DNAscope Germline variant calling with a machine learning model

Release 202112.07

Sentieon, Inc

Apr 24, 2023

Contents

1 Introduction

This documents describes the capabilities of DNAscope for germline calling using a machine learning model. If you have any additional questions, please contact the technical support at Sentieon® Inc. at support@sentieon.com.

2 Model Performance

Please refer to our github page https://github.com/Sentieon/sentieon-dnascope-ml to find out the latest DNAscope model's performance and runtime. Please refer to Sentieon's GitHub page[1] to download the latest model.

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

3 Using a machine learning model in DNAscope

3.1 Goal of a machine learning model in DNAscope

From version 201808.01 onwards, DNAscope allows you to use a model to perform variant calling with higher accuracy by improving the candidate detection and filtering.

Sentieon® can provide you with a model trained using a subset of the data from the GiAB truth-set found in https://github.com/genome-in-a-bottle. The model was created by processing samples HG001 and HG005 through a pipeline consisting of Sentieon® BWA-mem alignment and Sentieon® deduplication, and using the variant calling results to calibrate a model to fit the truth-set.

3.2 Using a machine learning model with DNAscope

Two individual commands are run to call variants and to apply the machine learning model. The input BAM file should come from a pipeline where only alignment and deduplication have been performed, to match the model creation methodology.

```
PCRFREE=true #PCRFREE=true means the sample is PCRFree, change it to false for PCR samples.
if [ "$PCRFREE" = true ] ; then
    sentieon driver -t NUMBER_THREADS -r REFERENCE -i DEDUPED_BAM \
    --algo DNAscope [ -d dbSNP ] --pcr_indel_model none --model ML_MODEL TMP_VARIANT_VCF
else
    sentieon driver -t NUMBER_THREADS -r REFERENCE -i DEDUPED_BAM \
    --algo DNAscope [ -d dbSNP ] --model ML_MODEL TMP_VARIANT_VCF
fi
sentieon driver -t NUMBER_THREADS -r REFERENCE --algo DNAModelApply \
  --model ML_MODEL -v TMP_VARIANT_VCF VARIANT_VCF
```

Reminder

It is important to add option `--pcr_indel_model NONE` when running `DNAscope` if the data you are using is PCR Free.

Depending on whether PCR is involved, `DNAscope` uses different priors for finding significant INDEL variants, which could be controlled by the `--pcr_indel_model` option. The default `--pcr_indel_model` setting is for PCR samples. Thus it is important to set `--pcr_indel_model none` for PCR Free samples.

The following inputs are required for the command:

- **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.
- **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.
- **DEDUPED_BAM**: the location of the input BAM file.
- **TMP_VARIANT_VCF**: the location and filename of the variant calling output of DNAscope. This is a temporary file.
- **VARIANT_VCF**: the location and filename of the variant calling output. A corresponding index file will be created. The tool will output a compressed file by using .gz extension.
- **ML_MODEL**: the location of the machine learning model file. In the DNAscope command the model will be used to determine the settings used in variant calling.

The following inputs are optional for the command:

- **dbSNP**: the location of the Single Nucleotide Polymorphism database (dbSNP) that will be used to label known variants. You can only use one dbSNP file.

3.3 Using DNAscope to produce GVCF output files

From version 202112.04 onwards, DNAscope allows you to use a model to produce variant calls in the Genomic VCF (GVCF) format. The GVCF format contains additional information on sites that are homozygous for the reference allele in the sample being processed. Recently trained DNAscope models are required and using a DNAscope model trained with Sentieon version 202112.01 or earlier will result in an error.

Two individual commands are run to call variants and to apply the machine learning model. The input BAM file should come from a pipeline where only alignment and deduplication have been performed, to match the model creation methodology.

```
PCRFREE=true #PCRFREE=true means the sample is PCRFree, change it to false for PCR samples.
if [ "$PCRFREE" = true ] ; then
    sentieon driver -t NUMBER_THREADS -r REFERENCE -i DEDUPED_BAM \
    --algo DNAscope [ -d dbSNP ] --pcr_indel_model none --model ML_MODEL \
    --emit_mode gvcf TMP_VARIANT_GVCF
else
    sentieon driver -t NUMBER_THREADS -r REFERENCE -i DEDUPED_BAM \
    --algo DNAscope [ -d dbSNP ] --model ML_MODEL --emit_mode gvcf TMP_VARIANT_GVCF
fi
sentieon driver -t NUMBER_THREADS -r REFERENCE --algo DNAModelApply \
  --model ML_MODEL -v TMP_VARIANT_GVCF VARIANT_GVCF
```

Reminder

It is important to add option `--pcr_indel_model NONE` when running `DNAscope` if the data you are using is PCR Free.

Depending on whether PCR is involved, `DNAscope` uses different priors for finding significant INDEL variants, which could be controlled by the `--pcr_indel_model` option. The default `--pcr_indel_model` setting is for PCR samples. Thus it is important to set `--pcr_indel_model none` for PCR Free samples.

The following inputs are required for the command:

- **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.
- **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.
- **DEDUPED_BAM**: the location of the input BAM file.
- **TMP_VARIANT_GVCF**: the location and filename of the GVCF output of DNAscope. This is a temporary file.
- **VARIANT_GVCF**: the location and filename of the GVCF output. A corresponding index file will be created. The tool will output a compressed file by using .gz extension.
- **ML_MODEL**: the location of the machine learning model file. In the DNAscope command the model will be used to determine the settings used in variant calling.

The following inputs are optional for the command:

- **dbSNP**: the location of the Single Nucleotide Polymorphism database (dbSNP) that will be used to label known variants. You can only use one dbSNP file.

The GVCF output file can be genotyped either individually, or jointly with GVCFs from other samples using the GVCFtyper algo in Sentieon version 202112.06 and later and will output a single sample or multi-sample VCF.

```
sentieon driver -r REFERENCE --algo GVCFtyper \
  -v s1_VARIANT_GVCF -v s2_VARIANT_GVCF -v s3_VARIANT_GVCF VARIANT_VCF
```

Please check the Sentieon manual for additional details about the GVCFtyper algo, https://support.sentieon.com/manual/usages/general/#gvcftyper-algorithm.

Reminder

GVCFtyper can be used genotype DNAscope GVCFs from multiple sequencing platforms into a single multi-sample VCF.

GVCFtyper does not support joint genotyping of DNAscope GVCFs mixed with GVCFs produced by DNAscope without a machine learning model or DNAscope GVCFs mixed with GVCFs produced by other tools.

3.4 Limitations of the machine learning model

When using DNAscope with a machine learning model, most of the options for DNAscope should not be used, even though the software will not give an error if the options are present. In particular, the following option should be avoided:

- `--var_type BND`: The use of DNAscope with a machine learning model is incompatible with structural variant calling.

In addition, using an input BAM file created using a pipeline with additional stages such as INDEL realignment or BQSR will likely degrade the performance, as the impact of those stages was fitted into the model.