UMIvar: Unveiling Insights in UMI Deduplication Software Benchmarking through UMI-aware Variant Simulation

D. G. Camblor[1], J. Martín-Arana[1,2], F. Gimeno-Valiente[3], B. García-Micó[1,2], B. Martínez-Castedo[1,2], J.F. Català-Senent[4], F. Robledo-Yagüe[4], S. Zúñiga-Trejos[4], A. Cervantes[1,2], N. Tarazona[1,2]

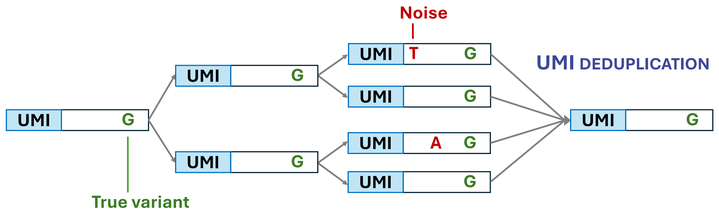
[1] *Department of Medical Oncology, INCLIVA Biomedical Research Institute, Valencia, Spain*
[2] *CIBERONC, Carlos III Health Institute, Madrid, Spain*
[3] *Cancer Evolution and Genome Instability Laboratory, University College London Cancer Institute, London, United Kingdom*
[4] *Bioinformatics Unit, INCLIVA Biomedical Research Institute, Valencia, Spain*

EACR2024-0891

EACR 2024 — European Association for Cancer Research

Background

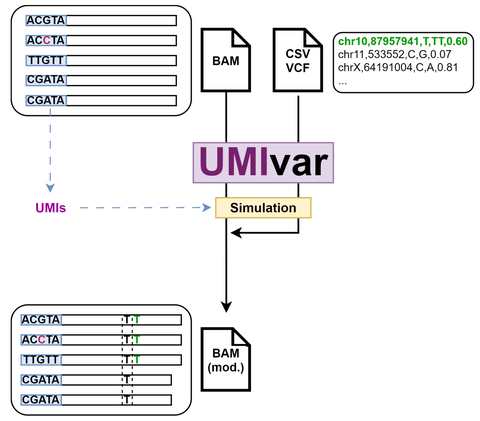
The use of **Unique Molecular Identifiers (UMIs)** in NGS variant calling analysis improves error correction and enhances the detection of low-frequency variants [1]. This is particularly important in relatively low depth settings (e.g., whole-exome sequencing) or in liquid biopsy variant calling (very low VAFs).



Different tools that handle UMI information are available, but they have not been properly compared in the variant calling setting: **there is a lack of accessible resources with ground truth variants and UMI information**. To address this, we developed **UMIvar**, a bioinformatic tool that simulates variants in alignment data while taking UMIs into account, specifically designed for variant calling benchmarking analyses.

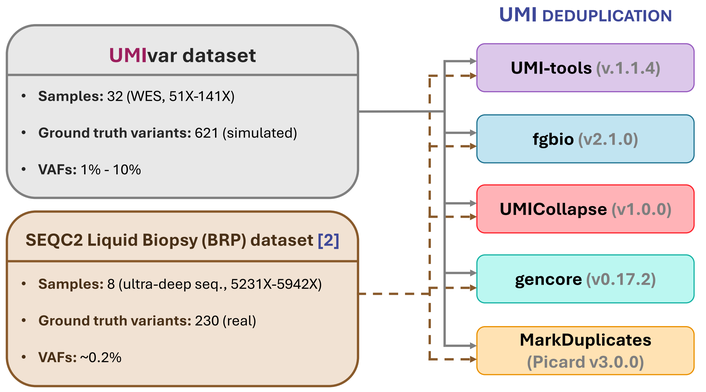
Methods

UMIvar was developed in Python. It takes as input UMI-tagged alignment data (**BAM** file) and the variants to be simulated (in a **CSV/VCF**).



The tool was initially evaluated by simulating 1283 variants (VAFs: 1-100%) in 5 BAM files and performing sensitive variant calling with VarDict.

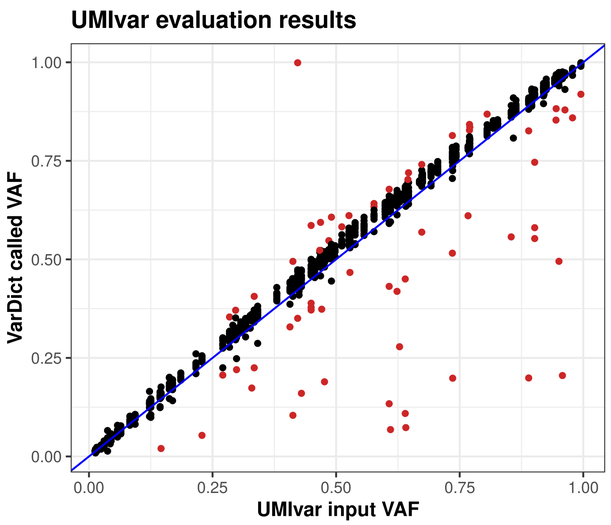
We then used **UMIvar** to benchmark the effect of five UMI deduplication tools in variant calling. A real dataset was also used.



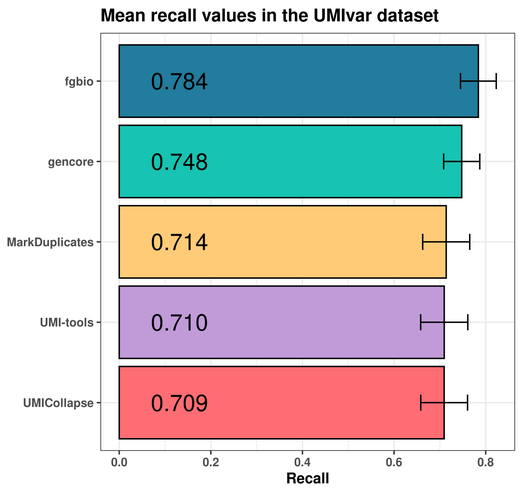
Variant calling was performed on each deduplication result using Lofreq, comparing:

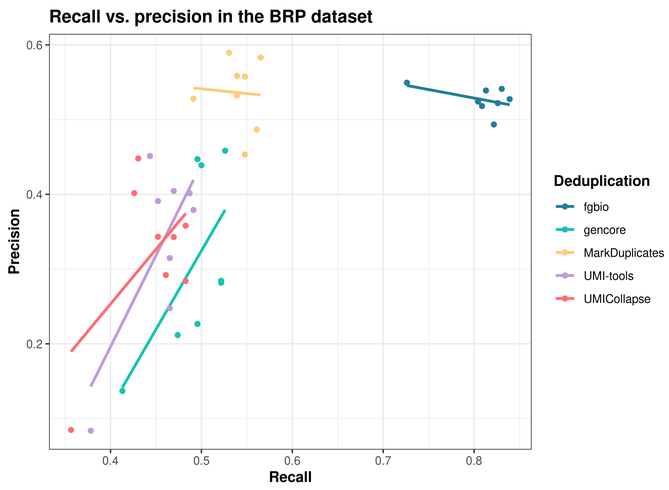
- **Recall:** TP / Total ground truth variants
- **Precision:** TP / TP + FP

Results



1138 ± 52.85	**36** ± 14.20
Variants called with VAF differences < 5%	Variants not detected





Results from the UMI deduplication software benchmarking showed that:

- **fgbio** led to a superior recall in both datasets, highlighting the efficacy of consensus-based deduplication algorithms.
- **fgbio** led to the second highest precision (error correction), only surpassed by **Picard MarkDuplicates**.
- The insights obtained from the **UMIvar** synthetic dataset were consistent with the BRP real dataset findings.

Conclusions

- **UMIvar** proved useful as a tool for UMI-related software benchmarking to select optimal variant calling approaches.
- Based on our results, **fgbio** deduplication would be advisable for variant calling (particularly at low frequencies).

References

[1] Kinde I, Wu J, Papadopoulos N, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 2011; 108: 9530–9535.

[2] Gong B, Deveson IW, Mercer T, et al. Ultra-deep sequencing data from a liquid biopsy proficiency study demonstrating analytic validity. Sci Data 2022; 9: 170.

UMIvar is available at: GitHub **dgcamblor/UMIvar**

Contact: dgonzalez@incliva.es