Copy-Number Variant Detection with PacBio Long Reads

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

Long Highly Accurate HiFi Reads

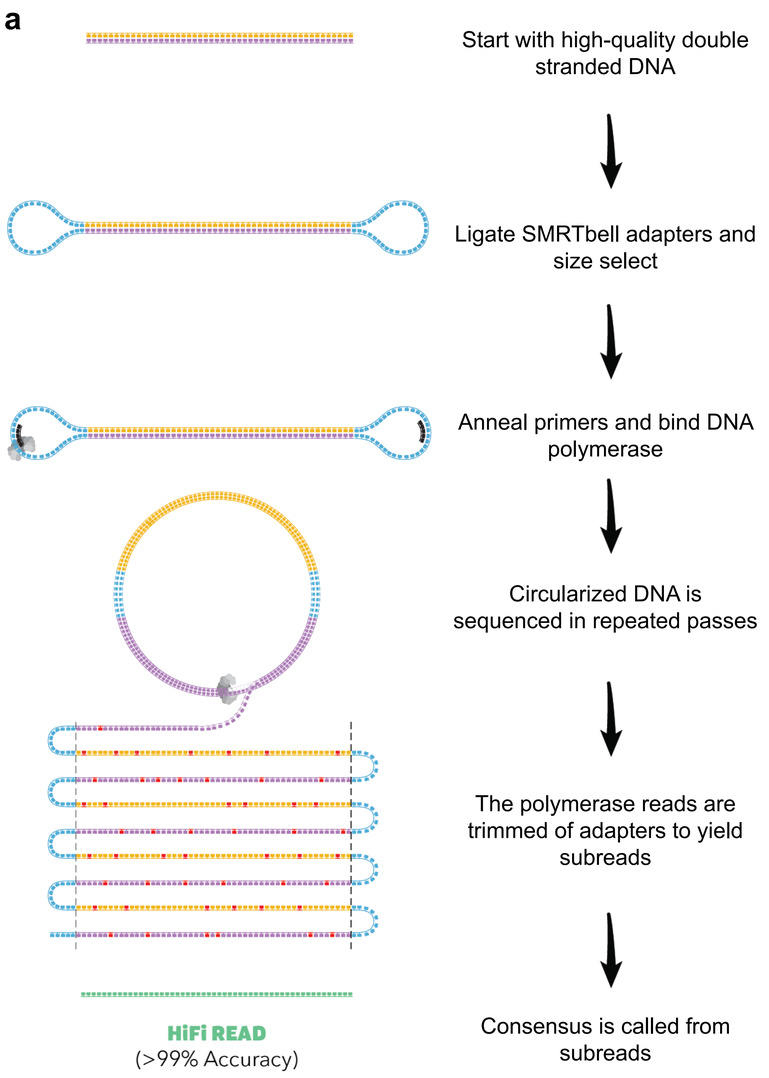


Figure 1. (a) PacBio Circular Consensus Sequencing produces HiFi reads[1] that are long (10-25 kb) and accurate (>99%). **(b)** HiFi reads align to difficult-to-map regions, including exons of the disease gene *STRC*.

Variant Detection with HiFi Reads

Variant calling with HiFi reads attains high precision and recall for single nucleotide variants (SNVs), indels, and structural variants (SVs).

Variant type	Precision	Recall
SNV	99.97%	99.95%
Indel (<50 bp)	99.10%	98.86%
SV (≥50 bp)	96.26%	94.93%

Table 1. Variant calling performance for 32-fold coverage of HG002 HiFi reads, measured against Genome in a Bottle (GIAB) benchmarks[2,3]. SNV and indel calling is with Google DeepVariant[4]. SV calling is with pbsv[5].

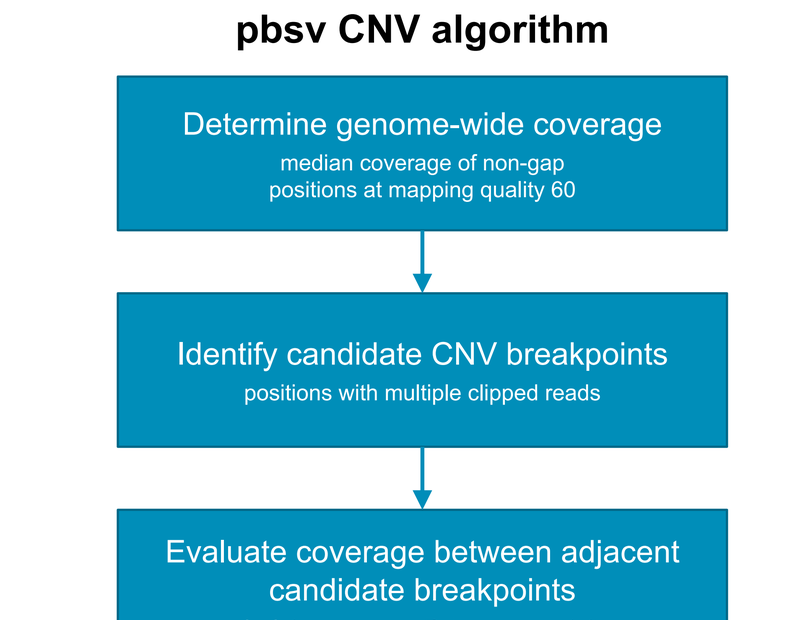
Copy-Number Variant (CNV) Detection with pbsv

Existing long read variant detection methods rely on *de novo* assembly or spanning reads. 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. 6)

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

pbsv CNV algorithm



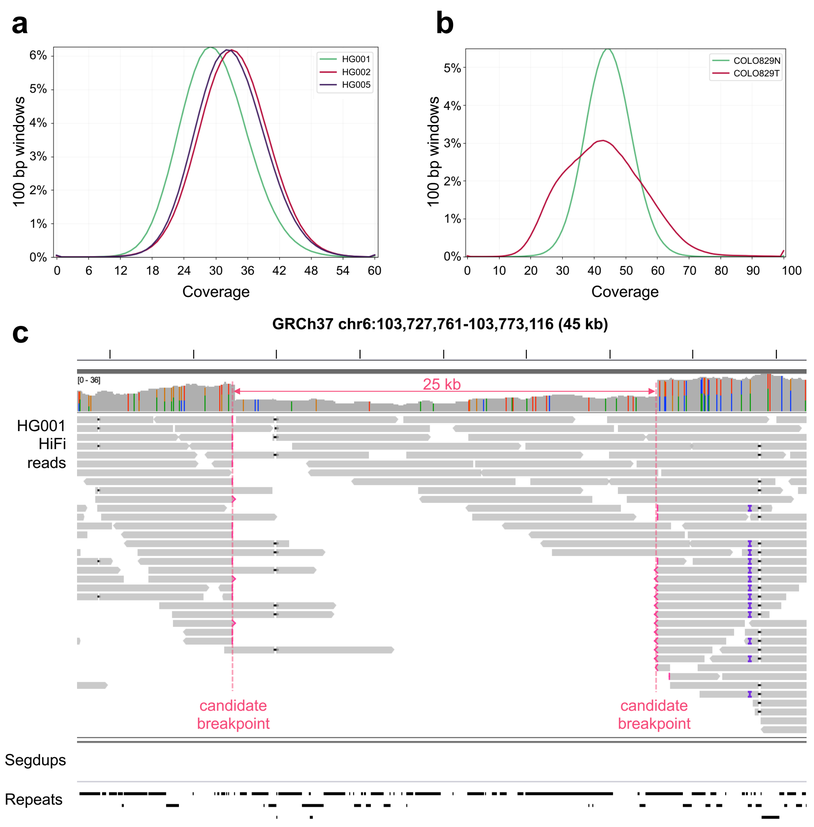


Figure 2. PacBio read coverage is Poisson distributed in autosomes for **(a)** samples from GIAB and **(b)** the normal sample from a tumor/normal pair. A tumor sample shows CNV regions of reduced and increased coverage. **(c)** To call CNVs, pbsv identifies candidate CNV breakpoints with multiple clipped reads and then evaluates read depth between adjacent breakpoints compared to the genome-wide typical coverage.

CNVs in HG001 and COLO829T

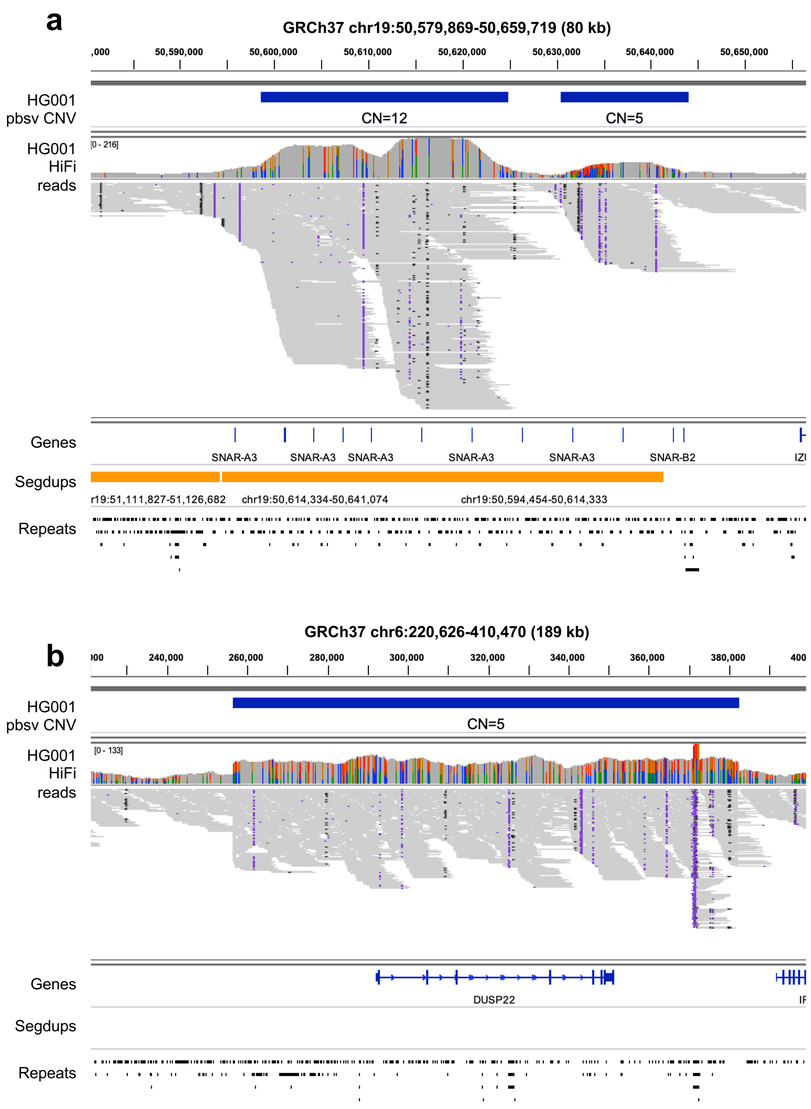


Figure 3. pbsv CNV calls in HG001 in **(a)** segmentally duplicated (segdup) sequence and **(b)** unique sequence.

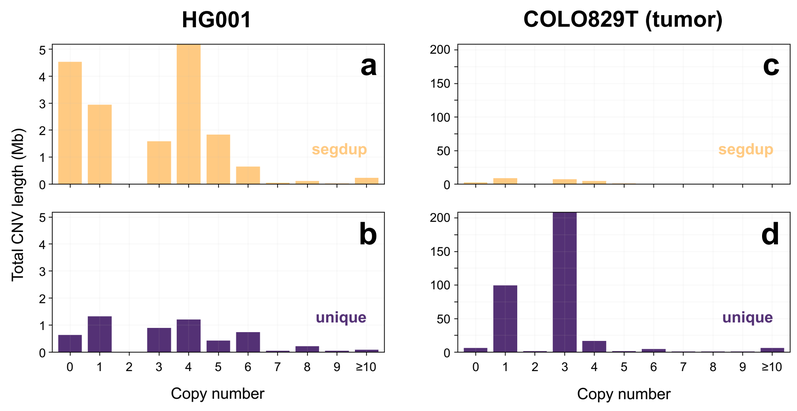


Figure 4. Genomic distribution of pbsv CNV calls, by copy number (normal = 2). **(a, b)** Most CNVs in the "healthy" genome HG001 involve segdups. **(c, d)** The tumor genome COLO829T has many more CNVs than HG001 and most fall outside of segdups.

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*. 84(2):125-140.
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. Poplin R et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 36(10):983-987.
5. https://github.com/PacificBiosciences/pbsv
6. Chaisson MJP et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 10(1):1784.

Thank you to Billy Rowell, David Scherer, and Pamela Bentley Mills for poster production support.