Structural Variant Calling for PacBio HiFi and Oxford Nanopore long reads

Release 202112.07

Sentieon, Inc

Apr 24, 2023

Contents

1 Introduction

This document describes using Sentieon® LongReads SV to call structural variants (SV) from PacBio HiFi and Oxford Nanopore long reads. Accurate long read sequencing technologies enable accurate discovery of large SV events that were previously inaccessible with short read approaches.

Sentieon® LongReads SV is able to take advantage of much longer read lengths to perform quick and accurate detection and genotyping for large SV evnts for both PacBio HiFi and Oxford Nanopore long reads input.

Sentieon® LongReads SV expects aligned BAM or CRAM file(s) as input and will output variants in VCF format. We recommend using Sentieon®'s accelerated Minimap2 and sorting to perform efficient and accurate alignment.

Currently, Sentieon® LongReads SV detects and genotypes large INDEL events, and outputs high-confidence concensenus INDEL base sequence. Support for other SV event types will be added in future release.

If you have any additional questions, please contact the technical support at Sentieon® Inc. at support@sentieon.com.

2 Using LongReads SV

To run Sentieon LongReads SV, run the following command:

```
sentieon driver -t NUMBER_THREADS -r REFERENCE -i INPUT_BAM \
  --algo LongReadSV [--min_sv_size MIN_SV_SIZE] [--min_map_qual MIN_MAP_QUAL]\
  --model MODEL OUT_SV_VCF
```

The following arugments are required for the command:

- `-t NUMBER_THREADS`: the number of computer threads that will be used in the calculation. We recommend that the number does not exceed the number of computing cores available in your system.
- `-r REFERENCE`: the location of the reference FASTA file. You should make sure that the reference is the same as the one used in the mapping stage.
- `-i INPUT_BAM`: The input alignment file should be an indexed BAM or CRAM file of PacBio HiFi or Oxford Nanopore reads aligned with `minimap2` or `pbmm2` for HiFi reads.
- `--model MODEL`: input model file that contains encoded preset configuration for either PacBio HiFi or Oxford Nanopore long reads. You need to select a model file that matches the platform of the input alignment file. Please refer to Sentieon's GitHub page[1] to download the latest SV model.

The command requires the following positional argument:

- `OUT_SV_VCF`: the location and filename of the SV calling output with `.vcf` or `.vcf.gz` extension. A corresponding index file will be created. The tool will output a compressed file for `.gz` extension.

The following arugments are optional for the command:

- `--min_sv_size MIN_SV_SIZE`: minimum SV size in basepairs to output (default: 40)
- `--min_map_qual MIN_MAP_QUAL`: minimum read mapping quality (default: 20)

3 Evaluate LongReads SV result

Sentieon® recommends using Sentieon® hap-eval for accurate evaluation and comparison of SV calling results based on assembled haplotypes. Hap-eval is an open-sourced VCF comparison engine for structual variant benchmarking. It is available for download at GitHub[2].

https://github.com/Sentieon/sentieon-models
https://github.com/Sentieon/hap-eval