



DNAScope Germline variant calling for HiFi reads

Release 202112.07

Sentieon, Inc

Apr 24, 2023

Contents

1	Introduction	1
2	Input data requirements	2
2.1	Aligned reads	2
2.2	The Reference genome	2
3	The Sentieon® DNAScope pipeline for PacBio® HiFi reads	2
3.1	Pipeline overview	2
3.2	Running the pipeline	2
3.3	Pipeline output	3
3.4	Other considerations	3

1 Introduction

This document describes using Sentieon® DNAScope to call germline variants from PacBio® HiFi reads. The PacBio® HiFi technology produces highly accurate long reads with base quality exceeding Q20 and average read lengths between 10 and 25kbases. Accurate long reads enable accurate variant calling across repetitive regions of the genome that were previously inaccessible with short read or noisy long read approaches.

Sentieon® DNAScope is able to take advantage of the improved quality and long read length of PacBio® HiFi reads to perform quick and accurate variant calling using specially calibrated machine learning models. The DNAScope pipeline for HiFi reads expects aligned HiFi reads as input and will output variants in VCF format.

In order to run this pipeline you need to use Sentieon software package version 202010.03 or higher; in addition, you will need to use a set of scripts that can be obtained by contacting Sentieon®. The pipeline also requires [python](https://www.python.org/)¹

¹<https://www.python.org/>

version >2.7 or >3.3, [bcftools](#)² version 1.10 or higher, and [bedtools](#)³. The 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

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 minimap2, setting `-x map-hifi` is recommend. This turns on the recommended minimap2 settings for alignment of HiFi reads.

2.2 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 pipeline for PacBio® HiFi reads

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.
- By suppling an optional MHC BED file, additional special handling can be applied to the MHC region to further increase variant calling accuracy.

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

3.2 Running the pipeline

Running the DNAscope pipeline for HiFi reads is done via a script that makes the necessary calls to individual Sentieon commands. A single command is run to call variants 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.

<http://samtools.github.io/bcftools/bcftools.html>

<https://bedtools.readthedocs.io/en/latest/>

<https://github.com/Sentieon/sentieon-models>

```
dnascope_HiFi.sh [-h] -r REFERENCE -i HIFI_BAM -m MODEL [-d dbSNP] [-B MHC_INTERVAL] [-b INTERVAL] [-t NUMBER_
↳THREADS] [-h] [--] VARIANT_VCF
```

The Sentieon® DNAscope pipeline for HiFi reads 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 HIFI_BAM`: the location of the input BAM file containing the HiFi reads mapped by pbmm2.
- `-m MODEL`: the location of a DNAscope HiFi model file.

The Sentieon® DNAscope pipeline for HiFi reads 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 MHC_INTERVAL`: interval in the reference that includes the MHC regions, in BED file format. Supplying this file will enable the special handling of the MHC region.
- `-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.
- `-h`: print the command-line help and exit.

The Sentieon® DNAscope pipeline for HiFi reads 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 pipeline for HiFi reads 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`).

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.