

Benchmarking short-read germline structural variant calling highlights advantages of using ensembles of tools and small impact of graph genome alignment

Yuliu Guo^{1,2}, Emilie Sofie Engdal¹, Miyako Kodama¹,
Drew Kaley Ann Thompson¹, Jiayi Yao¹, Alban Laus Obel Slabowska¹,
Mònica Aguilà-Sans¹, Eliana Buenaventura¹, Anna Reimer Hansen¹,
Andreas Ørslev Rasmussen¹, Frederik Otzen Bagger^{1*}

¹MDxCORE, Rigshospitalet, Copenhagen University Hospital, Copenhagen,
Denmark.

²Section for Computational and RNA Biology, Department of Biology, University
of Copenhagen, Denmark.

*Corresponding author(s). E-mail(s): frederik.otzen.bagger@regionh.dk;
Contributing authors: yuliu.guo@regionh.dk; emilie.sofie.engdal@regionh.dk;
miyakokodama@gmail.com; drew.kaley.ann.thompson@regionh.dk;
jiayi.yao.01@regionh.dk; alban.laus.obel.slabowska@regionh.dk;
monica.aguila.i.sans@regionh.dk; ingrid.eliana.buenaventura.ruiz@regionh.dk;
anna.reimer.hansen@regionh.dk; andreas.oerslev.rasmussen@regionh.dk;

Supplementary Information

- Supplementary Figure [S1](#): Performance metrics for coding regions and regulatory regions for each SV caller.
- Supplementary Figure [S2](#): Distribution of F1 scores across genomic positions for each SV caller.
- Supplementary Figure [S3](#): Overlap of SV calls among callers, illustrated using a Venn diagram.

- Supplementary Figure S4: Impact of frequency filtering on Manta variant calls in NA12878.
- Supplementary Figure S5: Simplified rule graph (DAG) of the Snakemake pipeline.
- Supplementary Figure S6: Distribution of quality scores for true and false positive SV calls.

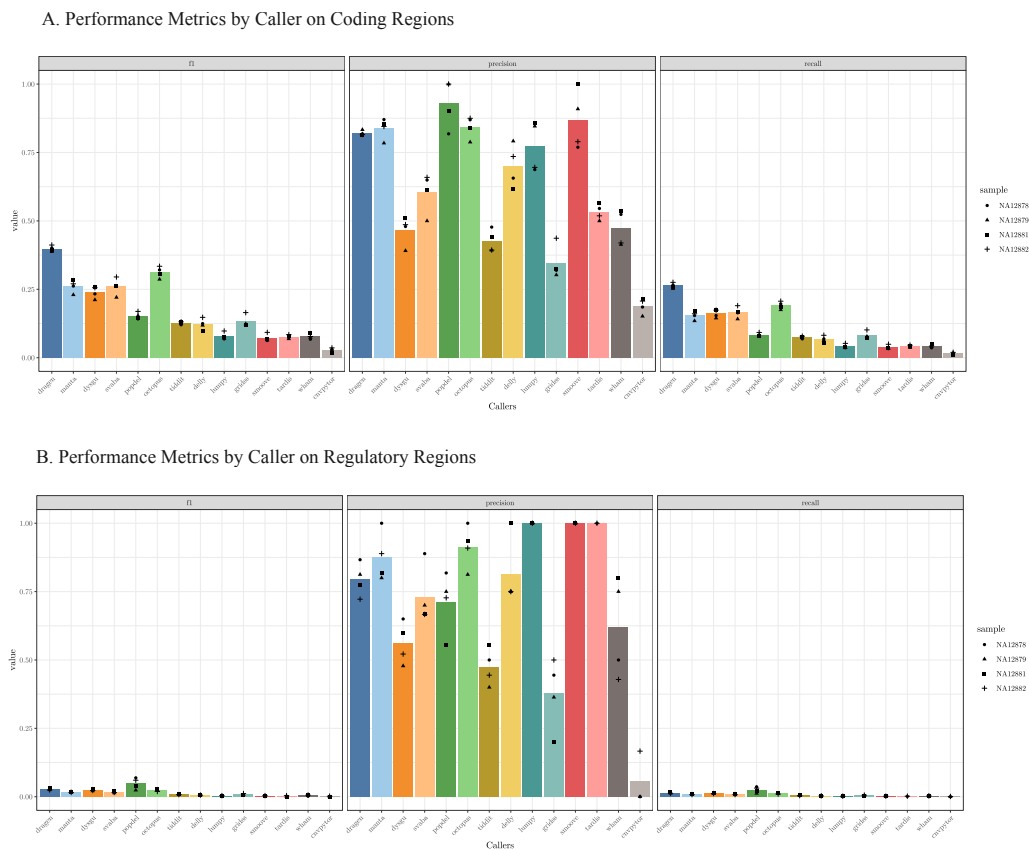


Fig. S1: Performance metrics on A) coding regions and B) regulatory regions for each SV caller. Bars represent the mean score for each caller, while the black line range represents the standard error across the four selected samples



Fig. S2: Distribution of F1 scores across genomic positions for each SV caller

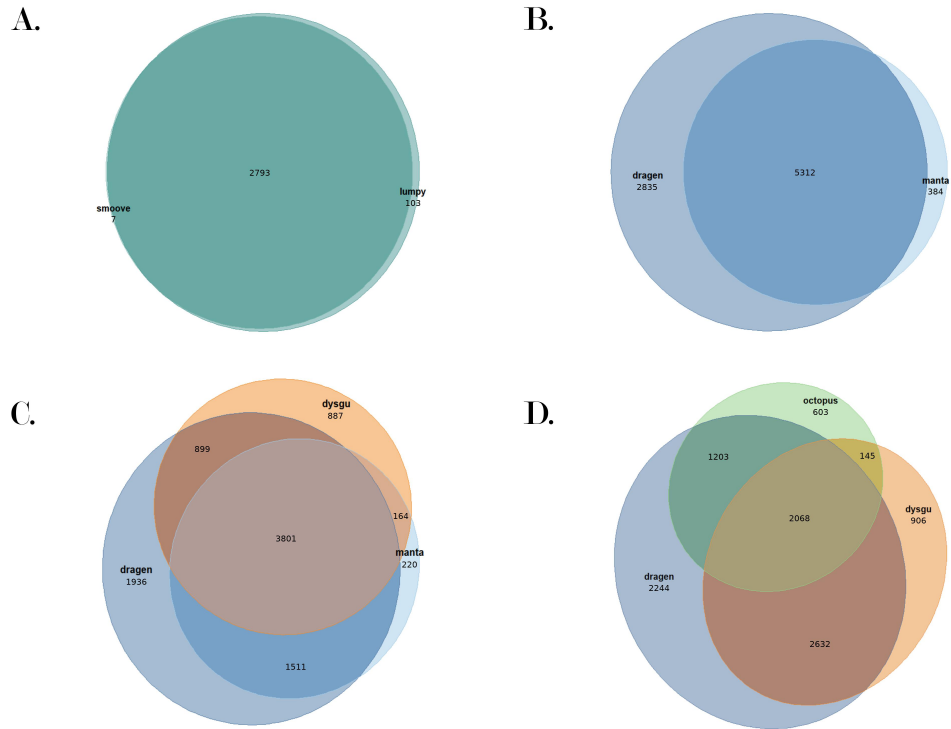


Fig. S3: Venn diagram showing the overlap among: A) Smoove and LUMPY B) DRAGEN and Manta C) The top three SV callers with the highest F1 scores: DRAGEN, Manta, and Dysgu D) top three SV callers with the largest number of unique calls: DRAGEN, Dysgu, and Octopus.

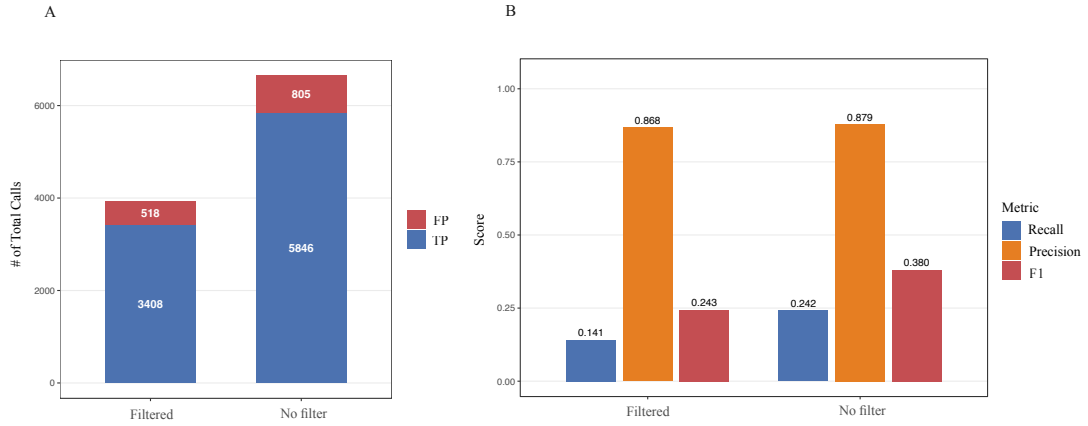


Fig. S4: A. Impact of frequency filtering on Manta variant calls. Number of TP and FP calls for NA12878 from Manta output with and without frequency filtering. B. Performance metrics of Manta output on sample NA12878 with and without frequency filtering.

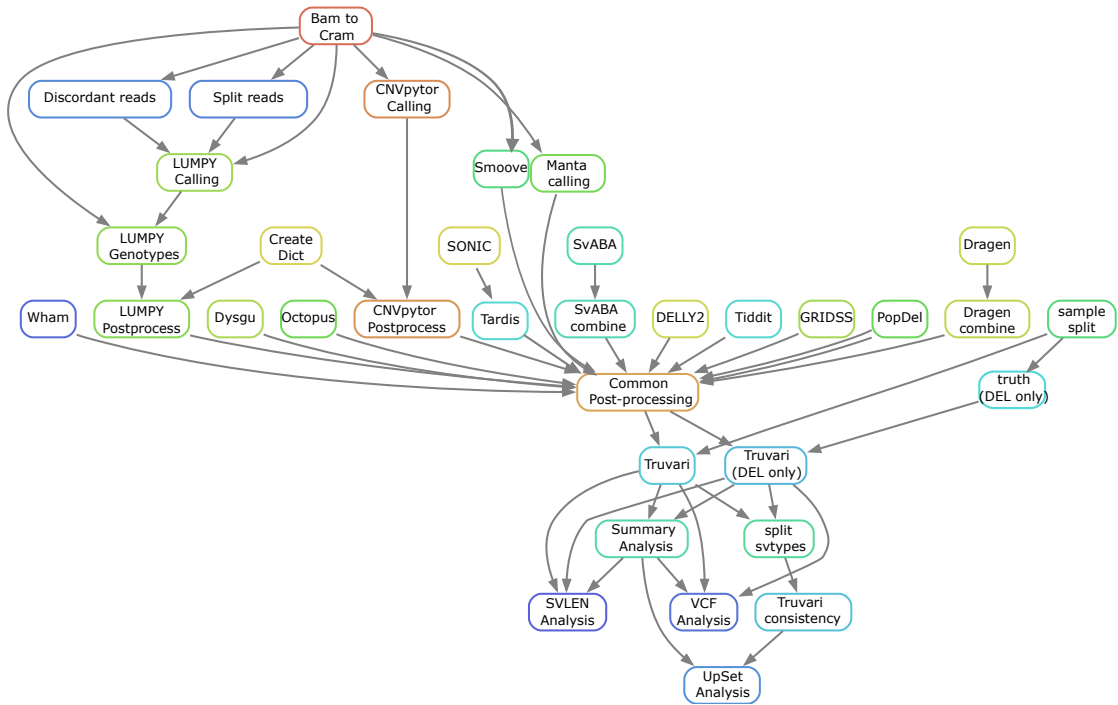


Fig. S5: Simplified rule graph (DAG) of the Snakemake pipeline, with common postprocessing steps collapsed for clarity.

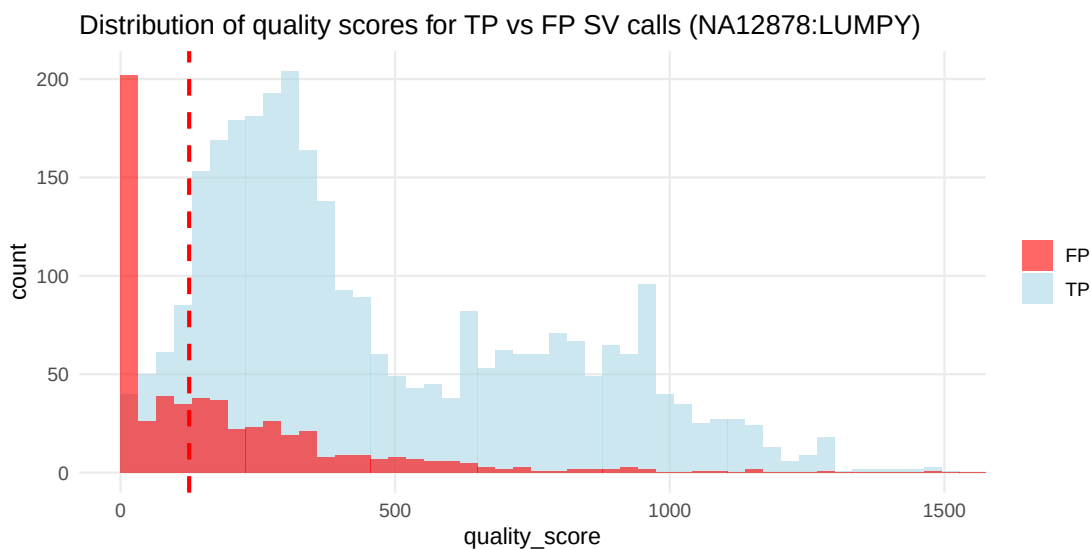


Fig. S6: Distribution of quality scores for true and false positive SV calls. Histogram showing the distribution of quality scores for true positive (TP, blue) and false positive (FP, red) structural variant calls derived from Truvari comparison against the pedigree truth set. The dashed line marked the quality score threshold of 125. The plot illustrates the separation between TP and FP calls and was used to empirically determine calling tool-specific quality thresholds for PASS variant classification.