Data-aware metacaller for improved gene fusion detection in RNA-seq

Iga Ostrowska, Tomasz Gambin
Institute of Computer Science, Warsaw University of Technology

1 INTRODUCTION

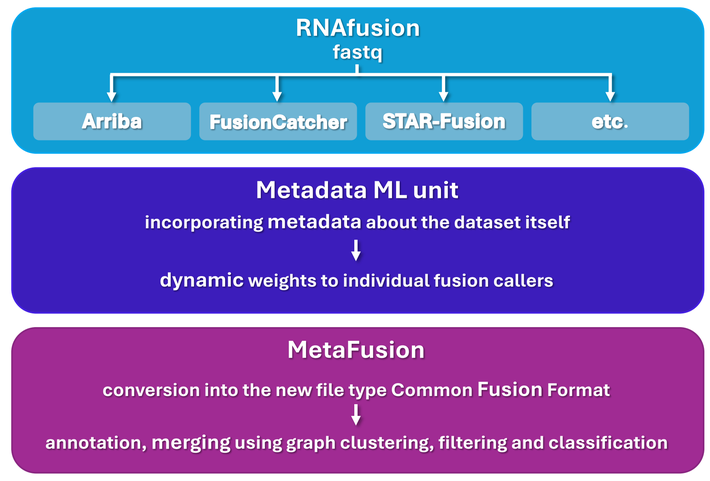
Gene fusions are critical drivers of tumorigenesis and serve as important diagnostic and therapeutic targets in oncology [1,2]. Detecting these events from RNA-seq data remains a complex bioinformatics task due to variable performance of existing tools and the influence of dataset-specific factors, such as sequencing protocol, read length, and cancer type [3, 4]. Current multi-caller and aggregation approaches improve robustness but do not account for dataset characteristics, resulting in suboptimal accuracy. To address this gap, we propose a **data-aware metacaller** that dynamically adapts fusion detection to the properties of the input data. Such an approach aims to enhance both the reliability and clinical relevance of fusion discovery in diverse cancer settings.

2 METHODS

We conducted a meta-analysis of **10 independent benchmarking studies**, identifying dataset-level features that strongly impact fusion caller performance. Based on these insights, we designed a hybrid system combining two established frameworks:

- **RNAfusion [5] (nf-core/rnafusion)**, a reproducible multi-caller pipeline for detecting gene fusions
- **MetaFusion [6]**, a graph-based ensemble method for result aggregation.

Our innovation introduces an additional metadata-aware layer that incorporates information such as sample type, sequencing characteristics, and quality metrics. We train a **machine learning model** on curated datasets from published studies to learn optimal weighting schemes for individual fusion callers, enabling adaptive prioritization according to dataset context.



The model was trained on real-world data comprising **56 cell lines representing multiple cancer types**, ensuring biological diversity and robustness of the learned weighting strategies.

3 RESULTS

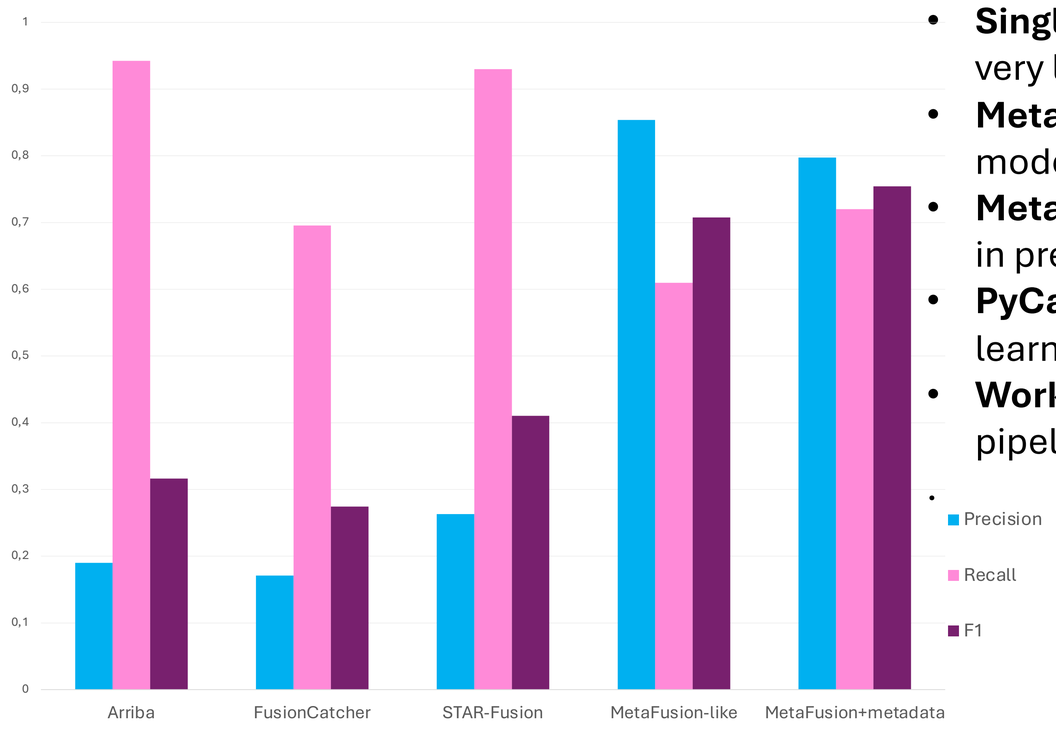


Figure 1. Comparison of Precision, Recall, and F1 metrics for Arriba, FusionCatcher, STAR-Fusion, MetaFusion-like, and MetaFusion-like+metadata, evaluated on a dataset of true sequencing samples from 56 cancer cell lines.

- **Single callers (Arriba, FusionCatcher, STAR-Fusion)** achieved high recall but showed very low precision and F1-scores, leading to many false positives.
- **MetaFusion-like aggregation** substantially increased precision and F1 while maintaining moderate recall.
- **Metadata-aware MetaFusion** further balanced performance, showing simultaneous gains in precision, recall, and F1 compared to both single callers and uniform aggregation.
- **PyCaret-based model selection** facilitated robust identification of suitable machine learning algorithms, ensuring reproducibility and adaptability across datasets.
- **Workflow compatibility:** the system integrates seamlessly with existing RNA-seq pipelines and scales across diverse cancer types.

Model	Recall	Prec.	F1
Random Forest Classifier	0,7199	0,7973	0,7542
Extra Trees Classifier	0,7193	0,7960	0,7531
K Neighbors Classifier	0,7134	0,7671	0,7355
Logistic Regression	0,6375	0,8395	0,7212
Decision Tree Classifier	0,6892	0,7487	0,7156

Table 1. Comparison of machine learning models in terms of Precision, Recall, and F1 score, evaluated on the fusion detection dataset from 56 cancer cell lines with metadata included.

4 DISCUSSION

Our findings demonstrate the importance of dataset-aware strategies in complex genomic analyses. By explicitly modeling the relationship between data characteristics and caller performance, the proposed framework overcomes limitations of current ensemble methods that treat all tools equally. Notably, incorporating cancer type and read-level features such as the number of split and junction reads led to encouraging improvements in F1-score, highlighting the value of feature-driven weighting. These results motivate further extensions of the system to include additional metadata, such as sequencing read length, sample size, or other specimen-specific parameters. Expanding the feature set may enable even greater adaptability across diverse experimental contexts and improve generalization to novel datasets.

5 CONCLUSSION

We present a **data-aware metacaller** that leverages metadata-driven machine learning to optimize gene fusion detection from RNA-seq data. This adaptive framework improves robustness, interpretability, and applicability across research and clinical contexts. By bridging benchmarking insights with predictive modeling, our system sets the stage for more reliable and customizable bioinformatics pipelines.

REFERENCES

1. Dorney R, et al. *Recent advances in cancer fusion transcript detection.* Brief Bioinform. 2023.
2. Ahmed J, et al. *Fusion challenges in solid tumors: shaping the landscape of cancer care in precision medicine.* JCO Precis Oncol. 2024.
3. Kerbs P, et al. *Fusion gene detection by RNA-sequencing complements diagnostics of acute myeloid leukemia and identifies recurring NRIP1-MIR99AHG rearrangements.* Haematologica. 2021.
4. Nikanjam M, et al. *Targeting fusions for improved outcomes in oncology treatment.* Cancer. 2020.
5. Apostolides M, et al. *MetaFusion: a high-confidence metacaller for filtering and prioritizing RNA-seq gene fusion candidates.* Bioinformatics. 2021.
6. nf-core/rnafusion. Version 3.0.2. Available at: https://nf-co.re/rnafusion/3.0.2

GitHub

Scan me!