VARium: a large-scale synthetic genome collection for multi-platform SV benchmarking

Yueyao Gao[1,2], Mark Fleharty[1,2], Katie Larkin[1,2], Sean Hofherr[1,2], Niall Lennon[1,2], Victoria Popic[1,2]

[1] Broad Clinical Labs [2] Broad Institute of MIT and Harvard Email: vpopic@broadinstitute.org

Abstract

Structural variants (SVs) are an important class of genomic variation with significant functional and clinical impact, yet their discovery remains an open problem. The performance of SV callers can vary as a function of variant type, size, genomic context, coverage, and sequencing platform. However, SV truthsets are scarce and typically contain a restricted set of SV types with skewed size and context distributions, hindering our ability to uncover systematic performance bias. Here, we introduce VARium: an extensive suite of synthetic genomes designed to assess the performance of SV discovery tools as a function of key domain parameters and confounders. We evaluated 22 SV callers, including alignment-based and assembly-based tools, on each genome from VARium using multiple sequencing platforms and varying read depth. As expected, we found that both recall and precision of most methods varied significantly across different benchmark settings, demonstrating that synthetic benchmarks can serve as a key tool for revealing systematic biases and limitations of SV methods, while providing upper-bound performance estimates for real datasets.

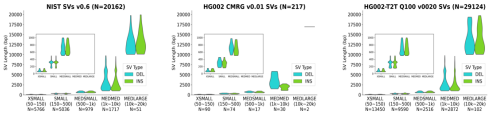
Motivation

Real SV benchmarks are scarce and include only a limited set of SV types – primarily deletions (DEL) and insertions (INS) – with skewed size distributions and uneven genomic context coverage.

	Number of variants	Number of variants in repeats	Number of variants in SEGDUPs
v0.6	20,162	14,090	195
CMRG	217	101	3
Q100	29,124	16,071	2,360

Number of variants stratified by genomic context in GIAB HG002 SV truthsets

Size distribution of SVs by type in GIAB HG002 SV truthsets



VARium v0 genome collection

Tier 1 (79 genomes)

Designed to stress-test simple SV recall as a function of type, size, and context

- 4 SV types: DEL, INS, DUP, INV
- 5 size ranges: 50-150bp, 150-500bp, 500-1kbp, 1k-10kbp, 10k-20kbp, 20k-100kb, 100k-1Mb
- 6 genomic contexts: UNIQUE, SEGDUP, NONUNIQUE, Alu*, L1HS*, TR* (*DEL only)

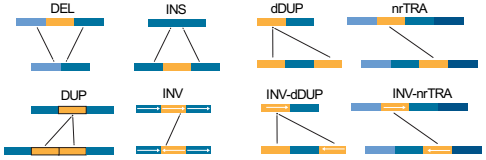
- 3 platforms: Illumina, PacBio, ONT
- 5 coverages: 30x, 15x, 10x, 5x, 0.5x

Tier 2 (40 genomes)

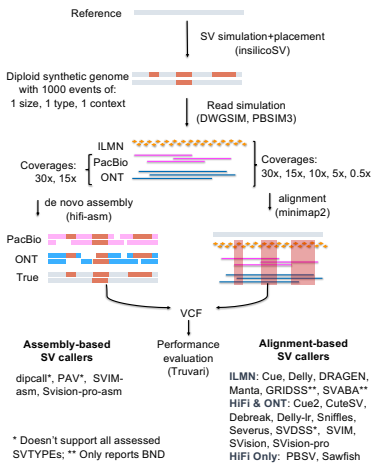
Designed to stress-test simple SV precision (in the presence of complex variants)

- 4 SV types: dDUP, nrTRA, INV-dDUP, INV-nrTRA
- 5 size ranges: 50-150bp, 150-500bp, 500-1kbp, 1k-10kbp, 10k-20kbp
- 2 dispersion distance regimes: 10k-50kbp, 1M-100Mbp

- 3 platforms: Illumina, PacBio, ONT
- 1 coverage: 30x



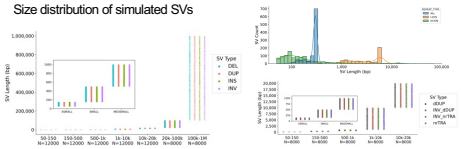
Data generation workflow



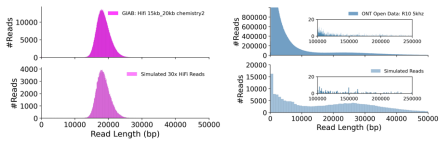
* Doesn't support all assessed SVTYPEs; ** Only reports BND

Synthetic WGS dataset characteristics

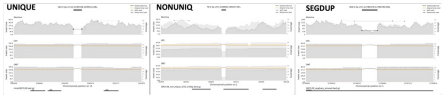
Size distribution of simulated SVs



Read lengths distribution for PacBio and ONT platforms

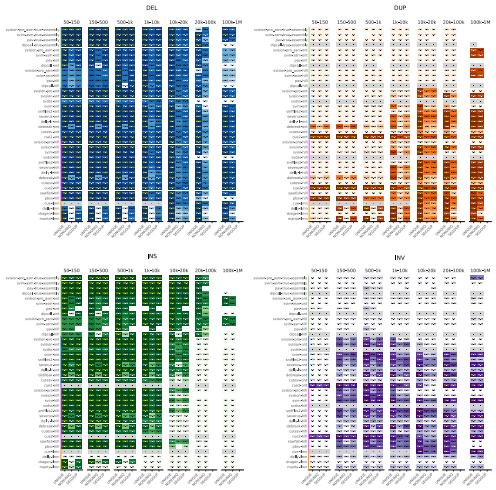


Example of a 150-500bp HOM DEL with placement in different genomic context

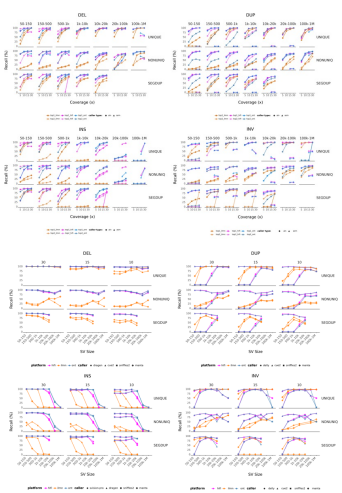


Tier 1 results

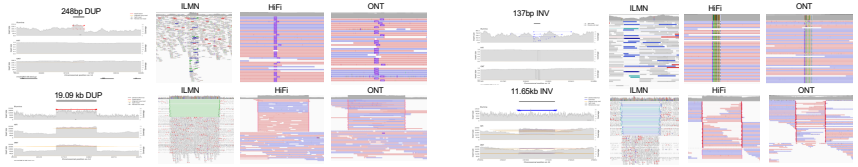
F1 scores stratified by SV type, size, context, and platform at 30x coverage



Recall stratified by SV type, size, context, platform, and coverage

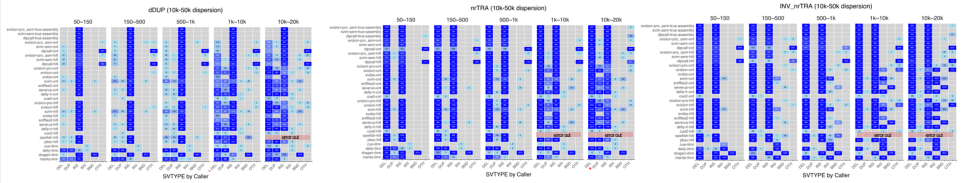


Examples of DUP and INV detection challenges across different event sizes



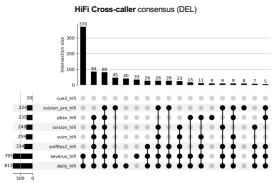
Tier 2 results

SV predictions at complex variant sites across three sequencing platforms

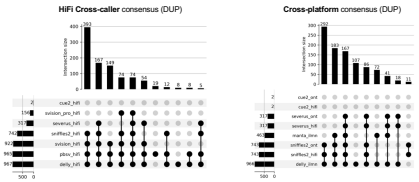


Cross-platform consensus of FP simple SV calls

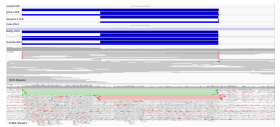
dDUP 1k-10k size, 10k-50k dispersion



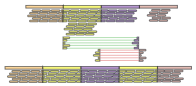
nrTRA 10k-20k size, 10k-50k dispersion



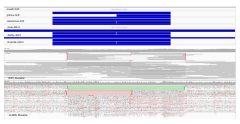
Simple SV predictions at a dDUP site



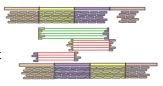
dDUP alignment signature



Simple SV predictions at an nrTRA site



nrTRA alignment signature



Conclusion

- 22 state-of-the-art SV callers show wide performance variability across conditions.
- Few callers achieve high recall across all type and size combinations; most perform well only in some regimes.
- Small DUPs and INVs are particularly challenging for most long-read callers due to alignment artifacts.
- Most callers incorrectly report simple events at different combinations of complex SV breakpoints. Notably, these FP calls are often shared across callers and platforms.

https://github.com/PopicLab/varium

Thank you to our funding sources:

NIH NHGRI BROAD INSTITUTE