Fusion gene discovery in single cells from high throughput long read single cell transcriptomes

Cheng-Kai Shiau[1,2], Hsiao-Yun Lin[1,2], Timothy Pan[1, 2, 3], Xiaodong Lu[4], Minhua Wang[1,2], Yueying He[1, 2], Lina Lu[1, 2], Jindan Yu[4], Ruli Gao[1,2,3,*]

[1] Department of Biochemistry and Molecular Genetics, Northwestern University – Feinberg School of Medicine, Chicago, IL, USA 60611

[2] Center for Cancer Genomics, Robert H. Lurie Cancer Center, Northwestern University – Feinberg School of Medicine, Chicago, IL, USA 60611

[3] The Driskill Graduate Program, Northwestern University – Feinberg School of Medicine, Chicago, IL, USA 60611

[4] Department of Urology, Emory University - School of Medicine, Atlanta, GA, USA 30322

* Correspondence: ruli.gao@northwestern.edu (R.G.)

Abstract

High throughput long-read single cell RNA sequencing allows for fusion gene discovery in thousands of cells in parallel, which is however constrained by the lack of robust computational tools. Here, we develop LongFUSE that employs the XOR logic operation and stringent filtering criteria for the accurate detection of cell-specific fusion genes and their splicing isoforms. LongFUSE outperforms existing tools in both simulation and real data, demonstrating high accuracy and robustness.

25 **Main**

26 Gene fusion plays critical roles in driving tumorigenesis[1-4]. Historically, fusion gene analysis has
27 been limited to bulk levels, resulting in a mixed understanding of their oncogenic functions and
28 activation models due to cellular heterogeneity. The emerging long-read single cell RNA
29 sequencing (LR-scRNAseq) enables direct sequencing of full-length fusion transcripts in
30 thousands of cells in parallel. Existing computational tools for fusion gene detection[5-8] showed
31 high false discovery rates due to inadequate consideration of the analytical challenges in LR-
32 scRNAseq, such as the existence of chimeric reads derived from mis-ligated fragments (*e.g.,*
33 chimera artifacts)[9, 10], transcriptional read-through[11], *trans*-splicing[12] and the large amount of cell-
34 to-cell variations.

35 To address these limitations, we develop LongFUSE, a robust computational tool for the
36 accurate detection of fusion genes and their splicing isoforms in individual cells from LR-
37 scRNAseq data. LongFUSE devises XOR logic circuit to achieve fast identification of fusion
38 candidates among large numbers of high-softclipping reads, followed by a selection process for
39 true fusion genes with stringent filtering criteria. Critically, LongFUSE rectifies chimera artifacts,
40 mitigating false discoveries of fusion events and erroneous assignment of cell identities.
41 Moreover, LongFUSE detects exact genomic breakpoints by examining reads of nascent RNAs
42 and assembles potential splicing isoforms of fusion genes from reads mapped to mature RNAs.
43 Lastly, LongFUSE quantifies the expression levels of fusion genes and their splicing isoforms in
44 individual cells with unique molecular identifiers (UMIs). These endeavors ensure the precise
45 discovery of fusion genes at both single cell and single molecule levels.

46 The computational workflow of LongFUSE starts from scanning and curation of false-
47 ligated chimeras that often occur during the library preparation steps of long read sequencing[9, 10]
48 (**Fig 1a**). After curation, all singleton reads harboring true cell barcodes and unique molecule
49 identifiers are re-aligned onto the reference genome (GRCh38) using minimap2[13], followed by
50 XOR logic operation to select candidate reads that harbor fusion genes (**Fig. 1b-c**). In this process,

51 the multiple mapping results against reference genome of each high-softclipping read were
52 aggregated and summarized into a unified binary string using XOR logic circuit (**Online Methods**),
53 which simplifies conditional logic looping and favors speedy searching among large numbers of
54 long-read data. To select candidates for fusion-harboring reads, we derived two parameters, XOR
55 coverage and XOR gap, based on the unified binary string. Reads with high XOR coverage and
56 low XOR gap are expected to be derived from fusion transcripts (**Fig 1b**), whereas reads with low
57 XOR coverage and/or high XOR gap are likely due to low genomic complexity (*e.g.,* repetitive
58 insertions of transposable elements into multiple genomic regions) (**Fig 1c**). Next, we compiled a
59 collection of filters (**Online Methods**) to remove false discoveries from non-quantitative factors,
60 such as chimeric readouts of read-through events or fusion partners which were homologous.
61 Lastly, the list of confident fusion gene pairs in each cell are characterized through quantification
62 of expression levels, detection of genomic breakpoints and assembly of splicing isoforms (**Online**
63 **Methods**) to support a range of downstream isoform analyses across cell populations.

64 To benchmark the performance of LongFUSE, we simulated fusion-harboring LR-
65 scRNAseq data by spiking in 16 chemically synthesized fusion transcripts[8] into our in-house LR-
66 scRNAseq data of a human PBMC sample at different depths (**Online Methods**). In addition to
67 fusion-harboring reads, we spiked in a portion of synthetic chimera reads into a subset of cells to
68 investigate false discovery due to sequnecing artifacts. Next, we applied LongFUSE and three
69 leading fusion detection tools, including CTAT-LR-fusion[8], JAFFAL[5], and FusionSeeker[14] onto
70 these data to perform fusion gene discovery in individual cells. Comparative analyses revealed
71 that LongFUSE outperformed all these tools (**Fig. 1d**) and successfully detected all spike-in fusion
72 genes in all cells across various sequencing depths and rescued the true cell identifiers from the
73 chimera reads (**Fig. 1e**). All three existing tools showed a low capability in rescuing chimea
74 artifacts, leading to false discovery of fusion events and erroneous assignment of cell identifiers.
75 Both CTAT-LR-fusion[8] and JAFFAL[5] showed reasonable performance, with CTAT-LR-fusion[8]
76 failing to detect *TMPRSS2-ERG* and JAFFAL[5] failing to detect *TPM3-NTRK1* and *EGFR-*

77 *SEPTIN14* (**Fig. 1e**). As a side note, CTAT-LR-fusion[8] had initially detected *TMPRSS2-ERG*

78 harboring reads, but was filtered out by the default criteria, resulting in the false negative discovery

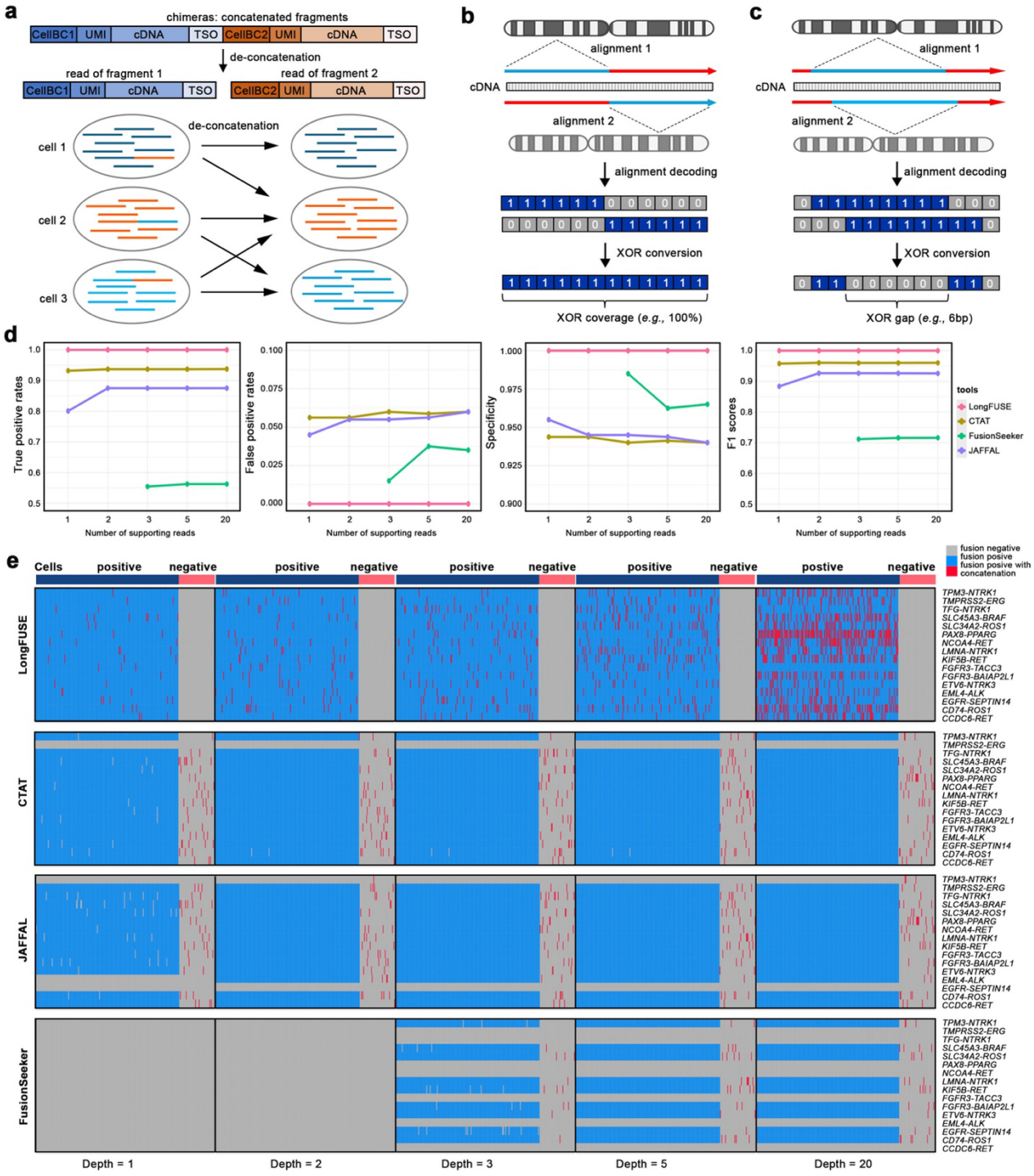

Figure 1 – Development of LongFUSE workflow and benchmarking its performance. **a**, Curation of chimera artifacts. XOR selection for candidate fusion genes (**b**), and repetitive low complexity elements (**c**). **d**, Comparison of LongFUSE performance with three existing tools. **e**, Heatmap of 16 spike-in fusion genes in single cells across depths and three existing fusion detection tools.

79 of a critical fusion event. Together, these data highlight the robustness and accuracy of LongFUSE
80 in detecting true fusion genes in single cells across various sequencing quality conditions.
81 Next, we applied LongFUSE onto an in-house LR-scRNAseq data of a prostate cancer
82 cell line, VCAP, known for harboring the *TMPRSS2-ERG* fusion. The sequencing data was
83 generated from fresh cell cultures using scNanoRNAseq, as previously published[15]. On average,
84 we obtained 26,983 consensus reads per cells, among which 3,775 showing high-softclipping
85 alignments were subjected to LongFUSE workflow. Remarkably, chimera reads concatenating
86 two or three random fragments accounted for large portions (63% ± 5% SD) of high-softclipping
87 reads (**Table S1**), while the percentages over total raw reads appeared lower (9% ± 0.5% SD)
88 (**Fig. 2a**). These data emphasized the critical importance of distinguishing chimera artifacts in
89 long-read single cell data analysis. After curation, all singleton reads were re-mapped onto the
90 reference and subjected to XOR operation to identify fusion candidates. The binned XOR
91 coverage showed a punctuated drop around 90%, suggesting a potential cutoff for fusion positive
92 events. Consistently, the XOR gap dropped to 10 nucleotides when XOR coverage increased
93 above 90% (**Fig. 2b**). To balance the tradeoff between sensitivity and specificity, we set the XOR
94 coverage parameter to be greater than 80% and XOR gap to be less than 30 nucleotides as the
95 initial criteria for selecting fusion candidates. Noteworthy, this step yielded a 60-fold enrichment
96 for fusion harboring reads from the high-softclipping read pools (**Fig. 2c-d, Table S1**), highlighting
97 the robustness of the XOR-based strategy in selecting true fusion genes. Next, fusion candidates
98 were subjected to a series of filtering criteria to remove additional false discoveries related to non-
99 quantitative factors, such as read-through events and homologous gene annotation artifacts.
100 Collectively, the miscellaneous collection of filtering steps demonstrated a 6-fold enrichment for
101 true fusion gene pairs in these cells (**Fig. 2c-d, Table S1**).
102 To facilitate analysis at cellular levels, LongFUSE classified fusion-positive cells based on
103 the existence of fusion-harboring reads and differentiated fusion-negative cells from ‘not defined’

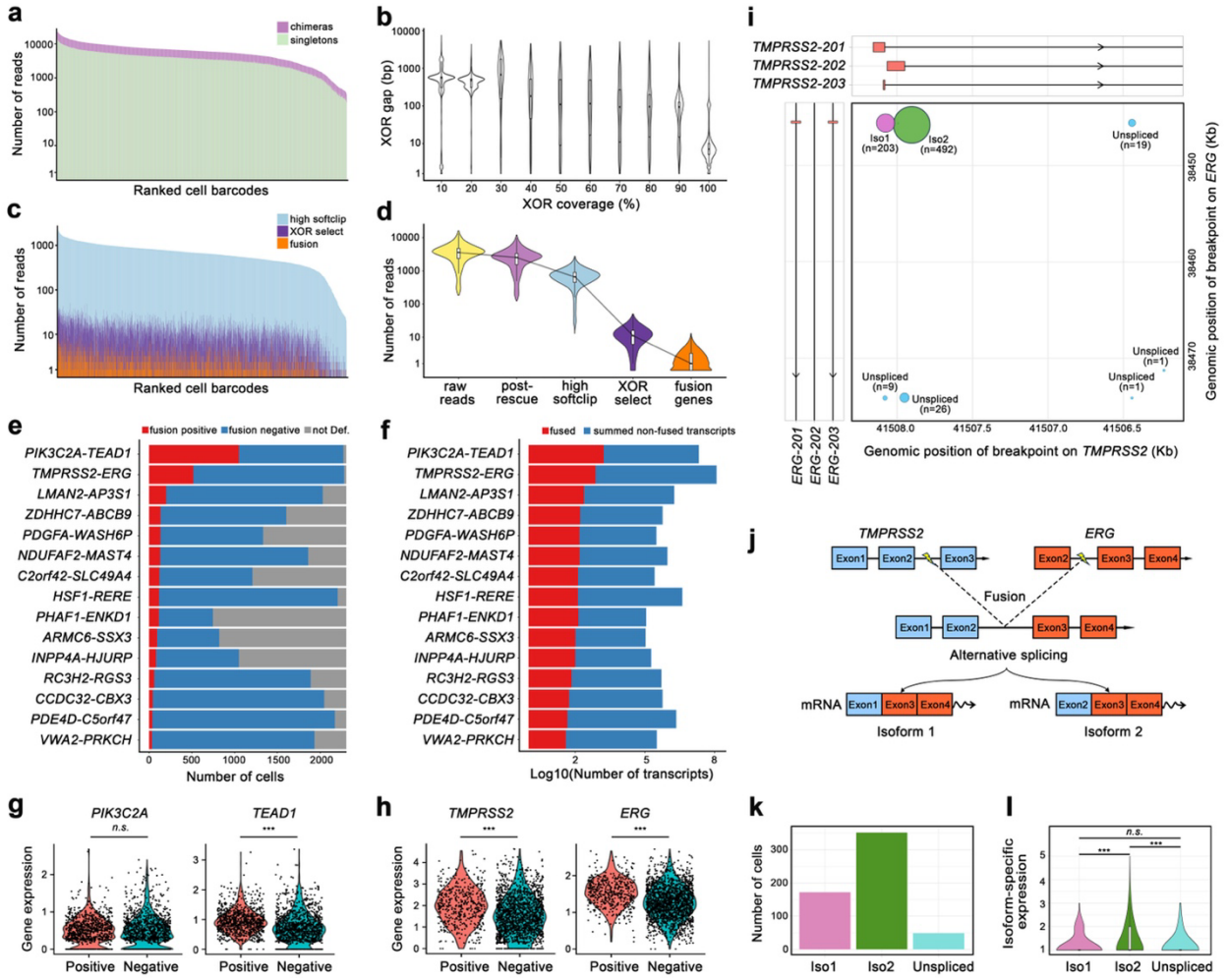

Figure 2 – Single cell fusion gene discovery in VCAP cell line. **a**, Composition of raw singleton and chimera reads in single cell. **b**, Distribution of XOR coverage and gap. **c**, Composition of post-curation read types. **d**, Sequential enrichment for fusion-harboring reads. **e**, Numbers of cells expressing fusion genes. **f**, Number of fusion transcripts and the summed wild-type transcripts. **g-h**, Expression levels of top two pairs of fusion partners. **i**, Supporting reads aligned to exact genomic positions. **j**, Inference of splicing isoforms. **k**, Cellular prevalence and **l**, expression levels of isoform-specific fusion transcripts.

cells by cross-checking the presence of wild-type transcripts of either fusion partners. Data showed that the *PIK3C2A-TEAD1* fusion was detected in 1,050 cells (*i.e.,* 38.5% cellular frequency), ranking it as the top expanded fusion event in this cell line, followed by the *TMPRSS2-ERG* fusion that was detected in 516 cells (*i.e.,* 18.9% cellular frequency) (**Fig. 2e**). Of note, the expression levels of fusion transcripts in fusion-positive cells were lower than the summed expression of both fusion partners in fusion-negative cells (**Fig. 2f**), but higher than the expression of individual fusion partners (**Fig. 2g-h**) with statistical significance (Wilcox *P-value*<0.001) except

111 for *PIK3C2A* (Wilcox *P-value* =0.09). These results showed-cased the unique power of
112 LongFUSE in studying the exact oncogenic dosage effects of fusion genes in single cells.

113 Next, we performed a detailed characterization of the *TMPRSS2-ERG* fusion that was
114 broadly reported in prostate cancer[3, 16, 17]. As expected, the precise breakpoints were detected in
115 the intronic regions between exons 2/3 of *TMPRS2* and exons 2/3 of *ERG* genes, supported by
116 a total of 56 pre-spliced reads (**Fig. 2i**). Interestingly, most reads were mapped to splice junction
117 regions with less than a 5-nucleotide distance to the closest exon boundaries, allowing for isoform
118 inference of the fusion transcripts. In total, 203 reads were mapped across the junctions between
119 exon 1 of *TMPRSS2* and exon 2 of *ERG*, while 492 reads covered junctions between exon 2 of
120 *TMPRSS2* and exon 2 of *ERG* (**Fig. 2i**). These reads were assembled into two major splicing
121 isoforms (Iso1 and Iso2) that resulted in distinct coding sequences of the *TMPRSS2-ERG* fusion
122 gene (**Fig. 2j**). The two isoforms showed differential cellular prevalence, with 203 cells having
123 utilized Iso1 contrasting to 352 cells that utilized Iso2 (**Fig. 2k**). Furthermore, Iso2 demonstrated
124 significantly higher expression compared to Iso1 (Wilcoxon *P-value*=6.13e-4). Together, these
125 results exemplify the unique strength of LongFUSE in supporting the detailed analysis of fusion
126 genes at both cellular and isoform levels.

127

128 In summary, we report a new and robust computational tool called LongFUSE for the
129 detection of fusion genes and their splicing isoforms in individual cells from high throughput LR-
130 scRNAseq data. LongFUSE outperforms existing tools, achieving high accuracy and robustness
131 through XOR logic operations and stringent filtering criteria. Importantly, it precisely reports coding
132 gene pairs, genomic breakpoints, expression levels, and splicing isoforms of fusion transcripts in
133 single cells, facilitating comprehensive downstream analyses of fusion genes across
134 heterogeneous cell populations.

135

136 **Data Availability**

The raw and processed scNanoRNAseq data generated in this study are deposited and freely accessible in the Gene Expression Omnibus (GEO) database under accession number GSE310978. The simulated fusion containing LR-scRNAseq data are deposited in Zenodo and can be accessed at: https://zenodo.org/records/17677124**.** The bulk long-read data of synthetic fusion transcripts were downloaded from Sequence Read Archive (SRA) under PRJNA1076207 (SRR27957791, SRR27957792, and SRR27957793).

Code Availability

All code and detailed documentation of LongFUSE methods are released at GitHub (https://github.com/gaolabtools/LongFUSE).

Acknowledgments

We thank research funding supports provided to R.G. by the National Institute of General Medical Sciences (NIH R35GM142539) and the National Heart, Lung, and Blood Institute (NIH 1R01HL160552).

Author Contributions

C.K.S. led the development of LongFUSE, performed the data analysis and wrote the manuscript. H.Y.L and M.W. performed sequencing experiments. T.P., X.L., Y.H., L.L., and J.Y. reviewed data, participated in discussions, and edited the manuscript. R.G. supervised the overall study, conceived the concepts, designed the computational tool, analyzed the data, and wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Online Methods

Cell line and tissue samples

165 The VCaP cell line was purchased from ATCC (cat. no. CRL-2876™). The PBMC sample was
166 collected from a heart transplantation patient at 22-month post-transplantation in Northwestern
167 Memorial Hospital in collaboration with Drs. Jane Wilcox and Arjun Shina. All activities of this
168 study complied with relevant ethical regulations of Institutional Biomedical Review Board at
169 Northwestern University.
170
171 **Long-read single cell RNAseq and data processing**
172 The long-read single cell RNAseq was performed using scNanoRNAseq method as we recently
173 published[15]. To do so, frozen cell stocks were thawed and resuspended in cell culture medium.
174 After cell filtering and counting, the single cell suspensions were loaded onto single cell capture
175 and barcoding platform (10X Genomics, Chromium iX, Single Cell 3' v3.1 protocol) to attach cell
176 barcodes (CellBCs) and unique molecular identifiers (UMIs) onto poly(A) tailed mRNAs of single
177 cells. Barcoded cDNAs were then enriched for library preparation using LSK114 kit (Oxford
178 Nanopore Technologies, ONT), and then sequenced on our in-house long read sequencer
179 (PromethION CapEX24, ONT) using R10.4 Q20+ flow cells, with one sample per flow cell. After
180 sequencing, the raw sequencing signals were base-called using Dorado (ONT) in super-accurate
181 mode) and stored in FASTQ format. The raw master FASTQ files were subjected to scNanoGPS
182 for demultiplexing into single cells and single molecules and mapping onto reference genome.
183 The reads harboring fusion genes are contained in high-softclipping reads that showed high
184 mappability onto multiple genomic regions. Single cell high-softclipping reads were extracted from
185 the mapped BAM files and subjected to LongFUSE workflow.
186
187 **De-concatenation of chimera reads of ligated fragments**
188 Long-read sequencing often generated unneglectable chimera reads of concatenated DNA
189 fragments that were caused by mis-ligation during library preparation steps[9, 10]. The chimera reads
190 contain multiple sets of barcodes, show high-softclipping mapping results and appear as

191 harboring fusion genes, causing significant amounts of false discovery in fusion gene detection.
192 Leveraging the expected patterns of barcode and adaptor sequences, LongFUSE identified and
193 broken down the chimeras into singleton reads, as such each full-length reads only contain one
194 cDNA fragment along with the true cell barcode and unique molecular identifiers. After curation,
195 all singleton reads were re-aligned against the reference genome with minimap2[13]. The singletons
196 with high-softclipping mapping results were then extracted and used for downstream detection of
197 fusion transcripts in single cells.
198
199 **Selection of fusion candidates with XOR logic operation**
200 To improve the speed and simplify logic looping, the XOR logic circuit principles were adopted to
201 select candidate fusion-harboring reads among high-softclipping reads. To do so, the two
202 mapping-quality strings corresponding to the top two mapped regions were first converted to
203 individual binary strings, where highly mappable loci were given 1 or 0 otherwise. Next, the two
204 binary strings of each read were unified to generate a new string through exclusive OR (XOR)
205 logic gate computing, in which 1 was assigned to a position if the values on the two initial strings
206 were different or 0 if the two initial values were same (both 0, or both 1). As such, the multiple
207 alignment conditions and outcomes were aggregated and summarized into a unified binary string,
208 which 1 represented high mappability of only one genomic region and 0 represented several
209 known and unknown possibilities, for instance, both high or both low mappability onto the top two
210 genomic regions or the certain loci had low data quality. This XOR operation simplified conditional
211 looping and greatly favored speedy search for fusion candidates among large numbers of long-
212 read data. Next, we derived two parameters to select candidate fusion-harboring reads, e.g., XOR
213 coverage and XOR gap. XOR coverage calculates the percentages of all 1s in the unified logic
214 strings, associating with the confidence of unique mapping to two distinct genomic regions. XOR
215 gap calculates the number of 0s between the boundaries of two consecutive 1s, associating with
216 the continuity between two genomic mapping and potential data quality related insertion and

217 deletions. Based on the empirical observations and trade-off between sensitivity and specificity,
218 XOR coverage ≥ 80% and XOR gap ≤ 30nt were set as the default cutoff and was left open for
219 users to adjust. In our tested data, these settings showed ~100X enrichment for fusion-harboring
220 reads.
221
222 **Miscellaneous filtering steps**
223 To boost speeds, singleton high-softclipping reads from unassembled scaffolds/contigs,
224 mitochondrial chromosomes, and other non-genomic regions were excluded from XOR operation.
225 Reads with discordant alignment directions or high-abundance multiple alignments were also set
226 aside to improve detection confidence. The candidate fusion gene pairs within 3 Kb regions of
227 two adjacent genes were removed to exclude potential read-through events. The candidate fusion
228 gene pairs belonging to paralog/homolog genes according to database were also filtered out to
229 mitigate annotation errors. Lastly, candidate fusion detected in less than certain numbers of cells
230 were filtered out to remove unknown random errors. The default cutoff was set as 3 cells, which
231 was open for users to adjust.
232
233 **Simulation of fusion containing LR-scRNAseq data**
234 The fusion containing LR-scRNAseq data were simulated by mixing spike-in fusion reads with an
235 in-house data of a human PBMC sample. The nanopore sequencing data of sixteen chemically
236 synthesized Spike-In fusion transcripts were downloaded from a previously published[8]. The
237 fusion-harboring reads with confident alignment were extracted, attached with barcode
238 sequencing using same library structures as of LR-scRNAseq, and then *in silico* mixed into the
239 PBMC dataset at different read depths (N=1, 2, 3, 5, and 20). Similarly, chimera reads were
240 synthesized to have multiple sets of cDNA fragments and barcodes and then injected into the LR-
241 scRNAseq data at different read depths.
242

Benchmarking with existing tools

The performance of LongFUSE was compared with three leading fusion gene detection tools, including CTAT-LR-fusion[8], JAFFAL[5], and FusionSeeker[14] using the simulated datasets. All three existing tools were run under default parameters. The calculation results of the sixteen Spike-In fusion genes in individual cells were compared between the four tools, by calculating true positive rate (TPR), false positive rate (FPR), specificity, and F1 score.

Detection of exact fusion breakpoints and splicing isoforms

To detect breakpoints, the XOR binary string was computed with AND logic gate separately against two alignment binary strings converted from two fusion partner genes. After AND logic gate computing, the last (right most) 1 was identified as the breakpoint of the upstream (left) alignment (denoted as gene A), while the first (left most) 1 was detected as the breakpoint for the downstream (right) alignment (denoted as gene B). The chromosomal positions of the two breakpoints were computed accordingly. To infer out possible splicing isoform, fusion reads were aligned against provided genome annotation (GTF). All possible isoforms fully matched were reported, along with matching isoform ID and exon ID combinations.

References

1. Rowley, J.D. Chromosome translocations: dangerous liaisons revisited. *Nat Rev Cancer* **1**, 245-250 (2001).
2. Chen, X. et al. Fusion gene map of acute leukemia revealed by transcriptome sequencing of a consecutive cohort of 1000 cases in a single center. *Blood Cancer J* **11**, 112 (2021).
3. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).
4. Tomlins, S.A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599 (2007).
5. Davidson, N.M. et al. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol* **23**, 10 (2022).
6. Jin, Z. et al. Single-cell gene fusion detection by scFusion. *Nat Commun* **13**, 1084 (2022).
7. Liu, Q. et al. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics* **21**, 793 (2020).
8. Qin, Q. et al. Accurate fusion transcript identification from long- and short-read isoform sequencing at bulk or single-cell resolution. *Genome Res* **35**, 967-986 (2025).

9. Li, Y. et al. A Genomic Language Model for Chimera Artifact Detection in Nanopore Direct RNA Sequencing. *bioRxiv* (2024).
10. Hard, J. et al. Long-read whole-genome analysis of human single cells. *Nat Commun* **14**, 5164 (2023).
11. Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T. & Steitz, J.A. Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell* **59**, 449-461 (2015).
12. Gingeras, T.R. Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206-211 (2009).
13. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
14. Chen, Y. et al. Gene Fusion Detection and Characterization in Long-Read Cancer Transcriptome Sequencing Data with FusionSeeker. *Cancer Res* **83**, 28-33 (2023).
15. Shiau, C.-K. et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nature Communications* **14**, 4124 (2023).
16. Demichelis, F. et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* **26**, 4596-4599 (2007).
17. Kron, K.J. et al. TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat Genet* **49**, 1336-1345 (2017).