ChimeraLM: A genomic language model to detect chimer artifa

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

Non-colinear Transcript (NLT) arising from structural variations challenge conventional linear analysis approaches in transcriptomics. Here we introduce the concept of Transcriptome Segment Graph (TSG) and present the first comprehensive infrastructure for graph-based transcript analysis. We develop the TSG file format as the first standardized encoding for transcript segment graphs, enabling systematic representation of NLT, and complex splicing patterns impossible to capture with linear methods. Our command-line toolkit provides the first comprehensive suite for TSG manipulation, analysis, and format conversion, while Aurora offers the first interactive visualization platform for transcript segment graphs. Together with our companion TSG caller scannls, this work establishes the foundational infrastructure for a new paradigm in transcriptomic analysis, providing the essential tools for systematic investigation of NLT structures.

Keywords: Transcriptomics, Graph-based analysis, Non-colinear transcripts

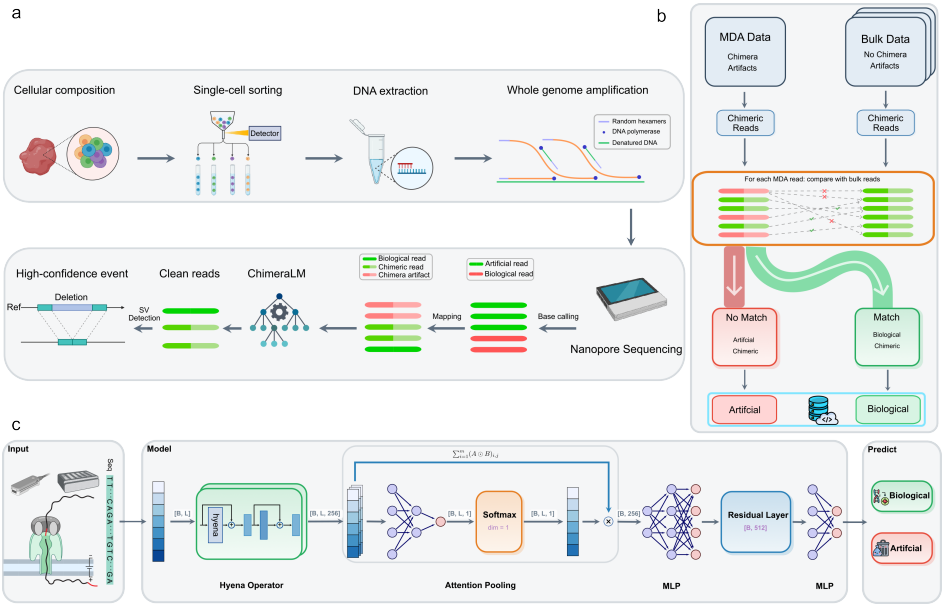

Fig. 1 Problem and Model

1 main

Transcriptomic structural variations, including gene fusions, circular RNAs, and complex alternative splicing, play critical roles in cancer, development, and disease.

2 Methods

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.

If your manuscript includes potentially identifying patient/participant information, or if it describes human transplantation research, or if it reports results of a clinical trial then additional information will be required. Please visit (https://www.nature.com/nature-research/editorial-policies) for Nature Portfolio journals, (https://www.springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/publishing-ethics/14214) for Springer Nature journals, or (https://www.biomedcentral.com/getpublished/editorial-policies#ethics+and+consent) for BMC.

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

1. First, chains of exons and splice junctions are identified from the data

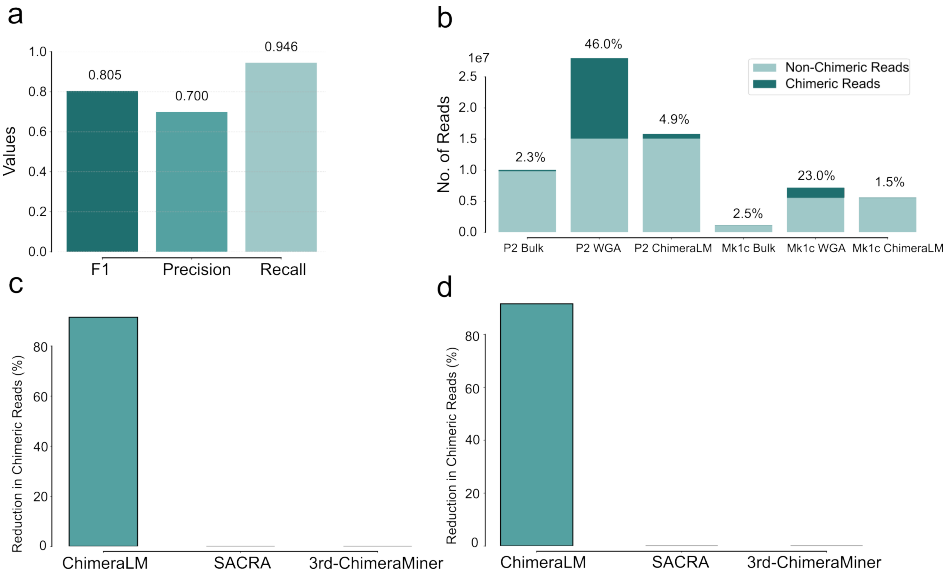

Fig. 2 Problem and Model

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

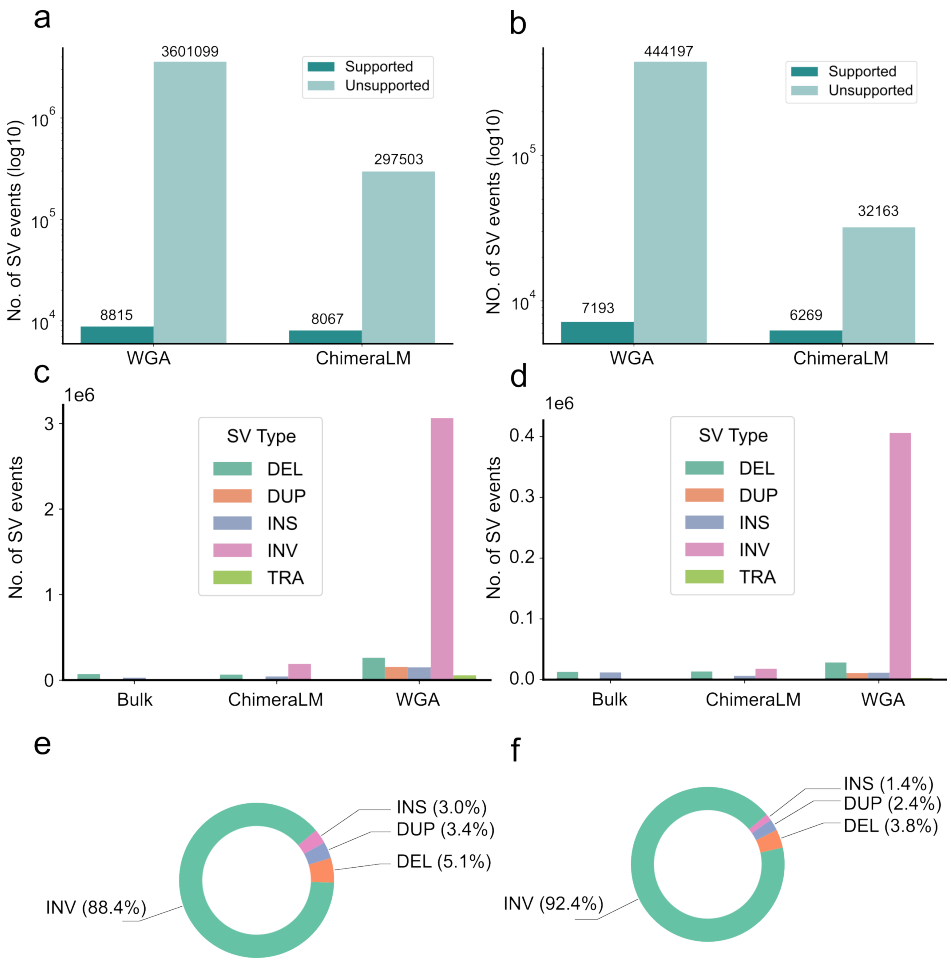

Fig. 3 Problem and Model

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

12670396.00 190309.00 101094.00 1777.00

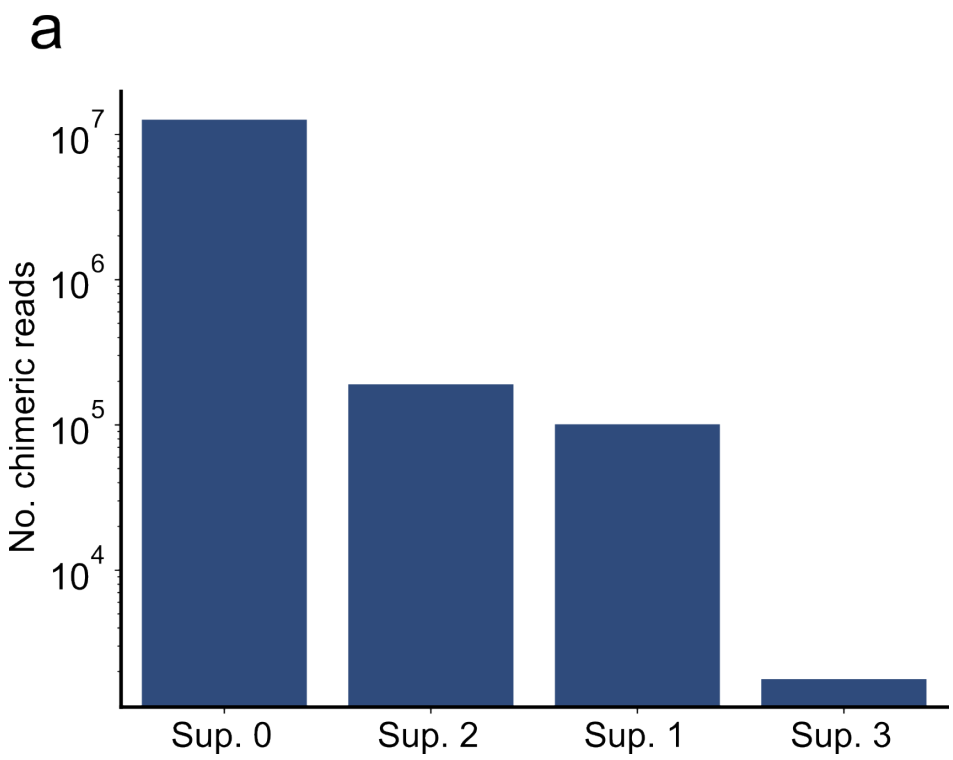

Fig. 4 Problem and Model

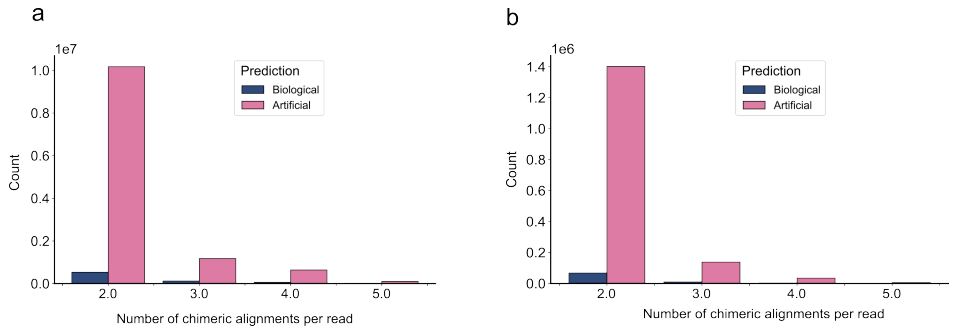

Fig. 5 Problem and Model

Acronyms

NLT Non-colinear Transcript 1

TSG Transcriptome Segment Graph 1

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