14] 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).
- [15] Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nature Communications* **14**, 5164 (2023).
- [16] Dippenaar, A. *et al.* Droplet based whole genome amplification for sequencing minute amounts of purified mycobacterium tuberculosis DNA. *Scientific Reports* **14**, 9931 (2024).

- 1059 [17] de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome
1060 amplification methods. *PLoS ONE* e105585 (2014).
- 1061
- 1062 [18] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification
1063 commercial kits using targeted sequencing. *Scientific Reports* 17171 (2021).
- 1064
- 1065 [19] Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion
1066 whole-genome amplification. *Proceedings of the National Academy of Sciences*
1067 11923–11928 (2015).
- 1068
- 1069 [20] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displace-
1070 ment amplification with whole-genome oxford nanopore sequencing. *Molecular*
1071 *Ecology Resources* e14094 (2025).
- 1072
- 1073 [21] Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification
1074 of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed
1075 rolling circle amplification. *Genome Research* 11, 1095–1099 (2001).
- 1076
- 1077 [22] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement
1078 amplification. *Computational and Structural Biotechnology Journal* 1688–1696
1079 (2023).
- 1080
- 1081 [23] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric
1082 sequences in long-read sequencing data. *Briefings in Bioinformatics* 24, bbad275
1083 (2023).
- 1084
- 1085 [24] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using
1086 single-molecule sequencing. *Nature Methods* 461–468 (2018).
- 1087
- 1088 [25] Smolka, M. *et al.* Detection of mosaic and population-level structural variants
1089 with sniffles2. *Nature Biotechnology* 1571–1580 (2024).
- 1090
- 1091 [26] Chen, Y. *et al.* Deciphering the exact breakpoints of structural variations using
1092 long sequencing reads with DeBreak. *Nature Communications* 283 (2023).
- 1093
- 1094 [27] Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped
1095 long reads. *Bioinformatics* 2907–2915 (2019).
- 1096
- 1097 [28] Jiang, T. *et al.* Long-read-based human genomic structural variation detection
1098 with cuteSV. *Genome Biology* 189 (2020).
- 1099
- 1100 [29] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and
1101 genotyping. *Nature Reviews Genetics* 12, 363–376 (2011).
- 1102
- 1103 [30] Kiguchi, Y., Nishijima, S., Kumar, N., Hattori, M. & Suda, W. Long-read
1104 metagenomics of multiple displacement amplified DNA of low-biomass human gut
phageomes by SACRA pre-processing chimeric reads. *DNA Research* 28, dsab019
1104 (2021).

- [31] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019). 1105
1106
1107
- [32] Mahmoud, M. *et al.* Structural variant calling: The long and the short of it. *Genome Biology* **20**, 246 (2019). 1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
- [33] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 287–297 (2025). 1110
1111
1112
1113
1114
1115
1116
1117
1118
- [34] Zhou, Z. *et al.* *DNABERT-2: Efficient foundation model and benchmark for multi-species genomes*, 1–24 (OpenReview.net, 2024). 1119
1120
1121
- [35] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution*, 43177–43201 (Curran Associates, Inc., 2023). 1119
1120
1121
- [36] Consens, M. E. *et al.* To transformers and beyond: Large language models for the genome (2023). [arXiv:2311.07621](https://arxiv.org/abs/2311.07621). 1119
1120
1121
- [37] Routhier, E. & Mozziconacci, J. Genomics enters the deep learning era. *PeerJ* **10**, e13613 (2022). 1122
1123
1124
- [38] Li, Y. *et al.* A genomic language model for chimera artifact detection in nanopore direct rna sequencing. *bioRxiv* (2024). URL <https://www.biorxiv.org/content/early/2024/10/25/2024.10.23.619929>. 1125
1126
1127
- [39] Poli, M. *et al.* *Hyena hierarchy: Towards larger convolutional language models*, 28043–28078 (PMLR, 2023). 1128
1129
1130
- [40] PLC., O. N. Dorado. <https://github.com/nanoporetech/dorado> (2023). 1131
1132
- [41] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embojournal* **17**, 10–12 (2011). 1133
1134
1135
- [42] Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 3094–3100 (2018). 1136
1137
1138
- [43] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* giab008 (2021). 1139
1140
1141
- [44] De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023). 1142
1143
- [45] Paszke, A. *et al.* *PyTorch: An imperative style, high-performance deep learning library*, 8024–8035 (Curran Associates, Inc., 2019). 1144
1145
1146
- [46] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning. GitHub repository (2019). URL <https://github.com/Lightning-AI/lightning>. 1147
1148
1149
1150

- 1151 [47] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (OpenRe-
1152 view.net, 2019).
- 1153
- 1154 [48] Yadan, O. Hydra - a framework for elegantly configuring complex applications.
1155 GitHub repository (2019). URL <https://github.com/facebookresearch/hydra>.
- 1156
- 1157 [49] Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for
1158 germline and cancer sequencing applications. *Bioinformatics* 1220–1222 (2016).
- 1159
- 1160 [50] Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end
1161 and split-read analysis. *Bioinformatics* i333–i339 (2012).
- 1162
- 1163 [51] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and
1164 indels by local assembly. *Genome Research* 581–591 (2018).
- 1165
- 1166 [52] Guo, Q., Li, Y., Wang, T.-Y., Ramakrishnan, A. & Yang, R. OctopusV and
1167 TentacleSV: A one-stop toolkit for multi-sample, cross-platform structural variant
1168 comparison and analysis. *Bioinformatics* btaf599 (2025).
- 1169
- 1170 [53] English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J.
1171 Truvari: Refined structural variant comparison preserves allelic diversity. *Genome
Biology* **23**, 271 (2022).
- 1172
- 1173 [54] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in
1174 python. *Nature Methods* 261–272 (2020).
- 1175
- 1176 [55] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science &
1177 Engineering* 90–95 (2007).
- 1178
- 1179 [56] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source
Software* 3021 (2021).
- 1180
- 1181
- 1182
- 1183
- 1184
- 1185
- 1186
- 1187
- 1188
- 1189
- 1190
- 1191
- 1192
- 1193
- 1194
- 1195
- 1196