Sawfish2: Integrating copy number segmentation with structural variant haplotype modeling to improve large-variant calling accuracy

Christopher T. Saunders, James M. Holt, Juniper A. Lake, Jonathan R. Belyeu, Zev Kronenberg, William J. Rowell, Michael A. Eberle

PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

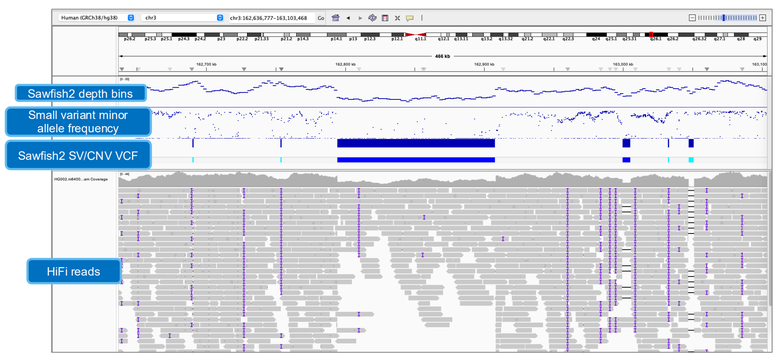
Summary

Sawfish is a general-purpose structural variant (SV) caller for HiFi sequencing data. It has already been shown to provide **best-in-class SV accuracy in both single-sample and joint-genotyping contexts**[1].

Sawfish2 adds depth-based CNV calling as a joint operation **integrated with sawfish's existing breakend-based SV calling** methods. This provides several benefits for HiFi WGS analysis:

- **A unified and consistent view of large-variants in each sample:** Sawfish2 provides a single VCF output of all SVs and CNVs, including merged records supported by both breakpoint and depth evidence.
- **More sensitive CNV calling:** Breakpoint evidence and structure is integrated into the depth segmentation process to boost sensitivity while retaining precision.
- **More precise SV calling:** Large unbalanced SVs are verified against the sample depth segmentation result and only reported as deletions or duplications when these signatures are consistent.

Integrated SV/CNV example



- 114kb deletion (CN1) from HG002
- In addition to integrated SV/CNV calls, **sawfish2 creates visualization tracks for depth and minor allele frequency** for each sample.
- **Sawfish2 merges all results to one VCF v4.4 output**, using the SVCLAIM standard. This includes CNVs solely supported by depth evidence, SVs supported by breakpoints, and joint SV/CNV calls supported by both.
- For joint SV/CNV calls, **all supporting read and depth information is unified on one record**:

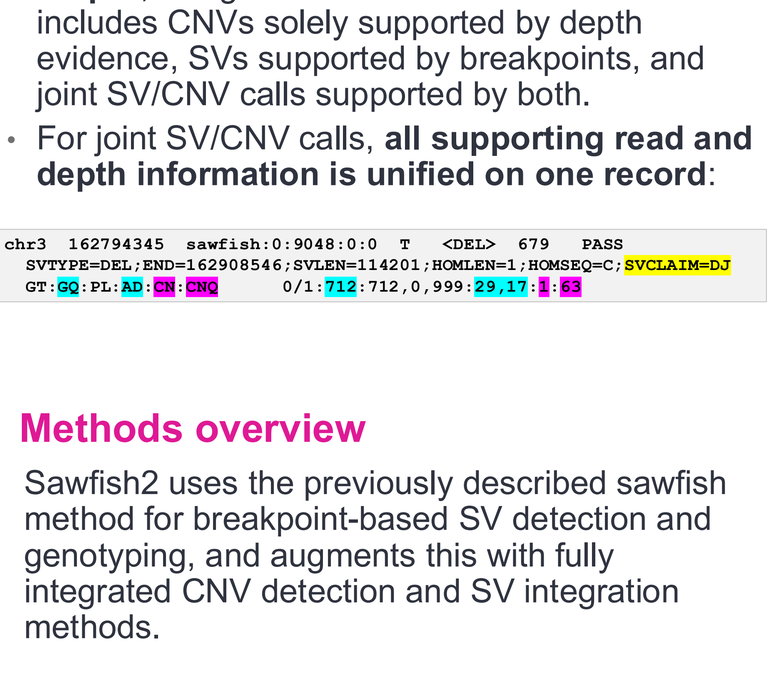
```
chr3  162794345  sawfish:0:9048:0:0  T  <DEL>  679  PASS
  SVTYPE=DEL;END=162908546;SVLEN=114201;HOMLEN=1;HOMSEQ=C;SVCLAIM=DJ
  GT:GQ:PL:AD:CN:CNQ      0/1:712:712,0,999:29,17:1:63
```

Methods overview

Sawfish2 uses the previously described sawfish method for breakpoint-based SV detection and genotyping, and augments this with fully integrated CNV detection and SV integration methods.

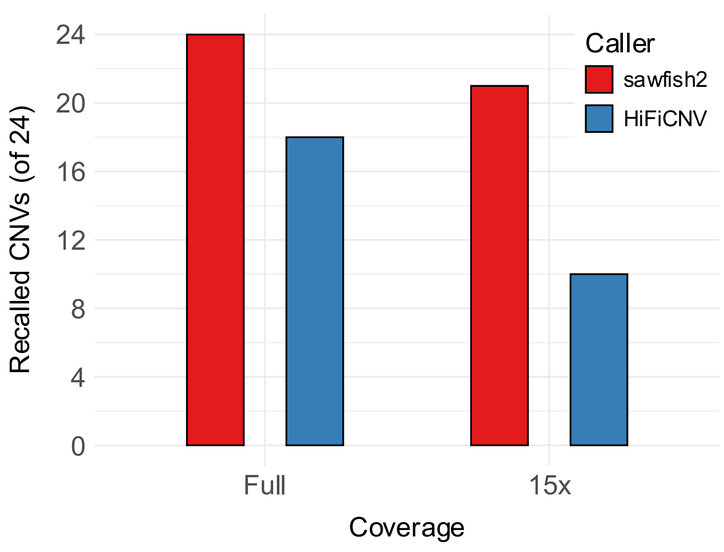
Discover Phase:
Per-sample 'draft' segmentation

Joint-call Phase:
SV/CNV integration



SV/CNV integration greatly improves recall of pathogenic CNVs

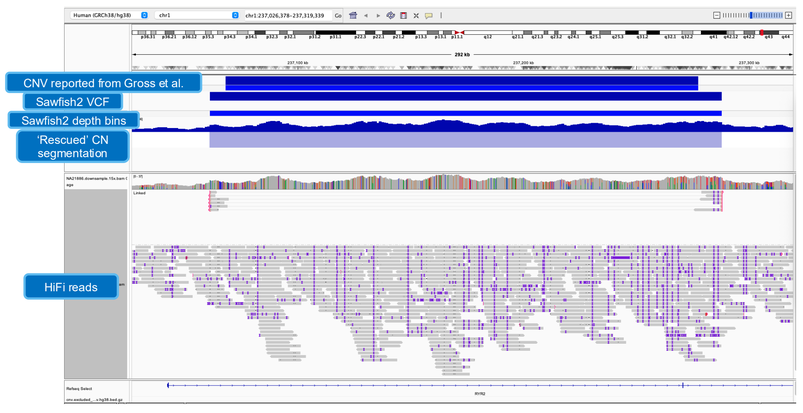
Sawfish2 uses SV breakend information and improved CNV quality scores to increase the sensitivity of CNV calling from HiFi WGS, especially at lower depth.



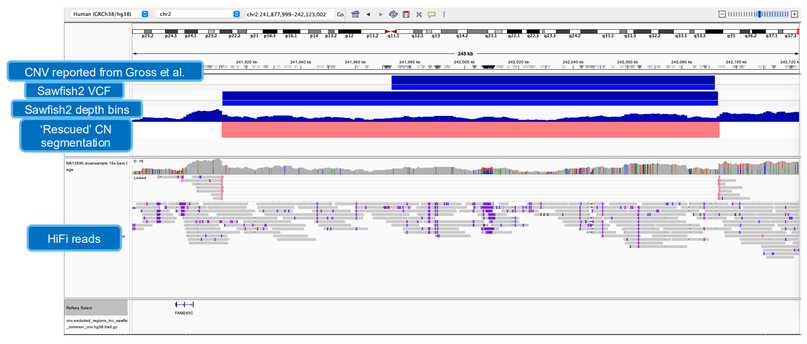
- Test set of known pathogenic CNVs from reference samples as described in Gross et al[2]. After all QC steps, this comprised 24 CNVs in 16 samples.
- All samples tested from HiFi WGS at ~30x, and at 15x (downsampled).
- CNV recall assessed by Truvari[3] for all passed variant calls from each method.

Sawfish2 rescues CNVs at 15x depth using SV breakpoint information

Rescued 227kb duplication (CN3) in NA21886



Rescued 183kb deletion (CN1) in NA13590



- Both CNV examples are undetected from depth-bin segmentation alone in either Sawfish2 or HiFiCNV
- Sawfish2 re-segments depth with copy number transition bonus for (select) breakpoints
- Both SV examples are rescued in sawfish2 re-segmentation step

Sawfish2 CNV quality values allow accurate retention of small CNVs

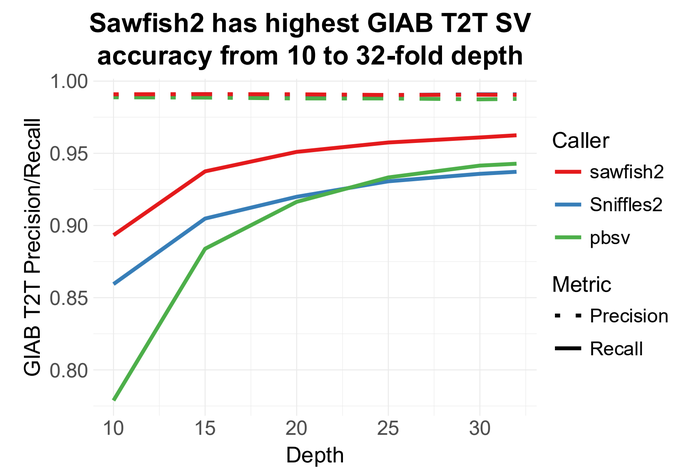
Quality for all sawfish2 autosomal CNVs in HG002

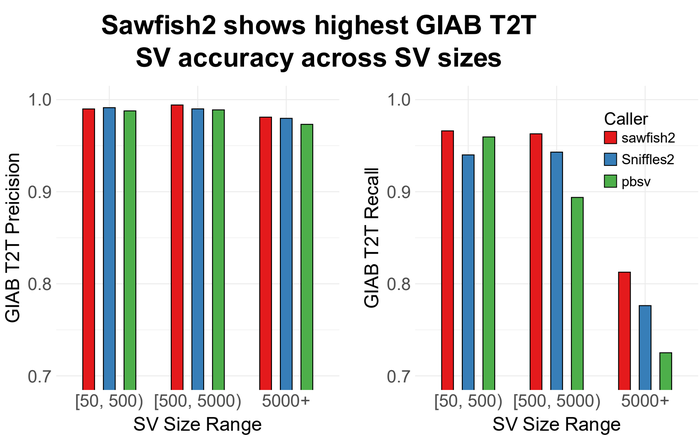
- New probabilistic copy number qualities allow dynamic filtration of CNVs by size and type as a function of sequencing coverage.
- Allows method to retain high precision without a hard size limit over all CNV types.

High SV recall and precision retained in integrated SV/CNV output

With integrated SV/CNV output, sawfish2 retains best-in-class SV accuracy compared to state-of-the-art long-read SV callers[4,5], across different sample read depths, size ranges, and for either comprehensive benchmarks or medically relevant genes.

Comprehensive (T2T) HG002 benchmark





- Accuracy assessed against the latest GIAB draft SV benchmark based on the T2T-HG002-Q100 diploid assembly[6], using Truvari[3], as described in the sawfish publication[1].

GIAB CMRG benchmark

Sawfish2 SV/CNV output retains the highest accuracy for SVs from the Challenging Medically Relevant Genes (CMRG) GIAB benchmark[7], as evaluated by Truvari bench:

Method	F1	Recall	#FN	Precision	#FP
sawfish2	0.993	0.991	2	0.995	1
Sniffles2	0.964	0.963	8	0.965	7
pbsv	0.971	0.958	9	0.985	3

Availability / Contact

latest release

app note

- Documentation, binary releases (Linux x64), and source code for sawfish and all new CNV-integration capabilities are available on GitHub: https://github.com/PacificBiosciences/sawfish
- Further methods and performance assessment details for sawfish SV calling available in our Bioinformatics App Note: https://doi.org/10.1093/bioinformatics/btaf136
- Sawfish is under active development. For questions or new use-case requests please reach out to us (csaunders@pacb.com).

References

1. Saunders, C. T., et al. (2025). Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. Bioinformatics, 41:4 (2025)
2. Gross, Andrew M. et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease, Genetics in Medicine, Volume 21, Issue 5, 1121 – 1130 (2019)
3. English, A.C., et al. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol 23, 271 (2022)
4. Smolka M., Paulin L.F., et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024 Jan 2. doi: 10.1038/s41587-023-02024-y.
5. pbsv: PacBio structural variant (SV) calling and analysis tools. URL: https://github.com/PacificBiosciences/pbsv
6. GIAB SV benchmark based off the T2T-HG002-Q100 assembly. URL: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.015-20240215/
7. Wagner J., et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol. 2022 May;40(5):672-680. doi: 10.1038/s41587-021-01158-1. Epub 2022 Feb 7. PMID: 35132260.