Challenges and Pitfalls in the Evaluation of Structural Variants Callers

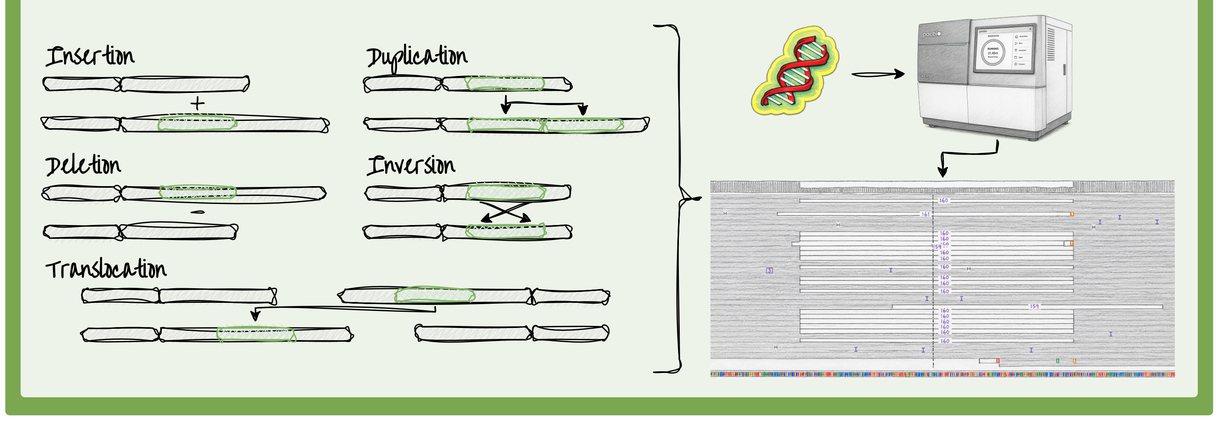
Luca Denti [1], *Thomas Krannich* [2], *Tomáš Vinař* [1], *Rayan Chikhi* [3], *Paola Bonizzoni* [4], *Broňa Brejová* [1], *Fereydoun Hormozdiari* [5]

[1] Comenius University in Bratislava, Slovakia; [2] German Cancer Research Center, Germany; [3] Institut Pasteur, France
[4] University of Milano-Bicocca, Italy; [5] University of California-Davis, USA

Motivation

Structural variants (SVs) are medium- and large-scale genomic alterations that shape phenotypic diversity and disease risk. Numerous methods have been proposed for the discovery of SVs, but their benchmarking has been inconsistent across studies, often resulting in *contradictory findings*. Primary sources of conflicting results lie in the inconsistent selection of reference genome version and ground truths, e.g., callsets curated by consortia and ad hoc callsets created from high-quality assemblies. Best practices and optimal evaluation strategies remain undefined and benchmarking results often vary, leaving the current performance assessment landscape *difficult to interpret*. More importantly, this has created *ambiguity* about the true capabilities of leading methods to accurately detect SVs, with far-reaching implications for routine genome analysis.

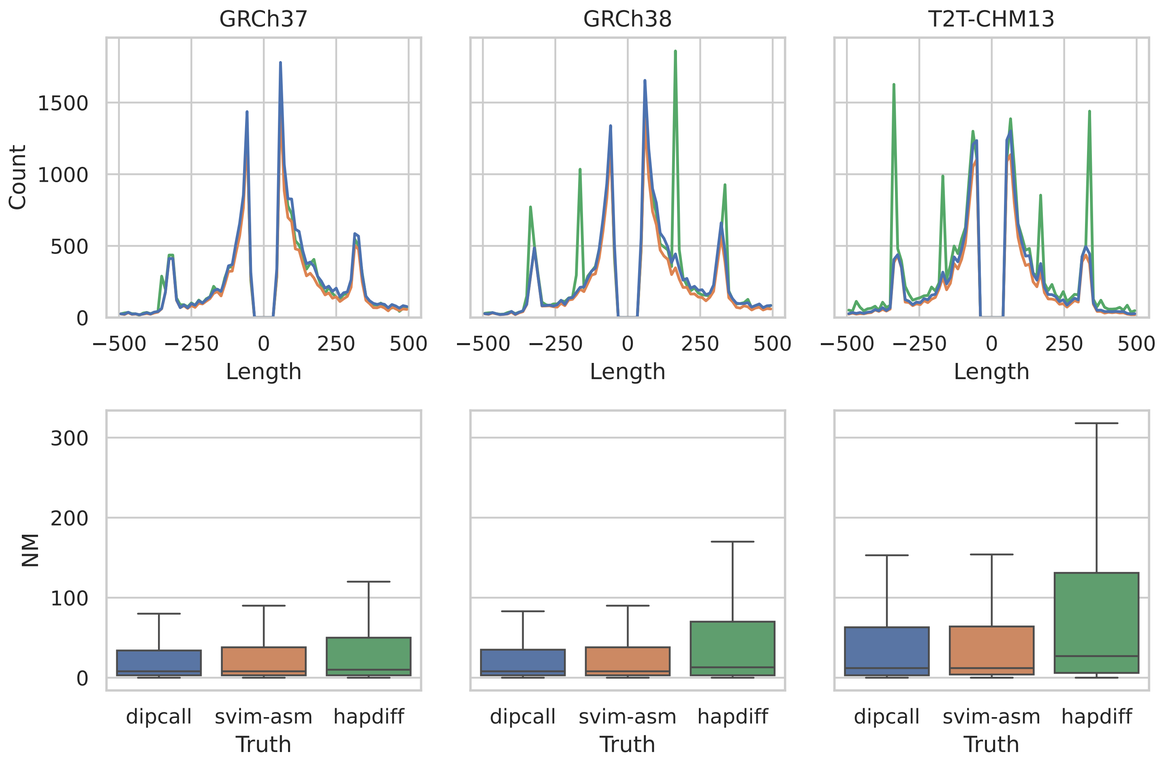
Background



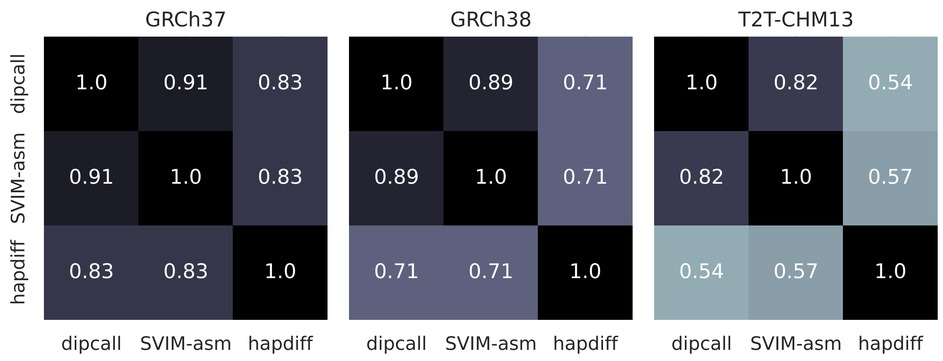
SV calling from high-quality *de novo* assemblies

***HG002 (GIAB Q100)**, HG002 (HPRC), NA12878 (Platinum Pedigree)*

(i) Leading methods produced different callsets, even when applied to the same assembly.



(ii) Similarity decreases as the completeness of the reference genome increases.



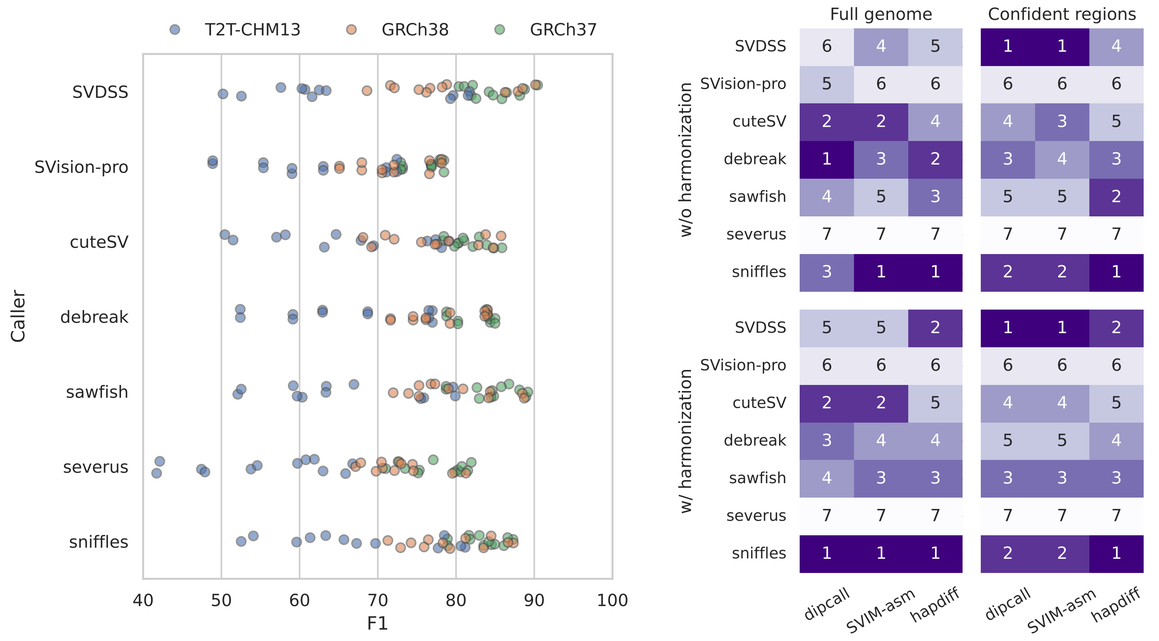
(iii) Differing computational strategies applied to address the biological complexity of certain genomic loci are one of the primary sources of these discrepancies:

- *alignment choices*
- *signal clustering (within and across haplotypes)*

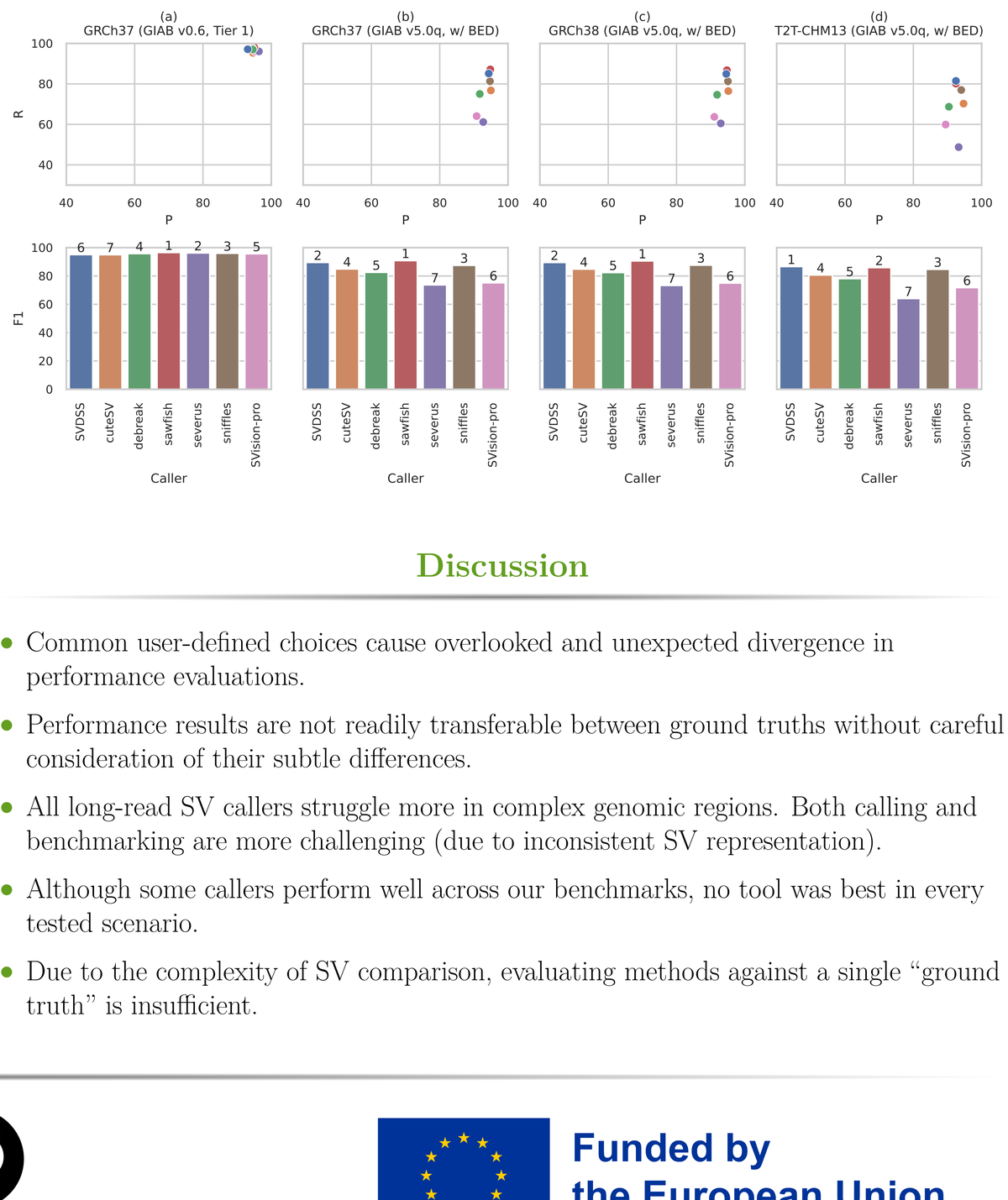
SV calling from PacBio HiFi reads

***HG002 (36x, HPRC)**, NA12878 (100x, Platinum Pedigree) - Sequel II*

(i) When evaluated against assembly-based callsets, F1-scores and rankings of callers vary depending on the choice of ground truth, reference genome, and benchmark parameters.



(ii) Similar results can be observed when considering curated SV callsets. Remarkably, the inclusion of more (challenging) SVs located in complex genomic regions substantially affects SV callers performances.



Discussion

- Common user-defined choices cause overlooked and unexpected divergence in performance evaluations.
- Performance results are not readily transferable between ground truths without careful consideration of their subtle differences.
- All long-read SV callers struggle more in complex genomic regions. Both calling and benchmarking are more challenging (due to inconsistent SV representation).
- Although some callers perform well across our benchmarks, no tool was best in every tested scenario.
- Due to the complexity of SV comparison, evaluating methods against a single "ground truth" is insufficient.

UNIVERZITA KOMENSKÉHO V BRATISLAVE

github.com/ldenti/svbench-fw

Funded by the European Union

This work is supported by the EU's Horizon programme under grant agreement No. 101180581 (ASVA-CGR)