Download

Oxford NANOPORE Technologies
APPLICATIONS

Comparison of structural variant calls from Oxford Nanopore haplotype-resolved and telomere-to-telomere genome assemblies

Sean McKenzie[1], Sergey Nurk[2], Alex Dawson[2], Sissel Juul[1] & Philipp Rescheneder[2]

[1]Oxford Nanopore Technologies Inc, New York, NY, [2]Oxford Nanopore Technologies plc, Oxford, UK

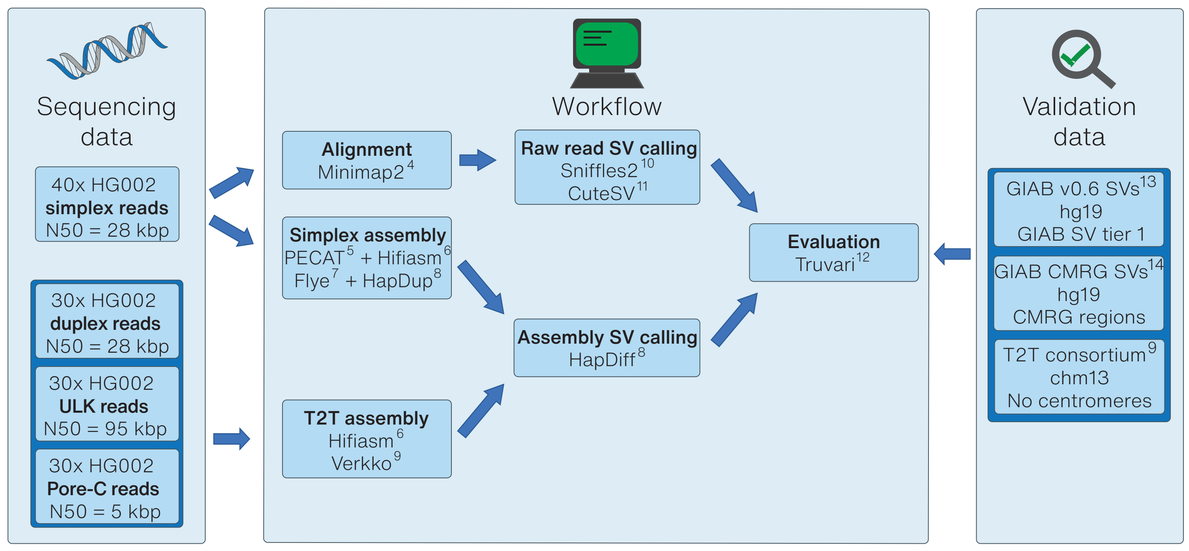
Contact: sean.mckenzie@nanoporetech.com

Abstract

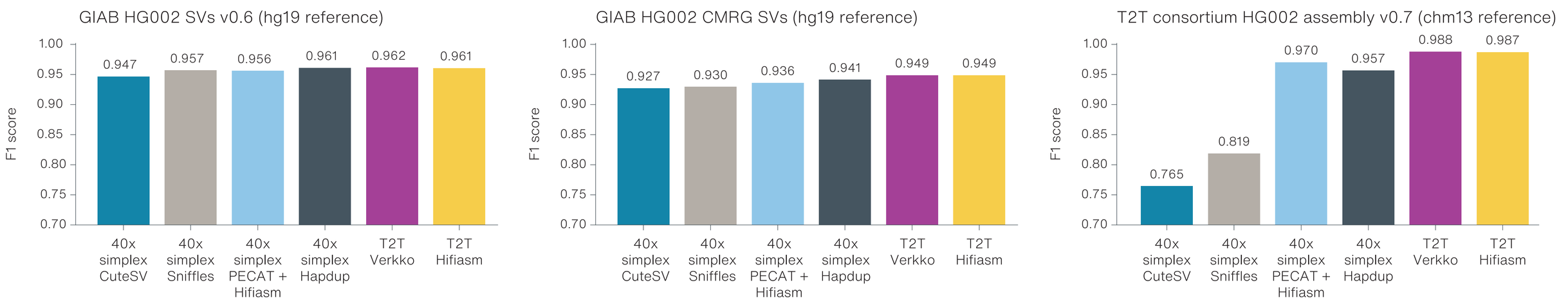
Structural variants (SVs), including insertions and deletions larger than 50 bp, duplications, inversions, and translocations, represent a major source of genetic variation in human populations[1,2]. These variants have historically been difficult to detect with array and short-read-based approaches, however, hindering efforts to understand their importance in human health and disease[1,2]. Modern long-read sequencing technologies have drastically improved our ability to identify and characterize SVs, recently culminating in complete "telomere-to-telomere" (T2T) *de novo* genome assemblies which can represent the full range of human sequence variation [3]. Generating T2T genomes remains challenging and costly and requires a combination of multiple sequencing data types. It is therefore important to compare T2T assembly variant calling to other available methods – including raw-read alignment and other haplotype-resolved assembly-based methods – to determine if and when T2T assembly is necessary.

Here we compare SV calls from raw-read, haplotype-resolved *de novo* assembly, and T2T *de novo* assembly approaches using Oxford Nanopore simplex and duplex whole-genome sequencing data in order to determine the most efficient SV discovery strategy.

1. Data and analysis

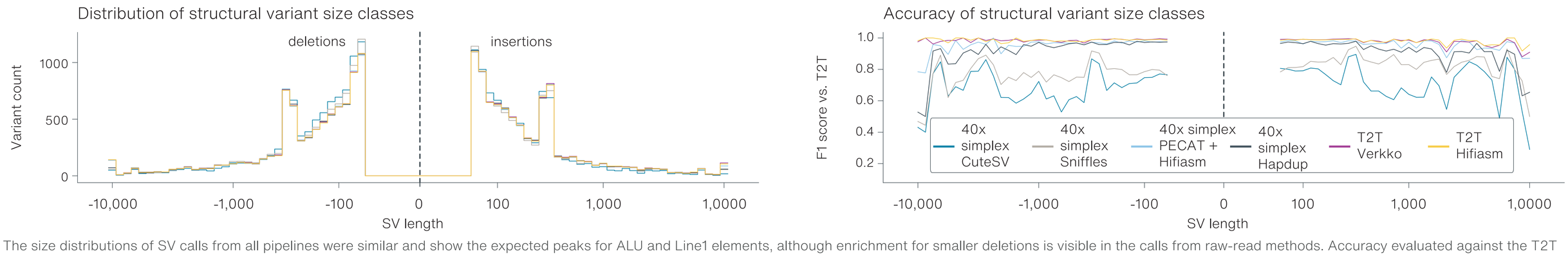


2. Overall accuracy of different structural-variant-calling methods



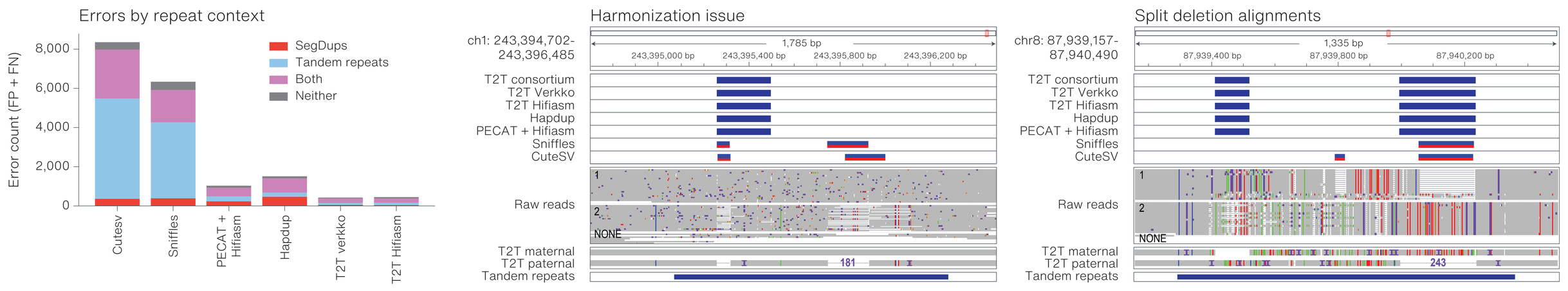
Comparison of variant calls with the Genome-in-a-Bottle (GIAB) v0.6 SV benchmark shows comparable performance among all SV-calling pipelines, with slightly higher accuracy observed in diploid *de novo* assembly approaches relative to Sniffles, which in turn showed better accuracy than CuteSV. A clearer hierarchy is visible when comparing to the GIAB challenging medically relevant genes (CMRG) SV benchmark, with T2T assemblies out-performing other diploid assemblies, all of which outperform raw-read approaches. Comparison against T2T consortium assembly v0.7 SV calls reveals an even more drastic difference between assembly approaches and raw-read approaches, while the gap between T2T assemblies and other diploid assemblies is comparatively modest in this evaluation.

3. Breakdown by length and type



The size distributions of SV calls from all pipelines were similar and show the expected peaks for ALU and Line1 elements, although enrichment for smaller deletions is visible in the calls from raw-read methods. Accuracy evaluated against the T2T consortium HG002 assembly SV calls was consistently high for assembly-based methods, while considerable variability in accuracy for different size classes was seen for raw read methods, with a drop off in accuracy at the largest size class.

4. Error context and challenges with harmonizing representations



The vast majority of errors (false positives (FP) and false negatives (FN)) from the raw read SV callers fall into tandem repeats, segmental duplications (SegDups), or both. Manual inspection of error-containing regions showed that this was mostly due to two phenomena: harmonization issues, where identical calls were represented too differentally for Truvari to match them; and split deletion alignments, where Minimap2 split deletions in the raw reads into multiple small deletions below Sniffles' and CuteSV's size thresholds.

References

1. Feuk, L., Carson, A.R., Scherer, S.W. (2006) Nature 7, 85-97
2. Ebert, P. et al. (2021) Science 372, eabf7117
3. Nurk, S. et al. (2022) Science 376, 44-53
4. Li H. (2018) Bioinformatics 34, 3094-3100
5. Nie, F. et al. (2023) bioRxiv doi:10.1101/2022.09.25.509436
6. Cheng, H. et al. (2021) Nature Methods 18, 170-175
7. Komogorov, M. et al. (2019) Nature Biotechnology 37, 540-546
8. Kolmogorov, M. et al. (2023) Nature Methods 20, 1483-92
9. Rautiainen, M. et al. (2023) Nature Biotechnology 41, 1474-82
10. Smolka, M. et al. (2023) bioRxiv doi:10.1101/2022.04.04.487055
11. Jiang, T. et al. (2020) Genome Biology 21, 189
12. English, A.C. et al. (2022) Genome Biology 23, 271
13. Zook, J.M. et al. (2020) Nature Biotechnology 38, 1347-55
14. Wagner, J. et al. (2022) Nature Biotechnology 40, 672-680

Conclusions

These findings show that all approaches for SV calling show very high accuracy in well-characterized regions of the genome, as variant calling within Genome in a Bottle high-confidence regions was comparable across the different approaches. De novo assembly approaches did show notably higher concordance with Telomere-to-Telomere Consortium validation variants in challenging genomic contexts outside the GIAB high-confidence regions, mostly due to issues with variant representation and discovery in tandem repeats. The gap could be closed by focused development of raw-read methods, to deliver improved handling of tandem repeats. T2T de novo assembly outperformed other de novo assembly methods, though the gains were marginal for the majority of SVs in most genomic contexts.