Comprehensive variant detection in a human genome with PacBio high-fidelity reads

William J. Rowell, Paul Peluso, John Harting, Yufeng Qian, Aaron Wenger, Richard Hall, David R. Rank
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract

Human genomic variations range in size from single nucleotide substitutions to large chromosomal rearrangements. Sequencing technologies tend to be optimized for detecting particular variant types and sizes. Short reads excel at detecting SNVs and small indels, while long or linked reads are typically used to detect larger structural variants or phase distant loci. Long reads are more easily mapped to repetitive regions, but tend to have lower per-base accuracy, making it difficult to call short variants. The PacBio Sequel System produces two main data types: long continuous reads (up to 100 kbp), generated by single passes over a long template, and Circular Consensus Sequence (CCS) reads, generated by calculating the consensus of many sequencing passes over a single shorter template (500 bp to 20 kbp). The long-range information in continuous reads is useful for genome assembly and structural variant detection. The higher base accuracy of CCS effectively detects and phases short variants in single molecules.

Recent improvements in library preparation protocols and sequencing chemistry have increased the length, accuracy, and throughput of CCS reads. For the human sample HG002, we collected 28-fold coverage 15 kbp high-fidelity CCS reads with an average read quality above Q20 (99% accuracy). The length and accuracy of these reads allow us to detect SNVs, indels, and structural variants not only in the Genome in a Bottle (GIAB) high confidence regions, but also in segmental duplications, HLA loci, and clinically relevant "difficult-to-map" genes.

As with continuous long reads, we call structural variants at 90.0% recall compared to the GIAB structural variant benchmark "truth" set, with the added advantages of base pair resolution for variant calls and improved recall at compound heterozygous loci.

With minimap2 alignments, GATK4 HaplotypeCaller variant calls, and simple variant filtration, we have achieved a SNP F-Score of 99.51% and an INDEL F-Score of 80.10% against the GIAB short variant benchmark "truth" set, in addition to calling variants outside of the high confidence region established by GIAB using previous technologies. With the long-range information available in 15 kbp reads, we applied the read-backed phasing tool WhatsHap to generate phase blocks with a mean length of 65 kbp across the entire genome. Using an alignment-based approach, we typed all major MHC class I and class II genes to at least 3-field precision.

This new data type has the potential to expand the GIAB high confidence regions and "truth" benchmark sets to many previously difficult-to-map genes and allow a single sequencing protocol to address both short variants and large structural variants.

Sample Preparation and Primary Analysis

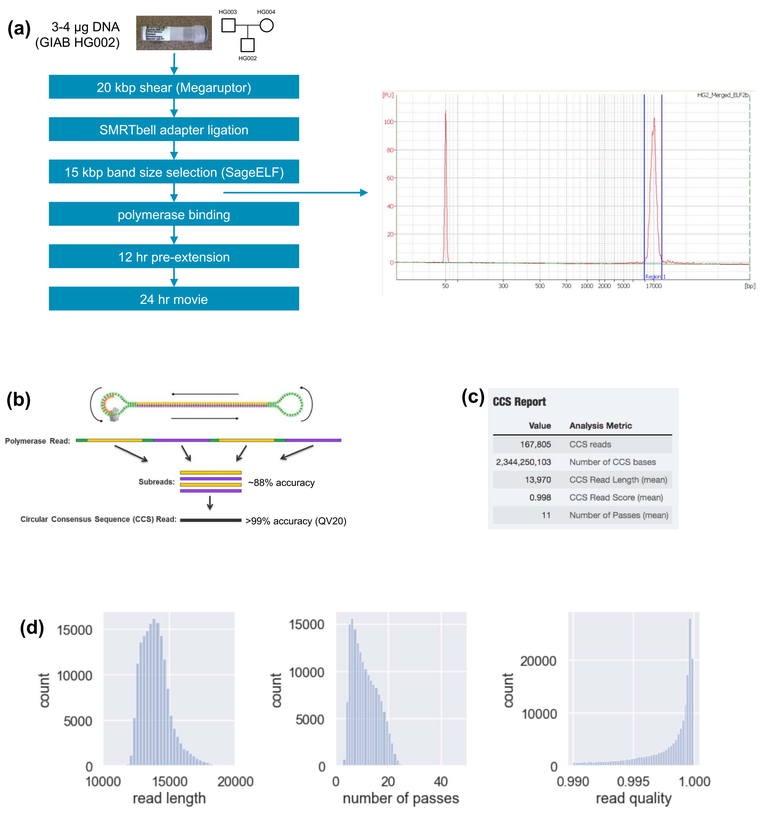

Figure 1. Generation of high-fidelity long reads. (a) Following shearing and SMRTbell library creation, the SageELF system was used to select the 15 kbp band, which was sequenced for 20 hours on the Sequel System. **(b)** High-fidelity reads were generated as the consensus of multiple sequencing passes over single SMRTbell templates. **(c)** Q20+ CCS reads from a representative single Sequel SMRT Cell. **(d)** High-fidelity reads have a tight read length distribution around 14 kbp, and read qualities approaching Q30.

Mapping

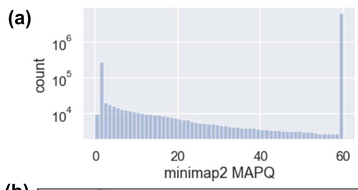

Figure 2. High-Fidelity reads have improved 'mappability' compared to short reads. (a) The majority of high-fidelity reads can be mapped at MAQP60. **(b)** 15 kb high-fidelity reads are mapped with high confidence to the 123 kb region centered on *CYP2D6*, which has several large segmental duplications.

Structural Variant Calling

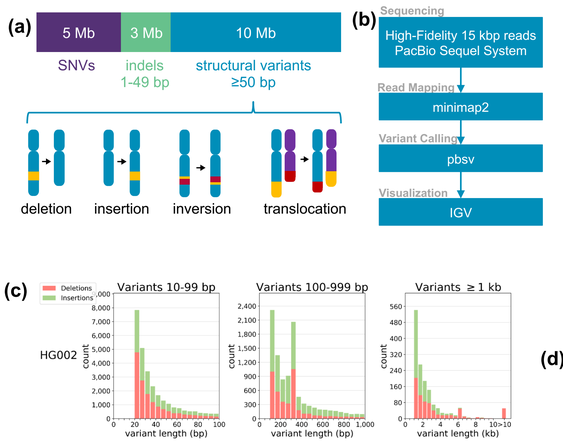

Figure 3. Structural variant detection. (a) Structural variants of 20 bp and larger can be detected **(b)** with minimap2 alignment followed by pbsv analysis. **(c)** ~39,000 variants are detected in HG002, and **(d)** we achieve high recall and precision compared to the GIAB SV Tier1 v0.6 benchmark.

Short Variant Calling

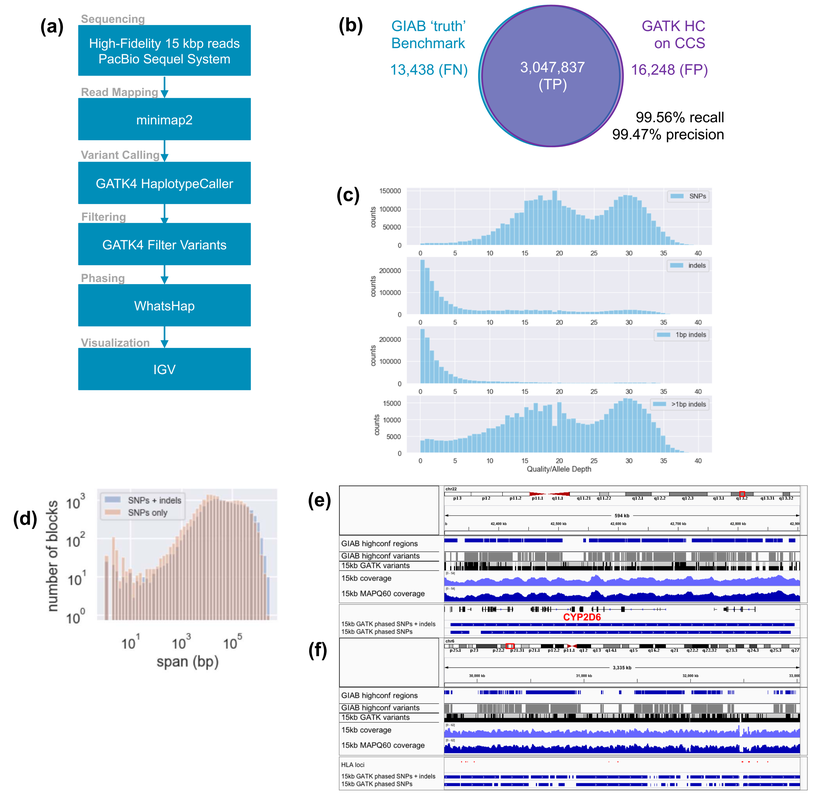

Figure 4. Short variant detection and phasing. (a) High-fidelity reads can be used with pre-existing short variant callers, **(b)** achieving SNP recall and precision comparable with short read technology at similar depths. **(c)** Unfiltered indels are dominated by a low quality peak corresponding to 1 bp indels, which can be addressed by splitting indels by size before applying filters. **(d)** Distribution of WhatsHap read-backed phase block lengths. **(e)** 500 kbp phase blocks covering the difficult-to-sequence CYP2D6 locus on chromosome 22. **(f)** Phasing of the HLA loci.

Clinically Relevant Genes

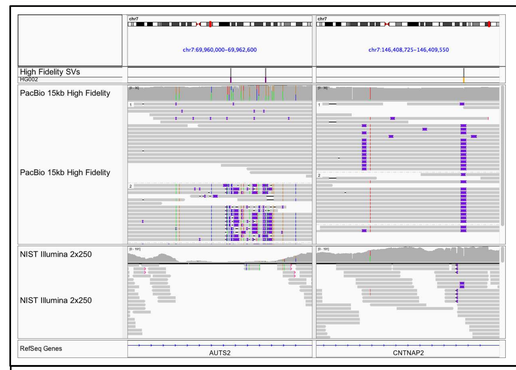

Figure 5. Structural variant detection. High-fidelity reads can be more easily aligned in repetitive regions, extending SV detection into previously difficult-to-map regions like simple repeats in the *AUTS2* locus (left panel), implicated in autism, and the *CNTNAP2* locus (right panel), implicated in epilepsy.

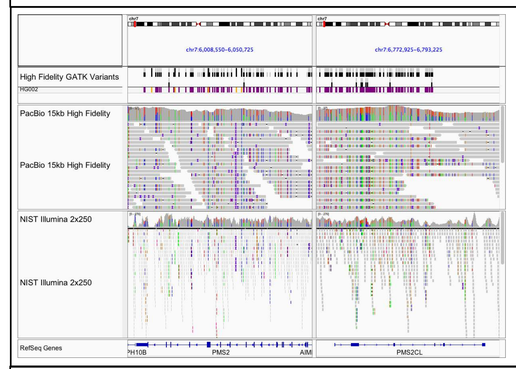

Figure 6. Short variant detection in *PMS2*/*PMS2CL*. High-fidelity reads can be assigned to the correct locus of the gene-pseudogene pair *PMS2* (left panel) and *PMS2CL* (right panel). Using a simple GATK workflow, short variants can be detected over the majority of the difficult-to-map *PMS2* locus, a gene associated with Lynch syndrome.

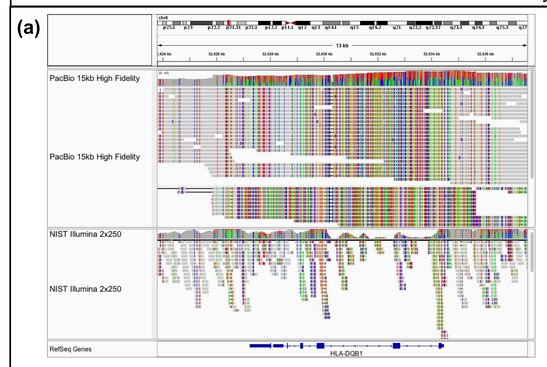

Figure 7. HLA typing. (a) High-fidelity reads are easily aligned to MHC class I and II genes, such as HLA-DQB1, and **(b)** allow typing to three fields or greater.

Conclusions

- High-fidelity long reads can be mapped definitively to more parts of the genome than short reads, including segmental duplications, tandem repeats, structural variants, and highly divergent sequences.
- Structural variant detection with PBSV is similar to that of continuous long reads (CLR), with the added benefit of base pair resolution of breakpoints.
- Because of the increased accuracy, high-fidelity long reads can be used with existing workflows to detect short variants with comparable SNP F1 score. The majority of indel miscalls are 1bp in length, and precision can be improved by splitting indels by size before applying filters.
- High-fidelity long reads give users the ability to detect both short variants and structural variants with the same dataset.

References

- **From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.** Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M, 2013 *CURRENT PROTOCOLS IN BIOINFORMATICS 43:11.10.1-11.10.33*
- **WhatsHap: fast and accurate read-based phasing.** Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, Tobias Marschall. bioRxiv 085050; doi: https://doi.org/10.1101/085050
- http://www.broadinstitute.org/igv
- https://www.pacb.com/support/software-downloads/