TNscope Somatic variant discovery with a matched normal sample using a machine learning model

Release 202112.06

Sentieon, Inc

Nov 07, 2022

Contents

1 Introduction

This documents describes the capabilities of TNscope® for somatic variant 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 Using a machine learning model in TNscope®

2.1 Goal of a machine learning model in TNscope®

TNscope® allows you to use a machine learning model to perform variant filtration to improve the accuracy of the results. The machine learning model methodology is described in https://www.biorxiv.org/content/early/2018/01/19/250647 and uses a set of sensitive settings in TNscope® to detect a higher number of variant candidates, followed by a model based variant filtration.

Sentieon® can provide you with a model trained using a set of mixture samples created using the GiAB truth-set found in https://github.com/genome-in-a-bottle.

2.2 Using a machine learning model with TNscope®

Three individual commands are run to call variants using high sensitive settings, to apply the machine learning model and to use BCFtools to set the model threshold. The input BAM file should come from a pipeline where only alignment, deduplication and BQSR have been performed, to match the model creation methodology.

```
sentieon driver -t NUMBER_THREADS -r REFERENCE \
  -i TUMOR_DEDUPED_BAM -q TUMOR_RECAL_DATA.TABLE \
  -i NORMAL_DEDUPED_BAM -q NORMAL_RECAL_DATA.TABLE \
  --algo TNscope --tumor_sample TUMOR --normal_sample NORMAL \
  --clip_by_minbq 1 --max_error_per_read 3 --disable_detector sv \
  --min_init_tumor_lod 2.0 --min_base_qual 10 --min_base_qual_asm 10 \
  --min_tumor_allele_frac 0.00005 TMP_VARIANT_VCF
sentieon driver -t NUMBER_THREADS -r REFERENCE --algo TNModelApply \
  --model ML_MODEL -v TMP_VARIANT_VCF VARIANT_VCF
bcftools filter -s "ML_FAIL" -i "INFO/ML_PROB > $ML_THRESHOLD" VARIANT_VCF \
  -O z -m x -o FILTER_VARIANT_VCF
```

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.
- **TUMOR_DEDUPED_BAM**: the location of the pre-processed BAM file after deduplication for the TUMOR sample.
- **TUMOR_RECAL_DATA.TABLE**: the location where the BQSR stage for the TUMOR sample stored the result.
- **NORMAL_DEDUPED_BAM**: the location of the pre-processed BAM file after deduplication for the NORMAL sample.
- **NORMAL_RECAL_DATA.TABLE**: the location where the BQSR stage for the NORMAL sample stored the result.
- **TUMOR**: name of the SM tag in the BAM file for the tumor sample.
- **NORMAL**: name of the SM tag in the BAM file for the normal sample.
- **TMP_VARIANT_VCF**: the location and filename of the variant calling output of TNscope®. 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.
- **FILTER_VARIANT_VCF**: the location and filename of the variant calling output after setting the final threshold for filtering. Due to using option *-O z* the output file will be a bgzip compressed vcf.gz file.
- **ML_MODEL**: the location of the machine learning model file.
- **$ML_THRESHOLD**: the threshold for the probability that a variant is true according to the model. A recommended number is 0.81.