Improving structural variant calling accuracy using SBX duplex and longer simplex reads

Mahdi Golkaram[1], Jordan Eizenga[1], Mike Molnar[1], Melanie Kirsche[1], Ke Tang[1], Dilmi Perera[1], Samuel Higgins[1], Alberto Gatto[1], Huanjie Sheng[1], Biljana Simsic[1], Xixi Wang[1], George Mavromatis[1], Amal Chaturvedi[1], Daniel Zinder[1], Jagadeeswaran Chandrasekar[2], Grant Kingsley[2], McKenna Osentowski[2], Alec Sautter[2], Alex Lehmann[2], Alan Kimura[2], Emily Ormbrek[2], Kendall Berg[2], Taylor Lehmann[2], Joanne Leadbetter[2], Majid Babazadeh[2], Melud Nabavi[2], Matthew Lopez[2], Kevin Smith[2], Marc Prindle[2], Cynthia Cech[2], Megan Freer[2], Anasha Arryman[2], Ryan Toma[2], Mark Kokoris[2], John Mannion[1], Chen Zhao[1]

[1]Roche Sequencing Solutions, Inc. Santa Clara, CA, USA
[2]Roche Sequencing Solutions, Inc. Seattle, WA, USA

Copies of this poster obtained through QR (Quick Response), and/or text key codes are for personal use only and may not be reproduced without written permission of the authors.

P15.106.D

Introduction

Structural variants (SVs) are large-scale alterations in DNA, typically >50bp, often linked to diseases but also contributing to human diversity. Challenges persist in accurately calling SVs using short read sequencing data. Sequencing By Expansion (SBX) is an ultra-high-throughput, rapid sequencing technology which offers flexible modalities including SBX Duplex (SBX-D), ideal for detection of small variants (~Q40, ~240bp read length), and SBX Simplex Longer (SBX-SL), which can generate longer reads up to ~1500bp (with a distribution from ~200bp to ~1500 bp) under appropriate sample and library prep conditions, at a higher error rate (~Q20). Here, we introduce a novel method to detect SVs using whole genome SBX-D or SBX-SL methods.

Methods

Characterizing germline structural variants (SVs) with short-read sequencing remains highly challenging due to severe mapping and assembly limitations within complex loci. Because a large proportion of SVs occur within repetitive elements and segmental duplications, short fragments (~100–150 bp) often fail to map uniquely, leading to extensive multi-mapping or alignment ambiguity. This constraint is magnified during large insertion events; when an insertion exceeds the read length, reads derived from the novel sequence lack a genomic anchor and remain entirely unmapped. Furthermore, in complex, variant-rich regions, the lack of long-range read connectivity makes local de novo assembly exceedingly difficult. Consequently, algorithms must rely on indirect proxy signals (e.g., split-reads and discordant mate-pairs) instead of direct breakpoint resolution. Because these signals are highly vulnerable to reference bias and mapping artifacts, both sensitivity and precision are severely degraded[1].
Here, we introduce a novel SV detection framework optimized for single-end SBX-D and SBX-SL whole-genome data by leveraging read depth and split-read alignments. To maximize variant evidence, the pipeline first implements pan-genome graph mapping. By retaining reads that map to SV loci within the graph but lack a linear reference projection, this approach increases split-read signal and reduces unmapped reads by more than 60%, substantially improving sensitivity for long insertions (Fig. 1).

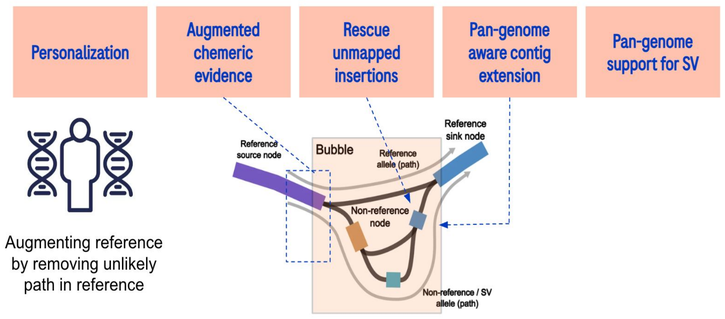

Figure 1. Leveraging pan-genome to improve SV evidence during mapping SBX reads to human genome reference. Image adopted from Groza et al[2].

Following alignment, a novel de novo local assembler integrates read-level evidence and pan-genome information to construct long contigs supporting alternative (ALT) haplotypes. These contigs are then mapped back to the linear reference using an optimized contig aligner (Fig. 2).

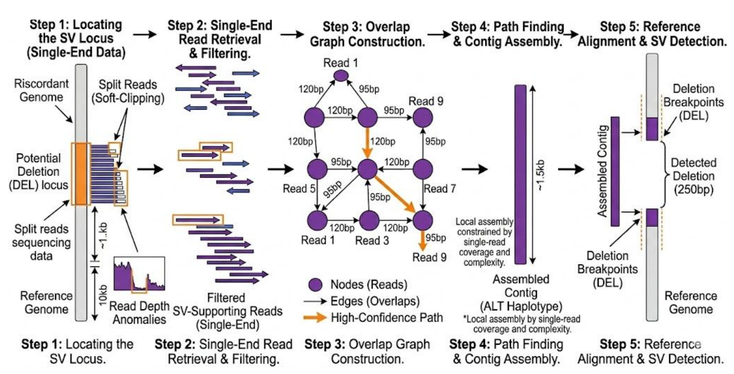

Figure 2. Workflow of overlap graph-based local *de novo* assembly of single-end SBX reads for germline SV detection.

Finally, to mitigate false positives arising from sequencing artifacts, features are extracted from the read and contig alignments to train a machine learning classifier (Fig. 3).

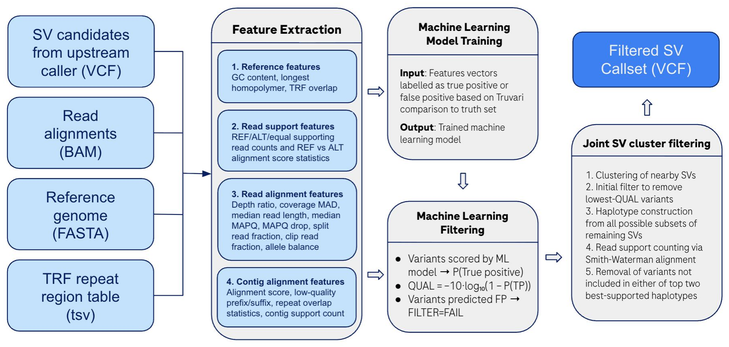

Figure 3. Machine learning based classification of SV candidates.

Results

To benchmark SV calling performance of the SBX platform, whole-genome sequencing was performed on the HG001 and HG002 reference cell lines using both the SBX-SL and SBX-D modalities. The ML model was trained on 30× coverage of HG001 and evaluated on 30× coverage of HG002. Specifically, target coverage was defined as deduplicated duplex depth for SBX-D and deduplicated total depth for SBX-SL. To prevent data leakage during the pangenome-guided read mapping phase, all Genome in a Bottle (GIAB) samples were strictly excluded from the pangenome reference graph (HPRC v2). Fig. 4 and 5 illustrate the SV detection performance of our framework evaluated against the HG002 NIST T2T Q100 benchmark (v1.1_v0.019, hg38) utilizing Truvari (v5.4).

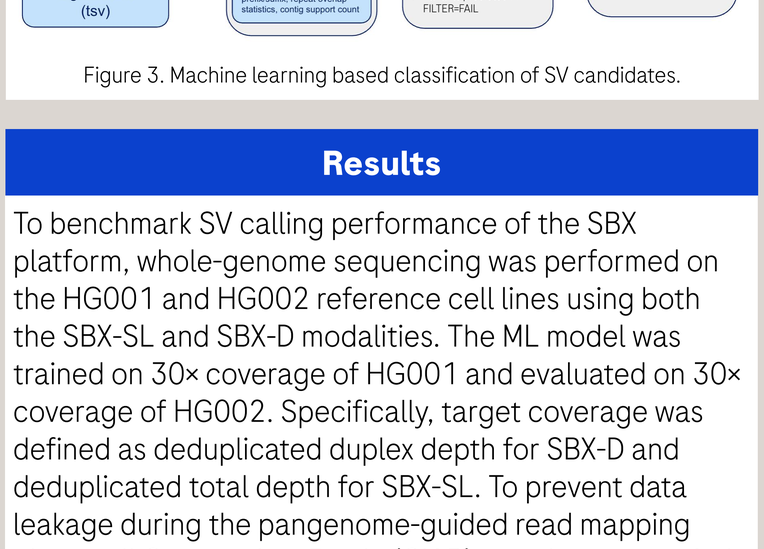

Figure 4. Performance evaluation of HG002 SV calling (INDELs) >50bp events. Benchmarked with Truvari v5.4 (-p 0 -P 0 for results without sequence similarity; --pick multi for all results) against T2T truth set. Non-HPRC benchmark was generated by removing all variants within 500bp of any SV in the HPRCv2.0 callset (after removing GIAB samples from the callset).

Overall, both the SBX-SL and SBX-D modalities demonstrate improved recall and F1-scores compared to current methodologies, with SBX-SL outperforming SBX-D due to its extended read lengths.
Notably, while this approach achieves exceptional accuracy for SVs represented within the HPRC pangenome graph, its performance in characterizing rare SVs remains highly competitive (Fig. 4b, d).

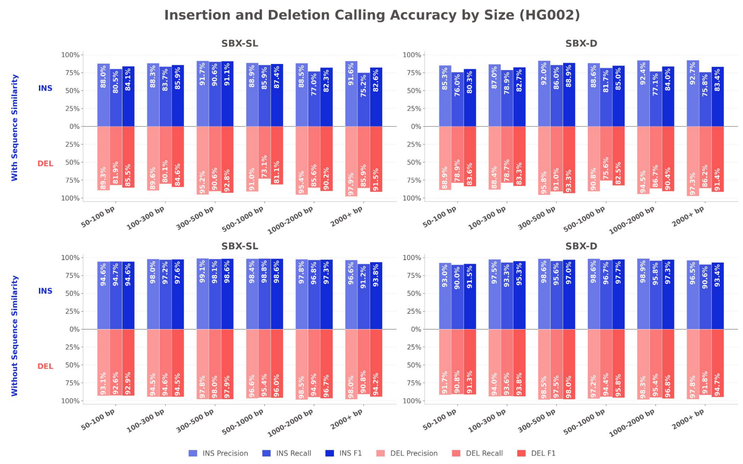

Figure 5. Breakdown of INS and DEL performance on HG002 by event size

Due to a lack of training data, inversion (INV) classification relied on manual heuristics (strand orientation and ALT support) and was evaluated against 90 HG002 benchmark variants[3]. Because HG002 lacks native translocations (TRAs), TRA detection was evaluated using 11 clinically relevant reciprocal rearrangements simulated by swapping chromosome tails in the GRCh38 reference. Ultimately, performance for both INVs and TRAs was significantly lower than for insertions and deletions, underscoring the persistent difficulty of resolving complex, copy-neutral rearrangements using short read sequencing data (Table1).

	True positives	False positives	Precision / Recall / F1
SBX-D INV	17 / 90	52	24.6% / 18.9% / 21.4%
SBX-SL INV	18 / 90	103	14.9% / 20.0% / 17.1%
SBX-D TRA	6 / 11	40	13.0% / 54.5% / 21.0%
SBX-SL TRA	10 / 11	51	16.4% / 90.9% / 27.8%

Table 1. Performance of SBX-D/SL in detection of germline TRA/INV events.

Conclusions

Here, we present a pan-genome-aware, machine-learning-based SV detection framework leveraging SBX-D and SBX-SL whole-genome sequencing data. This approach yields a significant increase in SV detection sensitivity, particularly for large insertions and variants within complex, variant-rich genomic regions. While this framework represents a significant advance, further research is required to improve the characterization of TRAs/INVs to fully overcome the remaining structural limitations inherent to short-read sequencing paradigms.

References

1. Cameron, Daniel L., et al. Nature Communications 10.1 (2019): 3240.
2. Groza, Cristian, et al. Nature Communications 15.1 (2024): 657.
3. Cheng S, Sedlazeck FJ. Benchmark for simple and complex genome inversions. bioRxiv (2025).

The sequencing by expansion (SBX) technology is in development and not commercially available. The content of this material reflects current study results and/or design goals. For Research Use Only. Not for use in diagnostic procedures.