Somatic Variant Calling for SNPs and INDELs

Release 202112.06

Sentieon, Inc

Nov 07, 2022

Contents

1 Introduction

This document describes somatic variant calling pipelines using TNscope® and TNseq® . TNscope® uses an improved variant calling algorithm to obtain higher accuracy and improved runtimes, while TNseq® matches the GATK's Mutect2 somatic variant calling with substantially improved runtime and parallelization.

For complete somatic variant calling pipelines, please visit our github page: https://github.com/Sentieon/sentieon-scripts/tree/master/example_pipelines/somatic.

2 Data processing without unique molecular identifiers (UMIs)

2.1 Step1. Alignment

```
# ****************************************
# 1a. Mapping reads with BWA-MEM, sorting for the tumor sample
# ****************************************
( sentieon bwa mem -R "@RG\tID:$tumor\tSM:$tumor\tPL:$platform" \
  -t $nt -K 10000000 $fasta $tumor_fastq_1 $tumor_fastq_2 || \
  echo -n 'error' ) | \
  sentieon util sort -o tumor_sorted.bam -t $nt --sam2bam -i -

# ****************************************
# 1b. Mapping reads with BWA-MEM, sorting for the normal sample
# ****************************************
( sentieon bwa mem -R "@RG\tID:$normal\tSM:$normal\tPL:$platform" \
  -t $nt -K 10000000 $fasta $normal_fastq_1 $normal_fastq_2 || \
  echo -n 'error' ) | \
  sentieon util sort -o normal_sorted.bam -t $nt --sam2bam -i -
```

2.2 Step2. PCR duplicate removal (skip for targeted amplicon sequencing)

```
# ****************************************
# 2a. Remove duplicate reads from the tumor sample
# ****************************************
sentieon driver -t $nt -i tumor_sorted.bam \
  --algo LocusCollector \
  --fun score_info \
  tumor_score.txt
sentieon driver -t $nt -i tumor_sorted.bam \
  --algo Dedup \
  --score_info tumor_score.txt \
  --metrics tumor_dedup_metrics.txt \
  tumor_deduped.bam

# ****************************************
# 2b. Remove duplicate reads from the normal sample
# ****************************************
sentieon driver -t $nt -i normal_sorted.bam \
  --algo LocusCollector \
  --fun score_info \
  normal_score.txt
sentieon driver -t $nt -i normal_sorted.bam \
  --algo Dedup \
  --score_info normal_score.txt \
  --metrics normal_dedup_metrics.txt \
  normal_deduped.bam
```

2.3 Step3. Base quality score recalibration (skip for panels)

```
# ****************************************
# 3a. Base recalibration for the tumor sample
```

(continues on next page)

(continued from previous page)

```
# ****************************************
sentieon driver -r $fasta -t $nt -i tumor_deduped.bam --interval $BED \
  --algo QualCal \
  -k $dbsnp \
  -k $known_Mills_indels \
  -k $known_1000G_indels \
  tumor_recal_data.table

# ****************************************
# 3b. Base recalibration for the normal sample
# ****************************************
sentieon driver -r $fasta -t $nt -i normal_deduped.bam --interval $BED \
  --algo QualCal \
  -k $dbsnp \
  -k $known_Mills_indels \
  -k $known_1000G_indels \
  normal_recal_data.table
```

2.4 Step4. Variant calling and filtration

Generation of the final variant callset is divided into two commands for variant calling and filtration. Depending on sample and library types, different options and filters should be used.

Please find the python script `tnscope_filter.py` for filtering at Sentieon's GitHub page at https://github.com/Sentieon/sentieon-scripts.

Sample type 1 - whole genome/exome sequencing

This section describes the pipeline for WGS/WES using TNseq®. For more information about TNseq®, please visit our manual https://support.sentieon.com/manual/TNseq_usage/tnseq/.

```
sentieon driver -r $fasta -t $nt -i tumor_deduped.bam -i normal_deduped.bam \
  [--interval $BED] \
  --algo TNhaplotyper2 \
  --tumor_sample $TUMOR_SM \
  --normal_sample $NORMAL_SM \
  [--pon $PON] \
  --germline_vcf $GERMLINE_VCF \
  output-tnhap2-tmp.vcf.gz \
  --algo OrientationBias \
  --tumor_sample $TUMOR_SM \
  output-orientation \
  --algo ContaminationModel \
  --tumor_sample $TUMOR_SM \
  --normal_sample $NORMAL_SM \
  --vcf $CONTAMINATION_VCF \
  --tumor_segments output-contamination-segments \
  output-contamination

sentieon driver -r $fasta \
  --algo TNfilter \
  -v output-tnhap2-tmp.vcf.gz \
  --tumor_sample $TUMOR_SM \
  --normal_sample $NORMAL_SM \
```

(continues on next page)

(continued from previous page)

```
  [--contamination output-contamination] \
  [--tumor_segments output-contamination-segments] \
  [--orientation_priors output-orientation] \
  output-tnhap2.vcf.gz
```

Sample type 2 - Target panel (hybridization capture)

This section describes TNscope® parameters for somatic variant calling from targeted panel sequencing (hybridization capture, 200-5000x depth, AF >= 1%). The thresholds for *–min_tumor_af* and *–min_depth* should be adjusted as needed.

```
sentieon driver -r $fasta -t $nt -i tumor_deduped.bam -i normal_deduped.bam \
  --interval $BED \
  --algo TNscope \
  --tumor_sample $TUMOR_SM \
  --normal_sample $NORMAL_SM \
  --dbsnp $dbsnp \
  --min_tumor_allele_frac 0.009 \
  --max_fisher_pv_active 0.05 \
  --max_normal_alt_cnt 10 \
  --max_normal_alt_frac 0.01 \
  --max_normal_alt_qsum 250 \
  --sv_mask_ext 10 \
  --prune_factor 0 \
  --assemble_mode 2 \
  [--pon panel_of_normal.vcf] \
  output_tnscope.pre_filter.vcf.gz

sentieon pyexec tnscope_filter.py \
    -v output_tnscope.pre_filter.vcf.gz \
    --tumor_sample $TUMOR_SM \
    -x tissue_panel --min_tumor_af 0.0095 --min_depth 100 \
    output_tnscope.filter.vcf.gz
```

Sample type 3 - Target panel (amplicon)

This section describes TNscope® parameters for somatic variant calling from targeted panel sequencing (amplicon, 200-5000x depth, AF >= 1%). The thresholds for *–min_tumor_af* and *–min_depth* should be adjusted as needed.

```
sentieon driver -r $fasta -t $nt -i tumor_deduped.bam -i normal_deduped.bam \
  --interval $BED --interval_padding 10 \
  --algo TNscope \
  --tumor_sample $TUMOR_SM \
  --normal_sample $NORMAL_SM \
  --dbsnp $dbsnp \
  --min_tumor_allele_frac 0.009 \
  --max_fisher_pv_active 0.05 \
  --max_normal_alt_cnt 10 \
  --max_normal_alt_frac 0.01 \
  --max_normal_alt_qsum 250 \
  --sv_mask_ext 10 \
  --prune_factor 0 \
  --assemble_mode 2 \
```

(continues on next page)

(continued from previous page)

```
  [--pon panel_of_normal.vcf ] \
  output_tnscope.pre_filter.vcf.gz

sentieon pyexec tnscope_filter.py \
    -v output_tnscope.pre_filter.vcf.gz \
    --tumor_sample $TUMOR_SM \
    -x amplicon --min_tumor_af 0.0095 --min_depth 200 \
    output_tnscope.filter.vcf.gz
```

Sample type 4 - ctDNA (tumor-only without UMI)

This section describes TNscope® parameters for ctDNA and other high depth cases (6000x-10000x depth, AF > 0.5%). The thresholds for *–min_tumor_af* and *–min_depth* should be adjusted as needed.

```
sentieon driver -r $fasta -t $nt -i tumor_deduped.bam --interval $BED \
  --algo TNscope \
  --tumor_sample $TUMOR_SM \
  --dbsnp $dbsnp \
  --disable_detector sv \
  --min_tumor_allele_frac 3e-3 \
  --min_tumor_lod 3.0 \
  --min_init_tumor_lod 1.0 \
  --assemble_mode 2 \
  [--pon panel_of_normal.vcf ] \
  output_tnscope.pre_filter.vcf.gz

sentieon pyexec tnscope_filter.py \
    -v output_tnscope.pre_filter.vcf.gz \
    --tumor_sample $TUMOR_SM \
    -x ctdna --min_tumor_af 0.005 --min_depth 400 \
    output_tnscope.filter.vcf.gz
```

3 UMI data processing with Sentieon® UMI and TNscope®

This section describes TNscope® parameters for ctDNA and other high depth samples with UMI-tagged reads (20000x-50000x depth, AF > 0.1%). The thresholds for *–min_tumor_af* and *–min_depth* should be adjusted as needed.

Please see our Application Note on Unique Molecular Identifiers for more details: https://support.sentieon.com/appnotes/umi/

Please find the python script `tnscope_filter.py` for filtering at Sentieon's GitHub page at https://github.com/Sentieon/sentieon-scripts.

3.1 Step1. Alignment and consensus generation

```
if [ "$DUPLEX_UMI" = "true" ] ; then
      READ_STRUCTURE="-d $READ_STRUCTURE"
fi
sentieon umi extract $READ_STRUCTURE $fastq_1 $fastq_2 | \
  sentieon bwa mem -p -C -R "@RG\tID:$tumor\tSM:$tumor\tPL:$platform" -t $nt \
  -K 10000000 $fasta - | \
```

(continues on next page)

(continued from previous page)

```
  sentieon umi consensus -o umi_consensus.fastq.gz
sentieon bwa mem -p -C -R "@RG\tID:$tumor\tSM:$tumor\tPL:$platform" -t $nt \
  -K 10000000 $fasta umi_consensus.fastq.gz | \
  sentieon util sort --umi_post_process --sam2bam -i - -o umi_consensus.bam
```

3.2 Step2. Variant calling with UMI consensus reads

```
sentieon driver -r $fasta -t $nt -i umi_consensus.bam --interval $BED \
  --algo TNscope \
  --tumor_sample $TUMOR_SM \
  --dbsnp $dbsnp \
  --min_tumor_allele_frac 0.001 \
  --min_tumor_lod 3.0 \
  --min_init_tumor_lod 3.0 \
  --pcr_indel_model NONE \
  --min_base_qual 40 \
  --resample_depth 100000 \
  --assemble_mode 4 \
  output_tnscope.pre_filter.vcf.gz

sentieon pyexec tnscope_filter.py \
    -v output_tnscope.pre_filter.vcf.gz \
    --tumor_sample $TUMOR_SM \
    -x ctdna_umi --min_tumor_af 0.001 --min_depth 1000 \
    output_tnscope.filtered.vcf.gz
```

4 Appendix

4.1 Description of optional command-line arguments

TNscope:

- –assemble_mode: assemble_mode=2 achieves a good balance between accuracy and speed. assemble_mode=4 maybe be used to handle more complex regions.
- –sv_mask_ext: Prevent SNP/Indel detection in the proximity of a SV break point.
- –max_fisher_pv_active: Turn on and set a threshold for the PV value filter, which measures the statistical difference between tumor and normal sample at a particular site.
- –min_tumor_allele_frac: Set the minimum tumor AF to be considered as potential variant site.
- –min_init_tumor_lod: Minimum tumor log odds ratio in initial variant calling pass.
- –min_tumor_lod: Minimum tumor log odds ratio to filter in the output VCF with the TLOD annotation.
- –prune_factor: Use prune_factor=0 to turn on the adaptive graph pruning algorithm.

4.2 FAQ

1. **The AD ratio does not match the reported AF in the output VCF:**
    - They are calculated and defined differently, for legacy reasons from GATK
    - AD: based on the pile-up of the input bam, before local assembly is done
    - AF: the ratio of the number of supporting reads for REF and ALT alleles, after local assembly is done
    - Filtered differently: Reads go through slightly different set of filters before the calculation
2. **Using the Panel of Normal file:**

- Panel of Normal can be used with or without matching normal samples
- When a variant is found in the PoN, it is flagged with the `panel_of_normal` filter
- The Panel of Normal is helpful when matching normal is not available
- Follow this link for instructions on generating your own Panel of Normal VCF file with TNscope: https://support.sentieon.com/manual/TNscope_usage/tnscope/#generating-a-panel-of-normal-vcf-file

3. **Role of COSMIC and dbSNP database file:**
    - They only affect Tumor-Normal cases, and only impact the filtering
    - Both COSMIC and dbSNP only determine the threshold of NOLD for `germline_risk`
4. **Complex variants:**
    - Sentieon provides an experimental python script for merging neighboring variants in the same phase. This script is only experimental, and is available at https://github.com/Sentieon/sentieon-scripts/blob/master/merge_mnp/merge_mnp.py.

5 References

- Freed, Donald, Renke Pan, and Rafael Aldana. "TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering." bioRxiv (2018): 250647. Link[1]
- Pei, Surui, et al. "Benchmarking variant callers in next-generation and third-generation sequencing analysis." Briefings in Bioinformatics (2020). Link[2]

©Sentieon Inc.
160 E Tasman Dr STE 208, San Jose, CA 95134-1619
www.sentieon.com

[1] https://www.biorxiv.org/content/10.1101/250647v1.full
[2] https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa148/5875142