Comprehensive structural and copy-number variant detection with long reads

Aaron M. Wenger, Armin Töpfer, Yuan Li, Luke Hickey
Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025

Variant Detection with HiFi Reads

PacBio highly accurate, long reads (HiFi reads) comprehensively detect variants in the human genome, including in difficult repetitive regions.

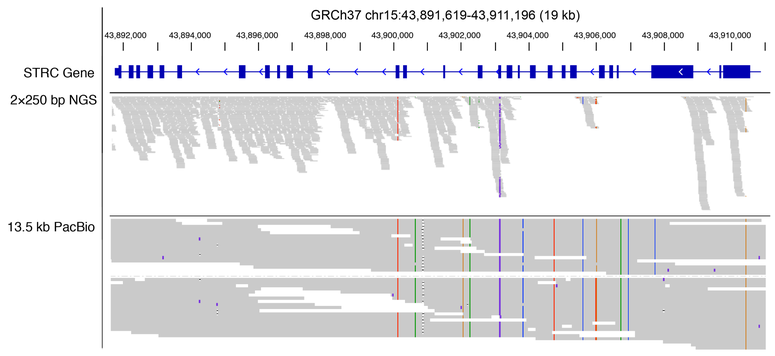
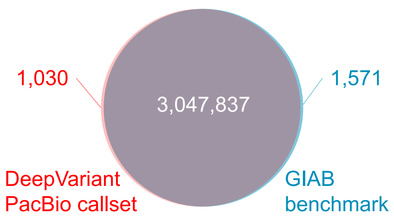


Figure 1. Accurate PacBio HiFi reads detect variants in difficult-to-map exons of the disease gene *STRC*[1].

Precision and recall, as measured against the Genome in a Bottle (GIAB) benchmarks[2,3], is high for single nucleotide variants (SNVs), indels, and structural variants (SVs).

SNVs

```
Ref GCAGGCAGCGACTACGTACGCTAACAGCGATCTCAG
Alt GCAGGCAGCGACTACGTCCTCTAACAGCGATCTCAG
```

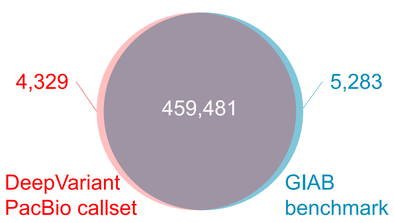


99.97% Precision
99.95% Recall

Figure 2. PacBio SNV calling performance for HG002 with 32-fold HiFi coverage (Rowell, poster 1866/W).

Indels

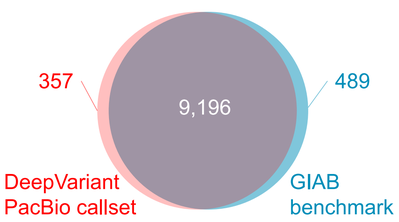
```
Ref GCAGGCAGCGACTACGTACGCTAACAGCGATCTCAG
Alt GCAGGCAGCGACTACGT-CGCTAACAGCGATCTCAG
```



99.10% Precision
98.86% Recall

Figure 3. PacBio indel calling performance for HG002 with 32-fold HiFi coverage.

SVs



96.26% Precision
94.93% Recall

Figure 4. PacBio SV calling performance for HG002 with 32-fold HiFi coverage.

Copy-number variant (CNV) calling with pbsv

Existing long read variant calling methods rely on *de novo* assembly or spanning reads to detect variants. These methods are effective for SVs but miss many CNVs that involve long segmental duplications.

> "We determined that 57% and 15% of the copy number variable bases within segmental duplications detected by dCGH and Genome STRiP, respectively, were not in contigs resolved by *de novo* assembly [of long reads]."
> – Chaisson et al. 2019 (ref. 4)

We extended the PacBio SV caller, pbsv[5], to detect CNVs using a combination of read clipping and depth.

pbsv CNV algorithm

Determine genome-wide coverage
median coverage of non-gap positions at mapping quality 60

↓

Identify candidate CNV breakpoints
positions with multiple clipped reads

↓

Evaluate coverage between adjacent candidate breakpoints
calculate z-score vs Poisson expectation

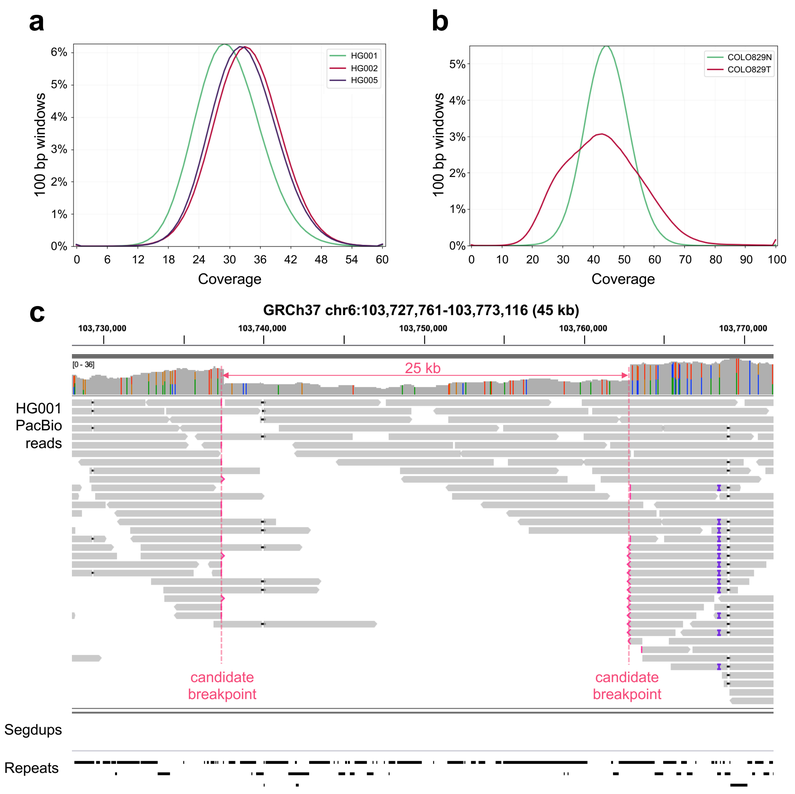


Figure 5. PacBio read coverage is Poisson distributed in autosomes for **(a)** samples from GIAB and **(b)** the normal sample from a tumor/normal pair (provided by WP Kloosterman). A tumor sample shows CNV regions of reduced and increased coverage. **(c)** To call CNVs, pbsv identifies candidate CNV breakpoints with multiple clipped reads and then evaluates read depth between adjacent breakpoints compared to the genome-wide typical coverage.

CNVs in HG001 and COLO829T

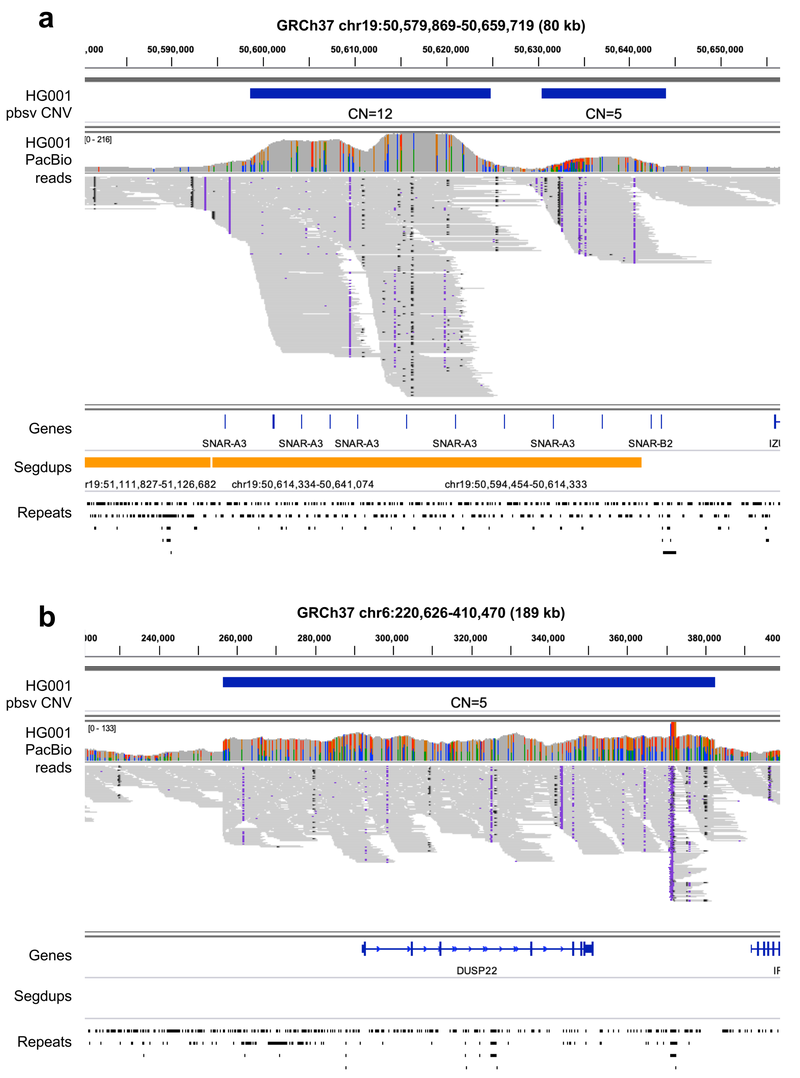


Figure 6. pbsv CNV calls in HG001 in **(a)** segmentally duplicated (segdup) sequence and **(b)** unique sequence.

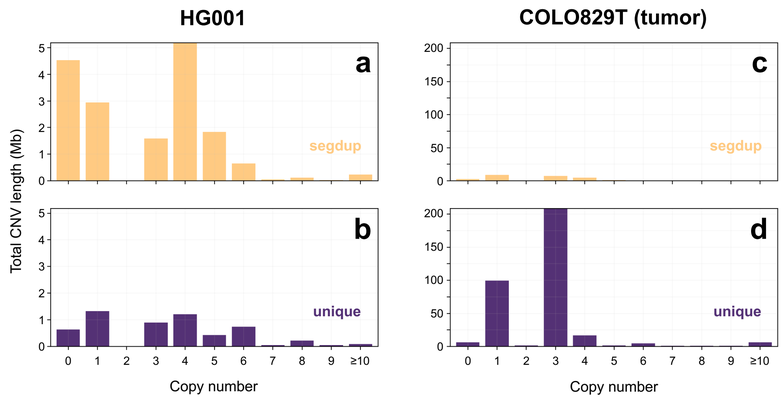


Figure 7. Genomic distribution of pbsv CNV calls. **(a, b)** Most CNVs in the "healthy" genome HG001 involve segdups. **(c, d)** The tumor genome COLO829T has many more CNVs than HG001 and most fall outside of segdups.

Summary

- PacBio HiFi reads comprehensively detect variants in a human genome.
- The pbsv variant caller identifies CNVs in HiFi reads using read clipping and depth signatures.

References

1. Wenger AM, Peluso P et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* doi:10.1038/s41587-019-0217-9.
2. Zook JM et al. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol.* 37(5):561-566.
3. Zook JM et al. (2019). A robust benchmark for germline structural variant detection. *bioRxiv.* doi:10.1101/664623. [Preprint]
4. Chaisson MJP et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun.* 10(1):1784.
5. https://github.com/PacificBiosciences/pbsv

Thank you to David Scherer, Kristin Robertshaw, and Pamela Bentley Mills for poster production support.