Leveraging Sample-Specific Strings to Enhance Fusion Transcript Detection

Luísa de Melo Barros Penze[1], **Lucas Peres Oliveira**[1], **João Meidanis**[1]

[1] Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Av. Albert Einstein, 1251 – 13083-852 – Campinas – SP – Brazil

{l238001,l265193}@dac.unicamp.br, meidanis@unicamp.br

Abstract. *Fusion transcripts are widely used biomarkers (e.g., in cancer diagnosis), but current methods to detect them from long-read sequencing data still yield a high number of false positives. In this study, we attack this problem by selecting reads containing absent strings in a reference transcriptome — sample-specific strings (SFS) — before applying a fusion detection method. We adapted an existing SFS retrieval algorithm to long-read RNA-seq data and applied it to select reads in eight simulated datasets. These reads were then fed to three transcript fusion detection methods — LongGF, JAFFAL, and CTAT-LR. Our results show that SFSs capture the reads with relevant information and, hence, improve the accuracy of most of these fusion detection methods.*

1. Introduction

Fusion transcripts arise from abnormal genetic events that join the segments of distinct genes and result in modified transcripts that often translate into defective proteins. These events are among the major drivers of cancer and, hence, are useful biomarkers for diagnosis and targets for therapy. Accurate detection of fusion transcripts is, therefore, a crucial step in cancer research and clinical practice. With recent improvements in long-read sequencing (LRS) technology, several computational methods have been developed to detect fusion transcripts from RNA LRS data [Liu et al. 2020, Chen et al. 2023, Davidson et al. 2022, Karaoglanoglu et al. 2022, Qin et al. 2025]. A challenge faced in these analyses is discerning the false positives retrieved by these methods.

In this paper, we show that the accuracy of two fusion transcript detection methods — JAFFAL [Davidson et al. 2022] and LongGF [Liu et al. 2020] — can be improved by restricting the set of input reads to those containing strings *absent* in a reference transcriptome [Khorsand et al. 2021, Denti et al. 2023]. We also attempted this strategy with a third fusion detection method — CTAT-LR [Qin et al. 2025] —, but it did not yield improvements. Specifically, we adapted the so-called "Ping-Pong algorithm" — originally proposed for double-stranded DNA — to single-stranded RNA LRS reads [Khorsand et al. 2021]. Then, we ran the three aforementioned fusion transcript detection methods on eight simulated datasets under two settings: using all the reads and using only those retrieved by the modified algorithm. Our experiments demonstrated that the number of false positives retrieved by LongGF and JAFFAL decreased in the latter setting, but at the cost of increased processing time. On the other hand, the number of false positives and false negatives retrieved by CTAT-LR increased under the same setting.

2. Methods

2.1. Sample-specific strings

Consider the alphabet $\Sigma = \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$. Let $T = T[1..n]$ and $W = W[1..m]$ be strings over Σ. A substring $T[i..j]$ of T, for $1 \leq i \leq j \leq n$, is *proper* if $i > 1$ or $j < n$. We say that W *occurs* in T if it is a substring of T; otherwise, it is *absent* in T. Furthermore, W is *minimal absent* in T if all proper substrings of W occur in T. A substring of W is *W-specific with respect to* T if it is minimal absent in T.

We fix T to be a reference transcriptome and W a read from a sequenced sample. There are two important reasons for using sample-specific strings (with respect to a reference transcriptome) in the context of this paper. First, they are not length-limited, making them ideal for spanning structural variants, such as those that originate fusion transcripts. Second, the minimality property allows them to be efficiently computed and stored because the number of sample-specific strings is at most $O(m)$, whereas the number of generic absent strings in T that occur in W can be $\Omega(m^2)$.

In order to compute sample-specific strings with respect to a reference transcriptome, we adapted the SFS retrieval algorithm implemented in the software tool SVDSS [Khorsand et al. 2021, Denti et al. 2023]. The original algorithm indexes T using the FMD-index, an implementation of the FM-index designed for double-stranded DNA data that considers reverse complements for efficient searches [Ferragina and Manzini 2000, Li 2012, Li 2024]. Since we are focusing on single-stranded RNA data, we optimized the code to use the functionality of a plain FM-index and ignore reverse complements. Code dependencies were updated to ensure compatibility and the modified version is available at the repository `https://github.com/meidanis-lab/ropebwt3`. Nevertheless, we used both the FMD- and FM-index-based algorithms in our experiments.

2.2. Simulated datasets

We used five simulated datasets that simulate data generated by nanopore sequencing [Davidson et al. 2022, Davidson 2021]. Each dataset was generated under different mean sequence qualities — 75%, 80%, 85%, 90%, and 95% — and have previously annotated fusion transcripts. All of the reads in these datasets cover a fusion transcript; hence, to better reflect real datasets, we created three additional datasets based on the 75%-, 85%-, and 95%-sequence-quality datasets by inserting synthetic reads with no fusions. These reads were generated using the Badread simulator using the human reference transcriptome, version hg38, as input [Wick 2018]. The reads from the original datasets account for approximately 1% of the augmented datasets, while the synthetic reads account for the remaining 99%. Hereafter, we refer to the original datasets as *fusion-only data* and the augmented datasets as *mixed data*.

2.3. Fusion transcript detection pipeline

We implemented a pipeline to automate the execution of the SFS retrieval algorithms and of three fusion transcript detection tools — LongGF, JAFFAL, and CTAT-LR. These tools were selected because they provide thorough documentation and have been tested in previous studies [Chen et al. 2023, Karaoglanoglu et al. 2022].

The pipeline runs the aforementioned tools on the fusion-only and mixed data under three settings: without selecting SFS before fusion calling; selecting SFS before fusion calling with the original FMD-index-based algorithm; and selecting SFS before fusion calling with our FM-index-based algorithm. For brevity, we will refer to the SFS retrieval algorithms as filters (indicating that they remove the reads without SFS from the analysis). For each setting, we computed the total number of errors — the sum of the number of false positives and false negatives — and the total execution time, with and without the filters, averaged over three runs. In this context, false positives correspond to fusion transcripts retrieved by the tools, but absent in the ground truth (the previously annotated fusions), and false negatives for those not retrieved by the tools, but present in the ground truth.

3. Results

3.1. Accuracy of fusion detection tools

Figure 1 and Figure 2 show the number of false positives (FPs) and number of false negatives (FNs) computed for the mixed data and fusion-only data, respectively, under the settings described in Section 2.3.

In the mixed data (Figure 1), FNs retrieved by JAFFAL and LongGF remained stable in all scenarios, indicating that the filters had minimal impact on their ability to detect true fusions. For CTAT-LR, however, FNs increased significantly when the filters were applied. As for FPs, we observed that the filters reduced their retrieval by JAFFAL and LongGF. For the datasets with higher quality, this improvement was more evident, which is expected because the retrieved SFS are less likely to have originated from sequencing artifacts. In contrast, FPs retrieved by CTAT-LR increased when the filters were applied.

In the fusion-only data (Figure 2), FNs increased for all datasets after applying the filters, regardless of sequence quality. As for FPs, we observed the same pattern of the mixed data: it increased significantly for CTAT-LR, but reduced for JAFFAL and LongGF; the improvement on the latter was more evident for higher sequence qualities.

A possible explanation for the worse performance of CTAT-LR after restricting its input is an internal routine used in this tool that leverages contextual information on the vicinity of the reads spanning a fusion transcript's breakpoint. By removing reads without SFS, this mechanism may have been disrupted, hindering CTAT-LR's ability to accurately distinguish true fusions from artifacts. In contrast, LongGF and JAFFAL do not rely on this contextual information and, therefore, may have benefited from the selection of reads containing SFS.

3.2. Execution time measurement

We measured the execution time of LongGF, JAFFAL, and CTAT-LR in all datasets under the settings described in Section 2.3. Figures 3 and 4 indicate that both filters introduce an overhead to compute SFS, but they also reduce the tools' processing times. In the simulated fusion-only dataset, the impact of filtering on tool time varied. For LongGF, the FM-index filter yielded better results only under high-quality conditions. In lower-quality data, the filtering step (both types of filter) just introduced time overhead. Regarding CTAT-LR, the FMD-index filter performed better than the no-filter condition across all sequence qualities, while the FM-index filter improved tool time in only two sequence

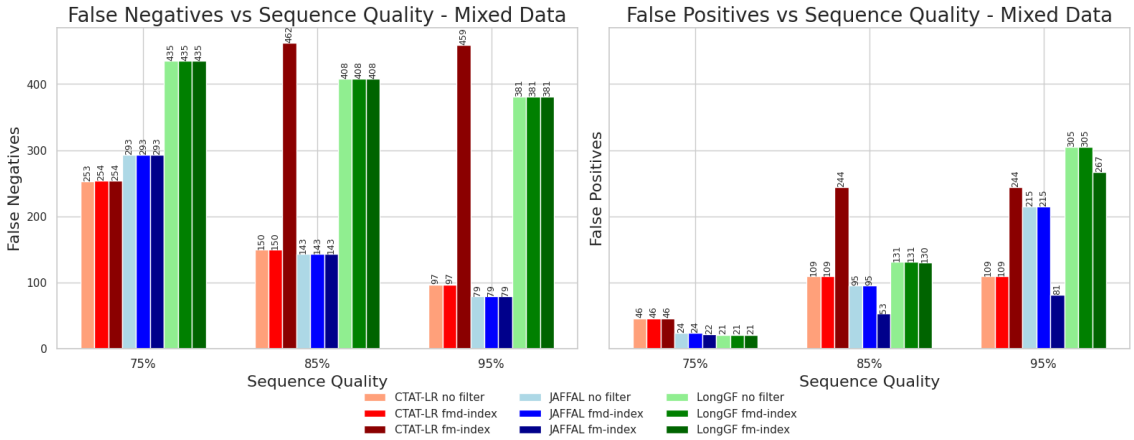

Figure 1. Total number of errors for mixed data. The vertical (y-axis) number scale is the same for both graphs

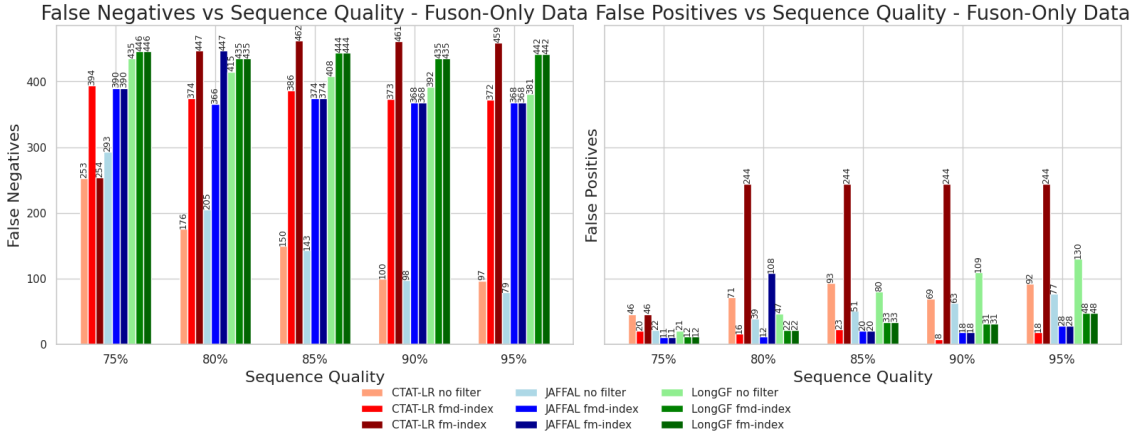

Figure 2. Total number of error for fusion-only data. The vertical (y-axis) number scale is the same for both graphs

quality scenarios. The same pattern was observed for JAFFAL, where FMD-index generally outperformed no filtering, and FM-index offered improvements only in a limited subset of conditions. In the case of the simulated mixed dataset, the FM-index filter showed broader effectiveness. For CTAT-LR and JAFFAL, the application of either filter — especially FM-index — resulted in lower tool execution times, indicating a favorable impact on performance. In contrast, LongGF did not benefit as consistently; in most mixed data scenarios, the no-filter condition led to better tool times than either filtering strategy. Additionally, for CTAT-LR, applying the FM-index filter also resulted in reduced total execution time, suggesting that the benefits of filtering extend beyond the tool's internal processing, improving the overall pipeline efficiency.

4. Conclusion and Future Work

This paper highlighted the potential of sample-specific strings (SFS) to enhance fusion transcript detection from LRS RNA-seq data. Our experiments showed that by restricting the input of JAFFAL and LongGF to reads that contain absent strings in a reference transcriptome, their accuracy improved on our mixed data, notably on reads with high

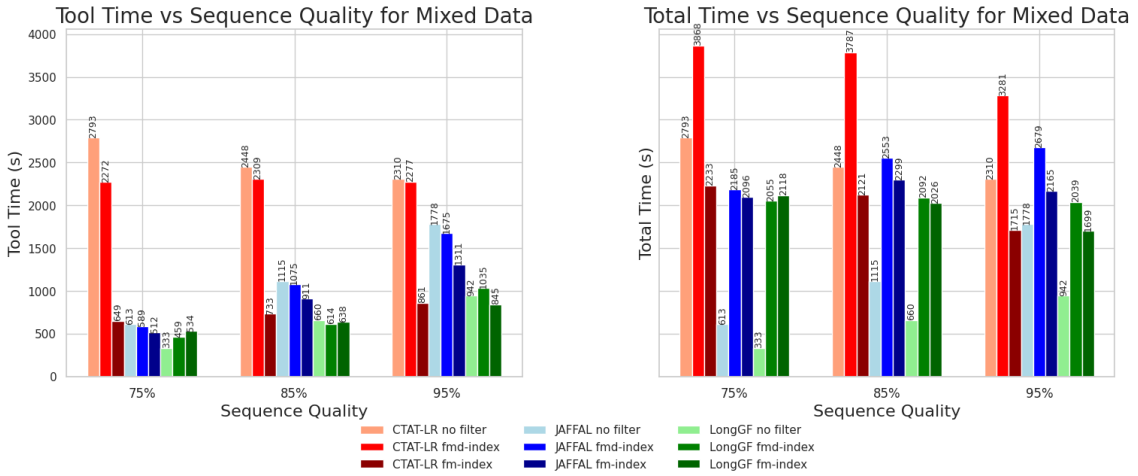

Figure 3. Execution time for mixed simulated data. 'Total time' refers to SFS filtering plus tool time. The vertical (y-axis) number scale is the same for both graphs

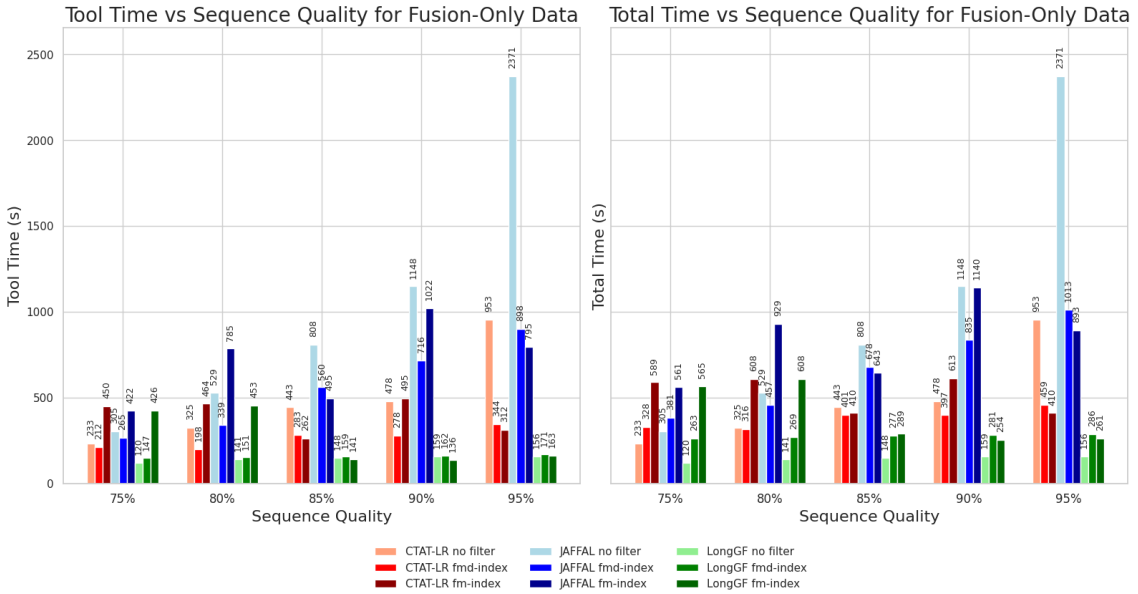

Figure 4. Execution time for fusion-only simulated data. 'Total time' refers to SFS filtering plus tool time. The vertical (y-axis) number scale is the same for both graphs

sequence qualities. This is an encouraging result because these inputs mimic real sequencing datasets more realistically. On the other hand, we observed a degradation in CTAT-LR when filtering the input reads, which warrants more study. A drawback of current SFS retrieval algorithms is their reliance on exact matching of strings, which limits their application to accurate reads. To widen the analysis of the proposed filtering strategy, future work will focus on applying it to real LRS RNA-seq data. Additionally, we will direct efforts toward further refining the FMD-index-based SFS retrieval algorithm for RNA-seq data.

5. Acknowledgements

Luísa de Melo Barros Penze is supported by grant #2024/14925-0, São Paulo Research Foundation (FAPESP); Lucas Peres Oliveira is supported by grant #2023/05887-5, São Paulo Research Foundation (FAPESP); João Meidanis is supported by grants #2018/00031-7 and #2024/01200-8, São Paulo Research Foundation (FAPESP).

References

Chen, Y., Wang, Y., Chen, W., Tan, Z., Song, Y., Human Genome Structural Variation Consortium, Chen, H., and Chong, Z. (2023). Gene Fusion Detection and Characterization in Long-Read Cancer Transcriptome Sequencing Data with FusionSeeker. *Cancer Research*, 83(1):28–33.

Davidson, N. (2021). Long read fusion simulation. `https://figshare.com/articles/dataset/Long_Read_Fusion_Simulation/14459007`. Accessed: 2024-08-10.

Davidson, N. M., Chen, Y., Sadras, T., Ryland, G. L., Blombery, P., Ekert, P. G., Göke, J., and Oshlack, A. (2022). JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biology*, 23(1):10.

Denti, L., Khorsand, P., Bonizzoni, P., Hormozdiari, F., and Chikhi, R. (2023). SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nature Methods*, 20(4):550–558.

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398.

Karaoglanoglu, F., Chauve, C., and Hach, F. (2022). Genion, an accurate tool to detect gene fusion from long transcriptomics reads. *BMC Genomics*, 23(1):129.

Khorsand, P., Denti, L., Human Genome Structural Variant Consortium, Bonizzoni, P., Chikhi, R., and Hormozdiari, F. (2021). Comparative genome analysis using sample-specific string detection in accurate long reads. *Bioinformatics Advances*, 1(1):vbab005.

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844.

Li, H. (2024). BWT construction and search at the terabase scale. *Bioinformatics*, 40(12):btae717.

Liu, Q., Hu, Y., Stucky, A., Fang, L., Zhong, J. F., and Wang, K. (2020). LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics*, 21(11):793.

Qin, Q., Popic, V., Wienand, K., Yu, H., White, E., Khorgade, A., Shin, A., Georgescu, C., Campbell, C. D., Dondi, A., Beerenwinkel, N., Vazquez, F., Al'Khafaji, A. M., and Haas, B. J. (2025). Accurate fusion transcript identification from long- and short-read isoform sequencing at bulk or single-cell resolution. *Genome Research*, 35(4):967–986.

Wick, R. R. (2018). Badread: Simulation of error-prone long reads. `https://github.com/rrwick/Badread`.