Detection and phasing of small variants in Genome in a Bottle samples with highly accurate long reads

William J. Rowell, Shreyasee Chakraborty, Primo Baybayan, Richard J. Hall
Pacific Biosciences, 1305 O’Brien Drive, Menlo Park, CA 94025

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.
- However, it has been challenging to call small variants from long reads due to higher error rates for raw reads.
- PacBio HiFi reads that are >99% accurate and 10-20 kb in length enable detection of small and large variants, increasing discovery power for human genetics research.
- We sequenced the Genome in a Bottle (GIAB) reference samples HG001, HG002, and HG005 to ~30-fold coverage to determine the optimal coverage depth for small variant detection and phasing.

Methods

a. BluePippin 20 kb linear template sequence is fold coverage to determine the optimal read sequencing has been applied
b. Methods

c. Long reads can be unambiguously mapped to more of the genome than short reads of comparable accuracy. Long reads can be unambiguously mapped to more of the genome than short reads of comparable accuracy.
d. With 8 passes, average accuracy is ~QV30 (99.9%).

Analysis

Align to reference genome with pbmm2
Randomly down-sample aligned reads
For each depth titration point, detect small variants with DeepVariant and phase with Whap.py
For each depth titration point, compare to GIAB small variant benchmark using Hap.py

Results

Figure 2. Bioinformatics workflow for read mapping and variant detection. ~30-fold coverage (six SMRT Cells SMRT Cells 8M with Sequel II System chemistry 1.0) of highly accurate (average 99.8%) 11 kb reads were mapped to the hg19 (hs37d5) reference with pbmm2. For each sample, aligned reads were randomly down-sampled 10 times for each 10% coverage depth increment. Single nucleotide variants (SNVs) and small indels (<50 bp) were detected using Google DeepVariant v0.8.0 and phased with Whap v0.18. Variant calls were evaluated against GIAB v3.3.2 benchmarks.

Table 1. Small variant benchmarks at ~15-fold coverage. (a) Precision and recall for SNVs and indels at different coverage titration levels. (b) Size of maximum phase block and phase block N50 at different coverage titration levels. All values are mean +/- standard deviation, n=10 for each sample at each coverage.

Table 2. Long, highly accurate reads can be mapped unambiguously through difficult regions. Mapped HiFi reads can easily be used to detect and phase small variants, even through repetitive sequence at the CYP2D6/CYP2D7 locus.

Figure 3. Small variant benchmarks and phasing statistics over titration of coverage depth. (a) Precision and recall for SNVs and indels at different coverage titration levels. (b) Size of maximum phase block and phase block N50 at different coverage titration levels. (c) DeepVariant calls on PacBio HiFi reads achieve >99% precision and recall for SNVs, increasing to >99.9% precision and recall at 30-fold coverage.

Conclusions

- At 15-fold coverage, DeepVariant calls on PacBio HiFi reads achieve >99% precision and recall for SNVs, increasing to >99.9% precision and recall at 30-fold coverage.
- At 15-fold coverage, DeepVariant calls on PacBio HiFi reads achieve >95% indel precision and recall, increasing to ~98% precision and recall at 30-fold coverage.
- We see very little increase in maximum phase block size or phase block N50 above 15-fold coverage.

References and Data Availability

6. PacBio HiFi reads generated on Sequel II System using Chemistry 1.0 are deposited in SRA under BioProjects PRJNA527279, PRJNA542705, and PRJNA544721.
7. Alignments to hg19/37 available from GIAB:
@nothingclever wrowell@pacbio.com