ChimeraLM: A genomic language model for detecting whole genome amplification artifacts in single-cell sequencing

Yangyang Li[1], Qingxiang Guo[1†], Rendong Yang[1,2*]

[1]Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, 60611, IL, USA.
[2]Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, 675 N St Clair St, Chicago, 60611, IL, USA.

*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu;
Contributing authors: yangyang.li@northwestern.edu;
qingxiang.guo@northwestern.edu;
†These authors contributed equally to this work.

Abstract

this is a abtract

Keywords: Whole Genomics Amplification, Genomic Language Model

1 Main

Single-cell genomics has revolutionized our understanding of cellular heterogeneity and development by enabling the characterization of individual cells rather than bulk populations [1, 2]. This approach has proven instrumental in uncovering rare cell types, tracking developmental trajectories, and identifying somatic mutations that drive disease progression. However, the limited amount of DNA present in a single cell, typically only a few picograms, poses significant technical challenges for comprehensive genomic analysis [3, 4].

To overcome this fundamental limitation, Whole Genome Amplification (WGA) has become an essential preprocessing step in single-cell genomic studies [5, 6]. Various

WGA techniques, including Multiple Displacement Amplification (MDA), Multiple Annealing and Looping-based Amplification Cycles (MALBAC), and other emerging methods, can amplify the entire genome from a single cell by several orders of magnitude, generating sufficient DNA material for high-coverage sequencing [7–9]. This amplification enables researchers to achieve the depth and breadth of coverage necessary for reliable variant calling, copy number analysis, and structural variation detection.

Despite its critical role in single-cell genomics, WGA introduces systematic artifacts that can significantly impact downstream analyses [10, 11]. Among the most problematic are chimeric sequences—artificial DNA constructs formed when DNA fragments from different genomic loci are erroneously joined during the amplification process [10, 11]. These chimeric artifacts can manifest as false-positive structural variations that do not exist in the original cell [10]. The presence of such artifacts poses a substantial challenge for accurate Structural Variation (SV) detection, potentially leading to misinterpretation of genomic rearrangements and their biological significance.

Current computational approaches for identifying WGA-induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [11, 12]. However, these methods often fail to distinguish between genuine structural variations and amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns or occur in repetitive genomic regions [13, 14]. The lack of robust artifact detection methods has limited the reliability of structural variant analysis in single-cell studies and hindered the full realization of single-cell genomics' potential.

To address these challenges, we developed ChimeraLM, a genomic language model specifically designed to detect chimeric artifacts introduced by whole genome amplification. By leveraging deep learning approaches to capture sequence patterns and contextual information in genomic reads [15–17], ChimeraLM can effectively distinguish between genuine biological sequences and WGA-induced chimeric artifacts. This approach represents a significant advancement in single-cell genomic analysis, offering improved accuracy in artifact detection and enabling more reliable structural variant analysis in single-cell studies. This methodology represents a significant advancement in single-cell genomic analysis, offering a principled approach to improve the reliability of structural variant detection and enable more precise characterization of genomic alterations in individual cells.

In this study, we present ChimeraLM, demonstrate its superior performance compared to existing methods, and illustrate its practical applications in genomic studies.

2 Results

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work. Authors are encouraged to include RIIDs where appropriate.

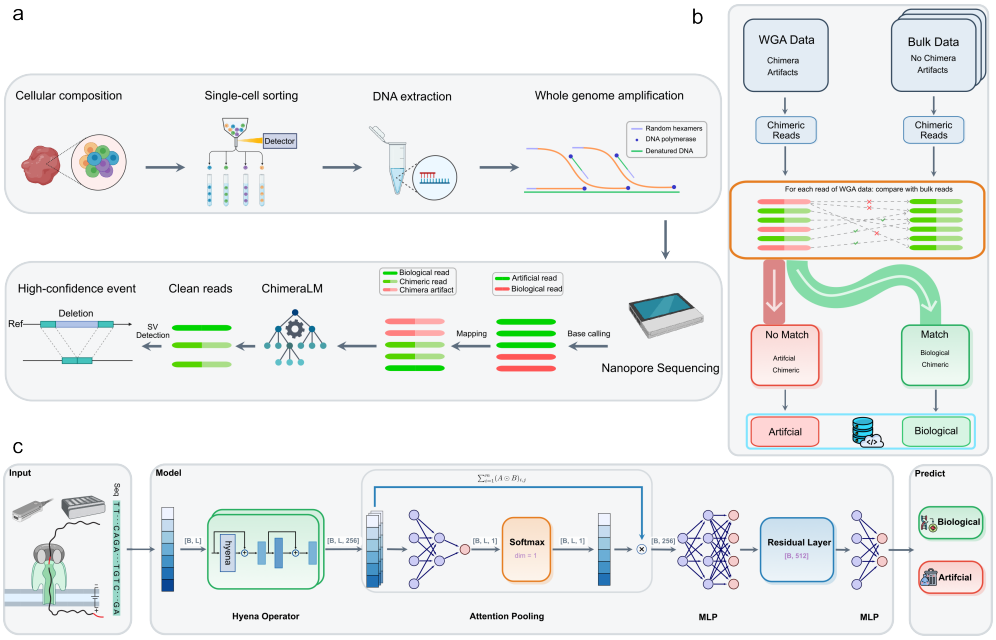

Fig. 1 Problem and Model

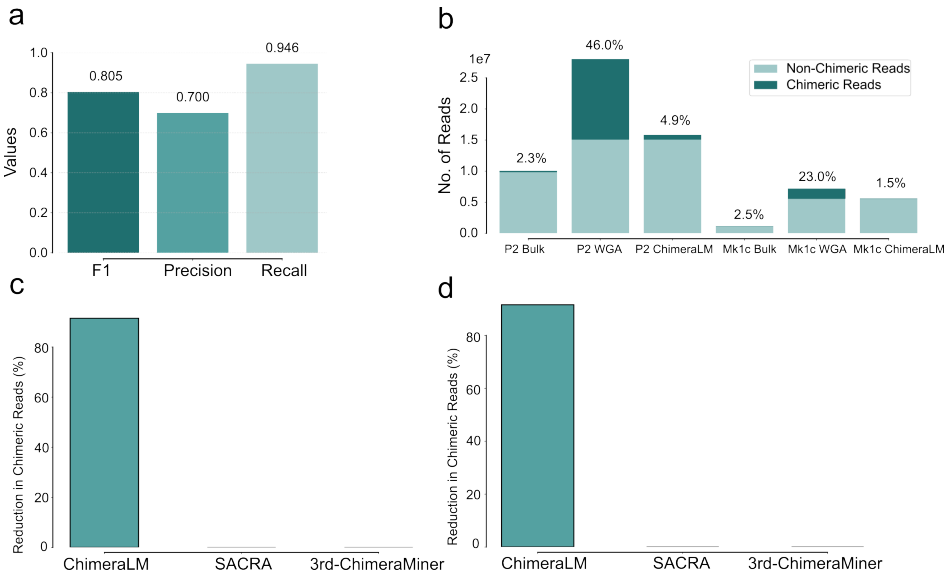

Fig. 2 Problem and Model

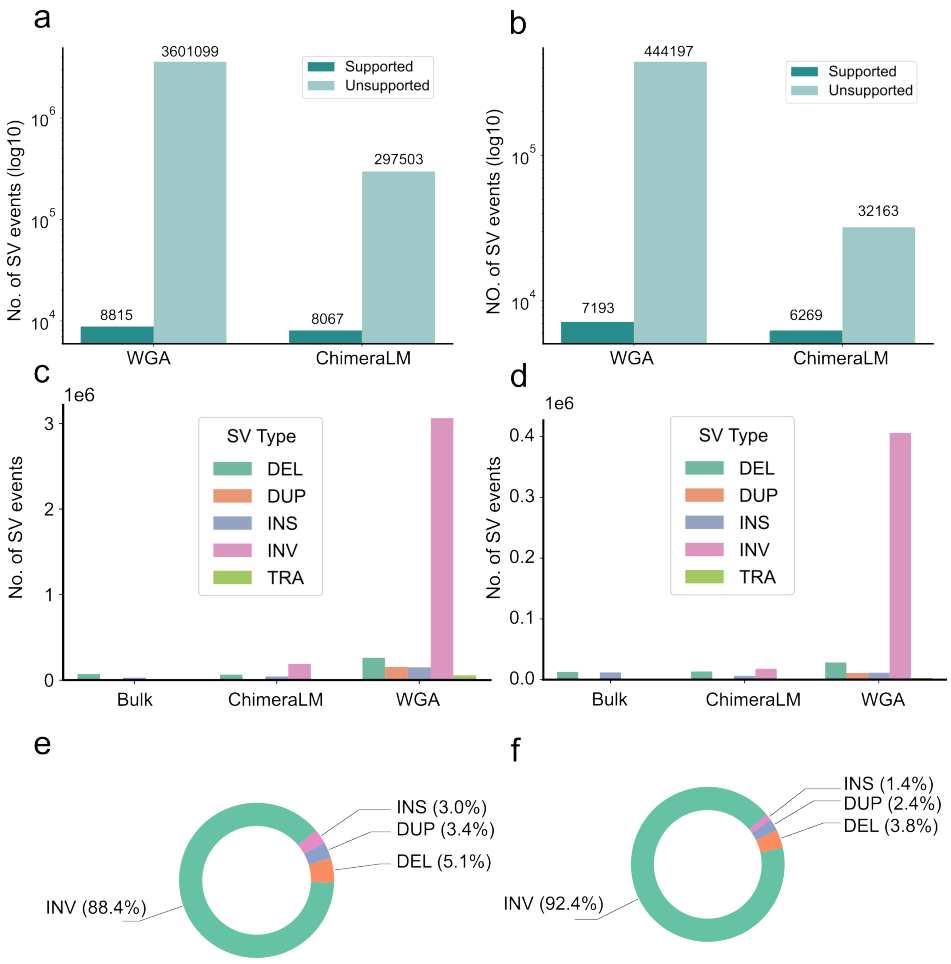

Fig. 3 Problem and Model

3 Methods

3.1 MDA sequencing

3.2 Train data construction

3.3 Model architecture

3.4 Model training

4 SV evaluation

Supplementary information. This separation aligns with how many transcript assembly algorithms work:

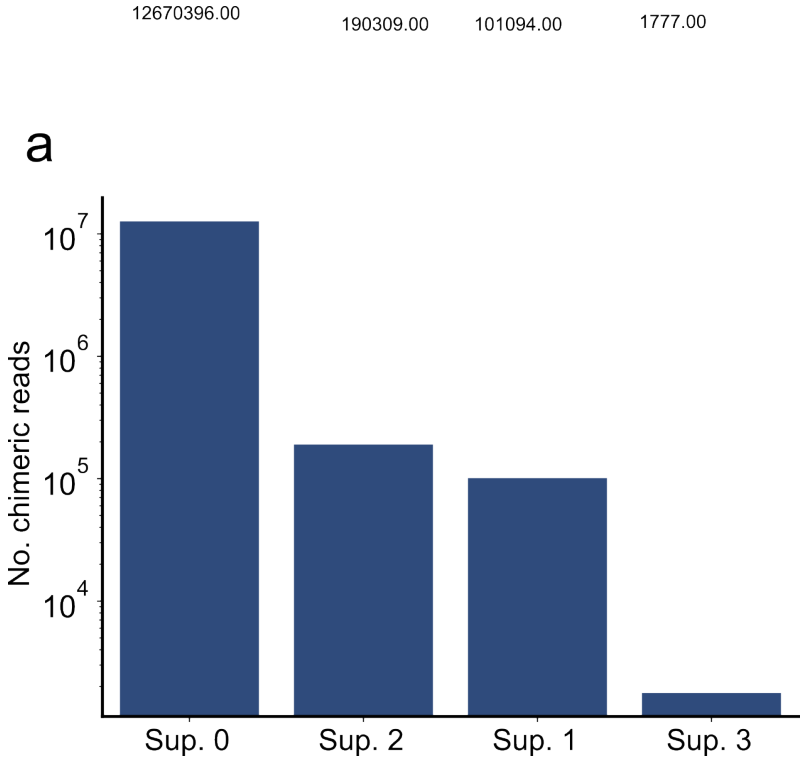

Fig. 4 Problem and Model

1. First, chains of exons and splice junctions are identified from the data
2. Then, potential transcripts are derived by traversing the graph in different ways
3. Finally, relationships between different transcript graphs are established

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
- Ethics approval and consent to participate

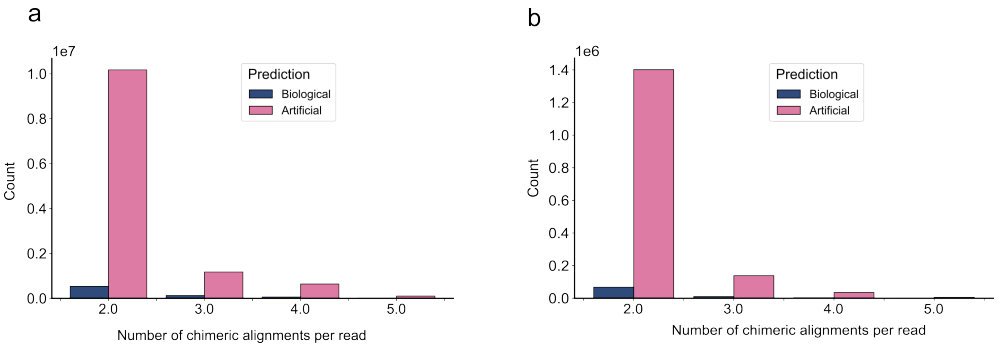

Fig. 5 Problem and Model

- Consent for publication
- Data availability
- Materials availability
- Code availability
- Author contribution

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Acronyms

MALBAC Multiple Annealing and Looping-based Amplification Cycles 2
MDA Multiple Displacement Amplification 2

SV Structural Variation 2

WGA Whole Genome Amplification 1, 2

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding

of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. *Communications Biology* **7**, 1288 (2024).

[2] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons. *Nature communications* **15**, 4220 (2024).

[3] Leung, M. L. *et al.* Highly multiplexed targeted dna sequencing from single nuclei. *Nature protocols* **11**, 214–235 (2016).

[4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**, 175–188 (2016).

[5] 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* **338**, 1622–1626 (2012).

[6] 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* **16**, 79–102 (2015).

[7] De Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PloS one* **9**, e105585 (2014).

[8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. *Scientific reports* **11**, 17171 (2021).

[9] Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences* **112**, 11923–11928 (2015).

[10] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal* **21**, 1688–1696 (2023).

[11] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275 (2023).

[12] 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).

[13] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology* **20**, 117 (2019).

[14] Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome biology* **20**, 246 (2019).

[15] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025).

[16] Zhou, Z. *et al.* Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006* (2023).

[17] Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems* **36**, 43177–43201 (2023).