

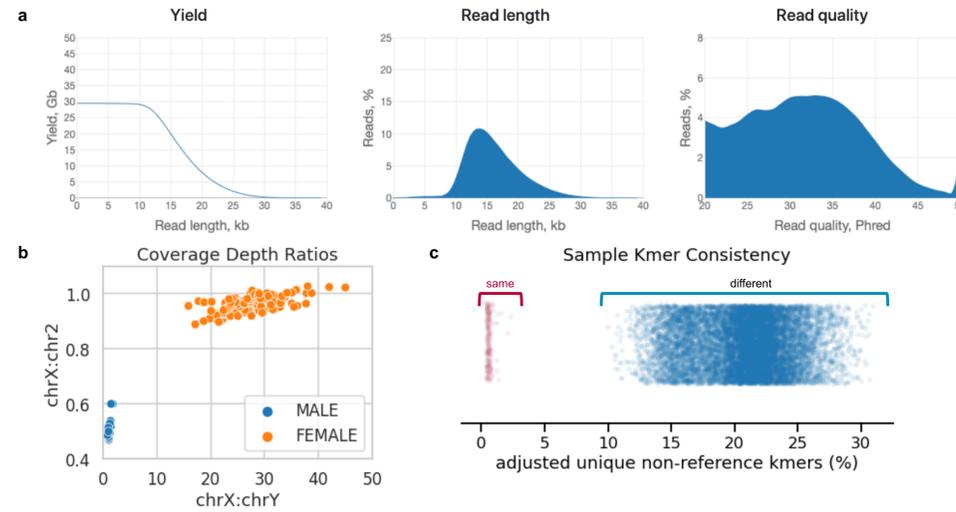
## PacBio HiFi reads provide comprehensive variant detection



**Figure 1. Circular Consensus Sequencing.** (a) A linear template sequence is (b) ligated to SMRTbell adapters. (c) DNA polymerase synthesizes complementary sequences to both strands of the original linear template, leading to (d) rolling circle sequencing and multiple passes of the original template. (e) CCS uses the noisy individual subreads to generate (f) a highly accurate consensus sequence (HiFi read).

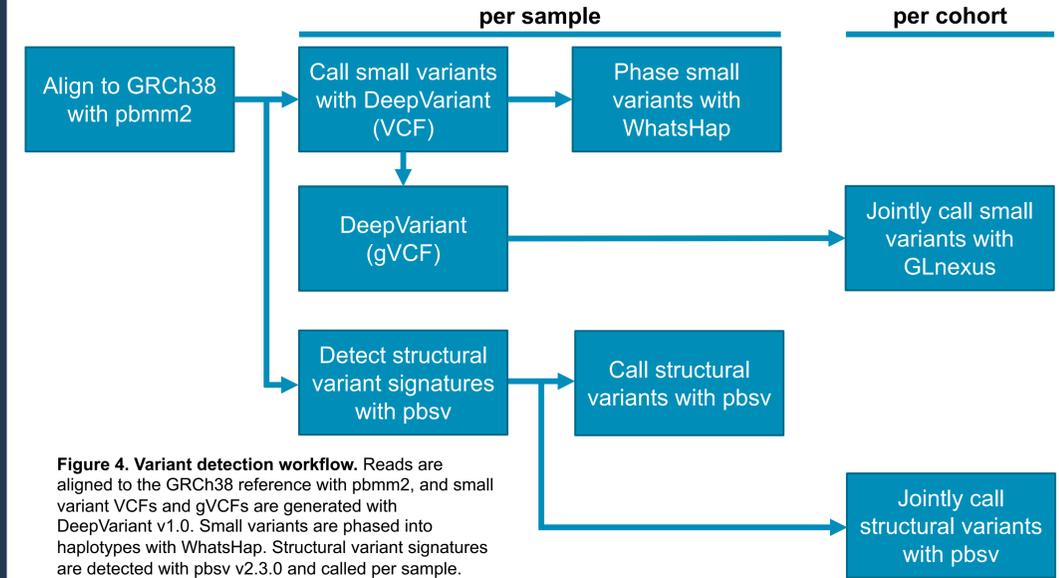
**Figure 2. HiFi reads detect more variants than short reads**  
HiFi reads detect over 300,000 single nucleotide variants (SNVs) and over 50,000 short indels missed by short reads, typically in regions that are difficult to map with short reads. HiFi reads detect more than 10,000 additional structural variants, including variants in difficult-to-map regions and long insertions.

## Sequencing metrics and quality control



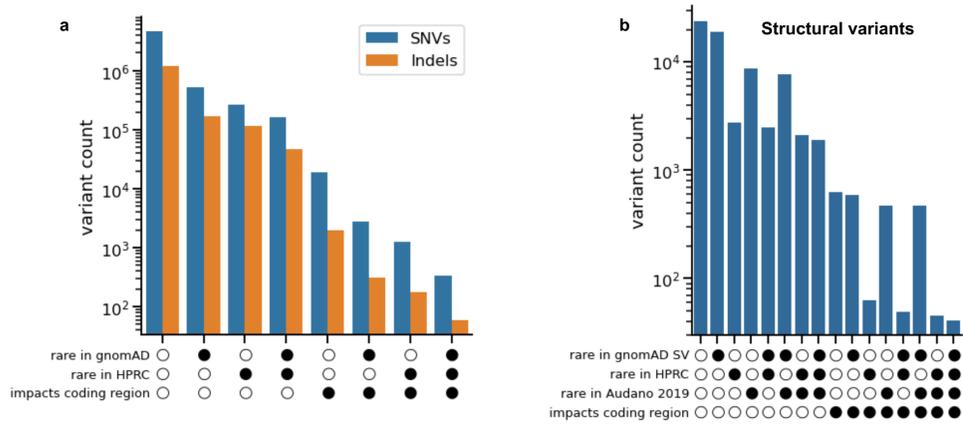
**Figure 3. Sequencing and sample quality control metrics.** (a) The quality of the sequencing runs is determined by assessing the HiFi yield, the read length distribution, and the HiFi read quality distribution. (b) Sex is inferred from the ratios of coverage depths of chrX:chrY and chrX:chr2. Cases where the inferred sex does not match the reported sex are flagged as a possible sample swap. (c) kmers are counted in reads for each SMRT Cell 8M, a subset of non-reference modimers is selected, and compared for each pair of SMRT Cells for a sample. Any pair with more than 3% unique non-reference kmers is flagged as a possible sample swap.

## Alignment and variant calling



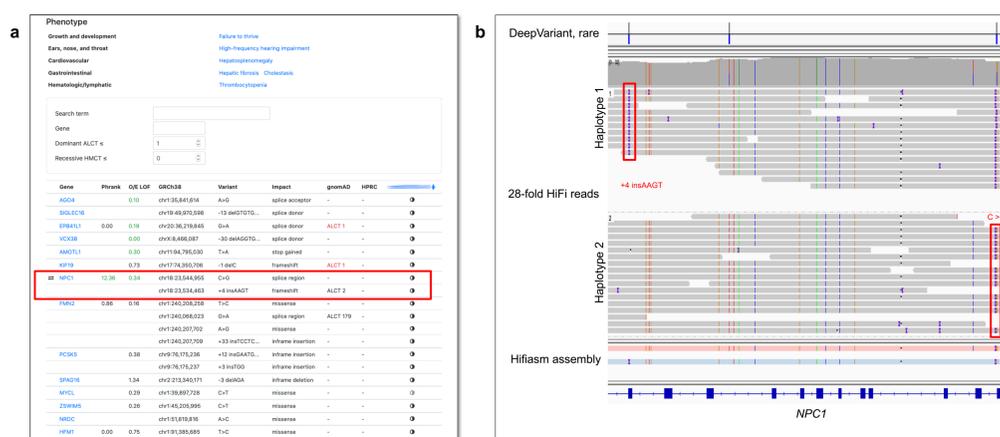
**Figure 4. Variant detection workflow.** Reads are aligned to the GRCh38 reference with pbmm2, and small variant VCFs and gVCFs are generated with DeepVariant v1.0. Small variants are phased into haplotypes with WhatsHap. Structural variant signatures are detected with pbsv v2.3.0 and called per sample. For multi-sample cohorts, small variants are jointly called from the DeepVariant gVCFs using GLnexus v1.2.7, and structural variants are jointly called with pbsv v2.3.0.

## Variant filtering and prioritization



**Figure 5. Identifying candidate pathogenic variants by filtering for rare variants predicted to impact a coding gene.** (a) Small variants are annotated with the allele count and number of homozygotes among the 71,702 short-read genomes in the gnomAD v3.0 catalog and 40 HiFi genomes from the Human Pangenome Reference Consortium (HPRC). Functional consequence is annotated based on Ensembl gene annotations. Filtering for variants that are rare in healthy samples (allele count < 3 & no homozygotes) reduces the number of candidate variants in a singleton by an order of magnitude. Additional filtering based on impact to a coding region reduces the number of variants by another two orders of magnitude, leaving hundreds of SNVs and tens of indels. (b) Structural variants are filtered based on similarity to variants in the 10,847 short-read genomes from gnomAD SV v2.1, 40 genomes from HPRC sequenced with HiFi, and 15 long-read genomes from Audano 2019. Filtering based on presence in these datasets reduces the number of variants by an order of magnitude. Additional filtering based on impact to a coding region reduces the number of variants by another order of magnitude.

## Variant review and visualization



**Figure 6. Variant review.** (a) Rare coding variants are annotated with the gnomAD observed/expected loss-of-function (O/E LOF) metric and the match between the case phenotype and the phenotypes of diseases caused by variants in the gene ("Phrank", Jagadeesh 2018). Results are presented in a web interface alongside patient phenotypes. Users can alter allele count thresholds to constrain/relax expectations for frequency of rare variants in gnomAD and HPRC genomes. The case shown has compound heterozygous rare variants in the *NPC1* gene, variants in which cause autosomal recessive Niemann-Pick disease, type C1, which is a good match to the case phenotype. The case has a C>G substitution in a splice region and a frameshifting insertion of AAGT in an exon over 10 kb downstream. (b) In IGV, phased haplotypes show that these two variants are on opposite alleles.

## Overall design and availability

All code is available at <https://github.com/williamrowell/pbRUGD-workflow> and <https://github.com/amwenger/pbRUGD-www>.

The workflow is implemented as three Snakemake workflows: 1) `process_smrtcells`, which aligns HiFi reads and generates SMRT Cell quality metrics, 2) `process_samples`, which calls variants, generates a HiFi assembly, and generates per-sample quality metrics, and 3) `process_cohorts`, which filters and prioritizes variants. Individual workflows can be triggered manually or by `cron`. Cohort and sample information are stored in a flexible `yaml` format, and results are reviewed through a custom web interface.

**Pipeline resources**  
<https://github.com/PacificBiosciences/pbmm2>  
<https://github.com/PacificBiosciences/pbsv>  
<https://github.com/chyip123/hifiasm>  
<https://github.com/dnanexus-rnd/GLnexus>  
<https://github.com/brentp/mosdepth>  
<https://github.com/google/deepvariant>  
<https://github.com/h3/calN50>  
<https://github.com/brentp/sliver>  
<https://github.com/gmarcisz/lellyfish>  
<https://github.com/whatsHap/whatsHap>  
<https://github.com/h3/minimap2>  
<https://github.com/amwenger/svpack>

**References**  
 Audano PA, Sulovari A, Graves-Lindsay TA, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* (2019), 176(3):663-675.e19. <https://doi.org/10.1016/j.cell.2018.12.019>  
 Cheng H, Concepcion GT, Feng X, et al. Haplotype-resolved de novo assembly with phased assembly graphs. *arXiv* (2020). <https://arxiv.org/abs/2008.01237>  
 Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarly to greatly improve Mendelian diagnostic disease prioritization. *Genet Med* (2019), 21: 464-470. <https://doi.org/10.1038/s41436-018-0072-y>  
 Karzewski KJ, Francioli LC, Tiao G et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020), 581: 434-443. <https://doi.org/10.1038/s41586-020-2308-7>  
 Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* (2012), 28(19): 2520-2522. <https://doi.org/10.1093/bioinformatics/bts480>  
 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (2018), 34:3094-3100. <https://doi.org/10.1093/bioinformatics/bty191>  
 Marcias G and Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* (2011), 27(6): 764-770. <https://doi.org/10.1093/bioinformatics/btr011>  
 Pedersen BS, Quinlan AR, Mosdeglia: quick coverage calculation for genomes and exomes. *Bioinformatics* (2018), 34(5): 867-868. <https://doi.org/10.1093/bioinformatics/btx699>  
 Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018), 36: 983-987. <https://doi.org/10.1038/nbt.4235>  
 Wagner J, Olson ND, Harris L. Benchmarking challenging small variants with linked and long reads. *bioRxiv* (2020). <https://doi.org/10.1101/2020.07.24.212712>  
 Yun T, Li H, Chang PC, et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *bioRxiv* (2020). <https://doi.org/10.1101/2020.02.10.942086v1>  
 Marcel Martin, Murray Patterson, Shipra Garg, et al. WhatsHap: fast and accurate read-based phasing. *bioRxiv* (2016). <https://doi.org/10.1101/085050>

**Acknowledgments**  
 Thanks to Shreya Chakraborty, Christine Lambert, and Primo Baybayan from PacBio for library prep and sequencing. Thanks to Tomi Pastinen, Neil Miller, and Emily Farrow from Children's Mercy Kansas City and the Genomic Answers for Kids program. Thanks to the Human Pangenome Reference Consortium for access to 40 human genomes sequenced with HiFi.

