DNAscope LongRead germline variant calling

Release 202308.03

Sentieon, Inc

Oct 23, 2024

Contents

1 Introduction

This document describes using Sentieon® DNAscope to call germline variants from PacBio® HiFi or Oxford Nanopore (ONT) long reads.

Sentieon® DNAscope is able to take advantage of the long read length of PacBio® HiFi and ONT reads to perform quick and accurate variant calling using specially calibrated machine learning models.

In order to run this pipeline you need to use Sentieon software package version 202308 or higher; in addition, you will need to use a set of scripts that can be obtained from, https://github.com/Sentieon/sentieon-scripts/tree/master/dnascope_LongRead. You will also need a model bundle for your platform. The latest model bundles can be found at, https://github.com/Sentieon/sentieon-models.

The pipeline also requires python[1] version >2.7 or >3.3, bcftools[2] version 1.10 or higher, and bedtools[3]. The `sentieon`,

[1] https://www.python.org/
[2] http://samtools.github.io/bcftools/bcftools.html
[3] https://bedtools.readthedocs.io/en/latest/

`python`, `bcftools`, and `bedtools` executables will be accessed through the user's `PATH` environment variable.

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

2 Input data requirements

2.1 Aligned reads - PacBio HiFi

As input, the pipeline will take PacBio® HiFi reads that have been aligned to a reference genome with `pbmm2` or `minimap2`. When aligning reads with pbmm2, setting `-c 0 -y 70 --preset HIFI` is recommend. These settings turn off pbmm2's legacy mapped concordance filter in favor of a gap compressed sequence identity filter for output alignments and turn on PacBio's® recommended settings for alignment of HiFi reads. When aligning reads with Sentieon minimap2, using the Sentieon model for HiFi reads is recommended. When aligning reads with minimap2, setting `-x map-hifi` is recommend. This turns on the recommended minimap2 settings for alignment of HiFi reads.

2.2 Aligned reads - ONT

The pipeline will accept Oxford Nanopore (ONT) long reads that have been aligned to the reference genome with `minimap2`. When aligning reads with Sentieon minimap2, using the Sentieon model for ONT is recommended. When aligning reads with the open-source minimap2, `-x map-ont` is recommended.

2.3 The Reference genome

DNAscope will call variants present in the sample relative to a high quality reference genome sequence. Besides the reference genome file, a `samtools` fasta index file (.fai) needs to be present. We recommend aligning to a reference genome without alternate contigs.

3 The Sentieon® DNAscope LongRead pipeline

3.1 Pipeline overview

The pipeline will perform two passes of variant calling from the input alignment file and will merge the generated VCFs to produce the final output file. The steps of the pipeline are:

- A first pass of variant calling, in which the pipeline will call variants present in the sample of interest.
- SNVs called in the first pass are then phased using the long-read information.
- A second pass of variant calling.
  - Across phased regions, variants are called from each haplotype separately.
  - Across unphased regions, variants are called with a more accurate diploid model.
- Variants from the first and second passes are then combined to generate a final callset.

The pipeline requires a DNAscope machine learning model. Please refer to Sentieon's GitHub page[4] to download the latest model.

[4] https://github.com/Sentieon/sentieon-models

3.2 Running the pipeline

Running the DNAscope LongRead pipeline is done via a script that makes the necessary calls to individual Sentieon commands. Different scripts are used for the HiFi and ONT pipelines.

A single command is run to call variants from PacBio HiFi reads and to apply the machine learning models. The input alignment file should be an indexed BAM or CRAM file of HiFi reads aligned with `pbmm2` or `minimap2`.

```
dnascope_HiFi.sh [-h] -r REFERENCE -i INPUT_BAM -m MODEL_BUNDLE [-d DBSNP_VCF] [-b DIPLOID_BED] [-t NUMBER_
↪THREADS] [-g]  [--] VARIANT_VCF
```

A single command is run to call variants from ONT reads and to apply the machine learning models. The input alignment file should be an indexed BAM or CRAM file of ONT reads aligned with `minimap2`.

```
dnascope_ONT.sh [-h] -r REFERENCE -i INPUT_BAM -m MODEL_BUNDLE [-d DBSNP_VCF] [-b DIPLOID_BED] [-t NUMBER_
↪THREADS] [-g]  [--] VARIANT_VCF
```

The Sentieon® DNAscope LongRead pipeline requires the following arguments:

- `-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 location of the input BAM or CRAM file.
- `-m MODEL_BUNDLE`: the location of the model bundle.

The Sentieon® DNAscope LongRead pipeline accepts the following optional arguments:

- `-d dbSNP`: the location of the Single Nucleotide Polymorphism database (dbSNP) used to label known variants. Only one file is supported. Supplying this file will annotate variants with their dbSNP refSNP ID numbers.
- `-b INTERVAL`: interval in the reference that will be used in the software, in BED file format. Supplying this file will limit variant calling to the intervals inside the BED file.
- `-t NUMBER_THREADS`: number of computing threads that will be used by the software to run parallel processes. The argument is optional; if omitted, the pipeline will use as many threads as the server has.
- `-g`: output variants in the gVCF format, in addition to the VCF output file. The tool will output a bgzip compressed gVCF file with a corresponding index file.
- `-h`: print the command-line help and exit.

The Sentieon® DNAscope LongRead pipeline requires the following positional arguments:

- `VARIANT_VCF`: the location and filename of the variant calling output. The tool will output a bgzip compressed VCF file with a corresponding index file.

3.3 Pipeline output

The DNAscope LongRead pipeline will output a bgzip compressed file (.vcf.gz) containing variant calls in the standard VCFv4.2 format along with a tabix index file (.vcf.gz.tbi). If the -g option is used, the pipeline will also output a bgzip compressed file (.g.vcf.gz) containing variant calls in the gVCF format along with a tabix index file (.g.vcf.gz.tbi).

3.4 Other considerations

Currently, the pipeline is only recommended for use with samples from diploid organisms. For samples with both diploid and haploid chromosomes, the `-b INTERVAL` option can be used to limit variant calling to diploid chromosomes.