Comprehensive variant detection in a human genome with highly accurate long reads

William J. Rowell*, Aaron M. Wenger†, Alexey Kolesnikov*, Pi-Chuan Chang‡, Andrew Carroll³, Richard J. Hall³
*Pacific Biosciences, Menlo Park, CA, USA
†Google Inc., Mountain View, CA, USA

Introduction

• Long-read sequencing has been applied successfully to assemble genomes and detect structural variants.
• Long reads can be unambiguously mapped to more of the genome than short reads of comparable accuracy.
• It has been difficult to call small variants from long reads due to higher read error rates.

Figure 1. Sample preparation and sequencing workflow. (a) Workflow steps. (b) Bioanalyzer trace for SMRTbell library of sheared fragments. (c) Bioanalyzer trace for SageELF fractions of Bioanalyzer trace for SMRTbell library. (d) 11 kb fraction was used for sequencing on three SMRT Cells 8M.

Methods

Figure 2. Circular Consensus Sequencing. (a) Linear template sequence is ligated to SMRTbell adapters. (b) DNA polymerase synthesizes complementary sequences to both strands of the original linear template, leading to (c) rolling circle sequencing and multiple passes of the original template. (e) CCS uses the noisy individual subreads to generate (f) highly accurate consensus sequence (HiFi read). (g) With 10 passes, CCS accuracy is approximately QV30 (99.9%).

Analysis

• Align on reference genome with pbsv

Figure 3. Bioinformatics workflow for read mapping and variant detection. Approximately 17-fold coverage (three SMRT Cells 8M) of highly accurate (average 99.8%) 11.1 kb reads were mapped to the hg19 reference with bbmap2. Single nucleotide variants (SNVs) and small indels (<50 bp) were detected using Google DeepVariant with a model trained for CCS. Structural variants (SVs) were detected with pbsv. Variant calls were evaluated against Genome in a Bottle (GAB) benchmarks.

• DeepVariant small variants

Table 1. Variant detection compared against Genome in a Bottle benchmarks. SNVs and indels (<50 bp) were detected using Google DeepVariant with a model trained for CCS. Structural variants (SVs) were detected with pbsv. Variant calls were evaluated against Genome in a Bottle (GAB) benchmarks.

Benchmarking

Table 2. Small variant phasing metrics. Small variants detected by DeepVariant were phased using long-range information available in CCS reads. Blocks size metrics were generated by WhatzHap.

Conclusions

• High accuracy and long read lengths allow detection of both small variants and structural variants.
• Three SMRT Cells 8M provided approximately 17-fold coverage of the human genome, which achieves highly accurate variant detection with DeepVariant and pbsv as compared against GAB benchmarks.
• Read lengths >10 kb and lack of GC bias provide improved mappability over short-read sequencing.

Figure 4. 5'7CR locus. Segmental duplications within this locus prevent the unambiguous alignment of short reads. HiFi reads can be aligned not only to all exons at this locus, but intronic sequence as well, allowing for small variants to be called by DeepVariant across the entire locus.

Figure 5. KRT76 locus. The length and accuracy of HiFi reads allow small variant calling even in reference-divergent regions within segmental duplications.

Figure 6. SPATA31C1 locus. Longer read lengths allow pbsv to detect a heterozygous 3.4 kb deletion covering the first exon.

Read more about highly accurate long reads in our preprint on bioRxiv.

Data and alignment

https://github.com/Illumina/hap.py
https://github.com/PacificBiosciences/SMRTstream

Variant Detection

https://github.com/google/deepvariant
https://github.com/PacificBiosciences/pbsv

References

Thanks to Paul Peluso and Shreyasee Chakraborty for data generation.