Functional Equivalent Pipeline from CCDG using Sentieon

Release 202308.03

Sentieon, Inc

Oct 23, 2024

Contents

1 Introduction

This documents describes how to implement the "Functional Equivalent Pipeline", also known as the CCDG pipeline standard, described in https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md and published in https://www.nature.com/articles/s41467-018-06159-4 using the Sentieon® tools. In order to match the version requirements of this pipeline, you should use the Sentieon® tools with version 201704 or higher. Starting with Sentieon® tools version 201911, the Sentieon® BWA was updated to version 0.7.17; BWA version 0.7.17 produces MC MateTags in its output, and samblaster addMateTags will not remove this MC tag and add its own MC tag to the BAM file, creating a duplicated MC tag.

2 Command line equivalence

2.1 Alignment

The alignment stage in the CCDG functional equivalent pipeline is done using BWA-MEM version 0.7.15:

```
FASTA=Homo_sapiens_assembly38.fasta
NT=$(nproc)
bwa mem -R "@RG\tID:$RGID\tSM:$SM\tPL:$PL" -t $NT -K 100000000 -Y $FASTA $FASTQ1 $FASTQ2 | \
samblaster --addMateTags -a | \
samtools view -Sbhu - | \
sambamba sort -n -t $nt --tmpdir tmp -o sorted.bam /dev/stdin
```

To run the equivalent command with Sentieon®:

```
sentieon bwa mem -R "@RG\tID:$RGID\tSM:$SM\tPL:$PL" -t $NT -K 100000000 -Y $FASTA $FASTQ1 $FASTQ2 | \
samblaster --addMateTags -a | \
util sort --sam2bam -i - -r $FASTA -t $nt -o sorted.bam
```

To run the equivalent command when using Sentieon® version 201911 or higher:

```
sentieon bwa mem -R "@RG\tID:$RGID\tSM:$SM\tPL:$PL" -t $NT -K 100000000 -Y $FASTA $FASTQ1 $FASTQ2 | \
sed 's|MC:Z:[^\t]*\t||' | \
samblaster --addMateTags -a | \
util sort --sam2bam -i - -r $FASTA -t $nt -o sorted.bam
```

2.2 Duplicate marking

The deduplication stage in the CCDG functional equivalent pipeline is done using Picard version 2.4 or higher:

```
java -Xmx48g -jar $picard MarkDuplicates I=sorted.bam METRICS_FILE=markdup_metrics.txt \
   ASSUME_SORT_ORDER=queryname QUIET=true COMPRESSION_LEVEL=0 O=/dev/stdout | \
sambamba sort -t $NT --tmpdir tmp -o markduped.bam /dev/stdin
```

To run the equivalent command with Sentieon®:

```
sentieon driver -t $nt -r $FASTA -i sorted.bam --algo LocusCollector --fun score_info tmp_score.gz && \
sentieon driver -t $nt -r $FASTA -i sorted.bam --algo Dedup --score_info tmp_score.gz \
   --output_dup_read_name --metrics dedup_metrics.txt tmp_dup_qname.txt.gz && \
sentieon driver -t $nt -r $FASTA -i sorted.bam --algo Dedup --dup_read_name tmp_dup_qname.txt.gz markduped.bam
```

The Sentieon® command uses a special 3-pass Deduplication flow to mark both primary and non-primary reads.

2.3 Base quality score recalibration with binning scheme

The BQSR stage in the CCDG functional equivalent pipeline is done using GATK3 or GATK4:

```
INTERVAL_ARG="-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 -L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
   -L chr11 -L chr12 -L chr13 -L chr14 -L chr15 -L chr16 -L chr17 -L chr18 -L chr19 -L chr20 -L chr21 -L chr22"
DOWNSAMPLE_ARG="--downsample_to_fraction .1"
KNOWN_MILLS_INDELS="Mills_and_1000G_gold_standard.indels.hg38.vcf.gz"
KNOWN_1000G_INDELS="Homo_sapiens_assembly38.known_indels.vcf.gz"
KNOWN_DBSNP="Homo_sapiens_assembly38.dbsnp138.vcf"
java -Xmx48g -jar $GATK_37 -T BaseRecalibrator -R $FASTA -I markduped.bam $DOWNSAMPLE_ARG $INTERVAL_ARG \
   -knownSites $KNOWN_MILLS_INDELS -knownSites $KNOWN_1000G_INDELS -knownSites $KNOWN_DBSNP \
   -o recal_data_37.table && \
java -Xmx15g -jar $GATK_37 -T PrintReads -R $fasta -I markduped.bam --BQSR recal_data_37.table -o recaled_37.bam \
   --globalQScorePrior -1.0 --preserve_qscores_less_than 6 --static_quantized_quals 10 \
   --static_quantized_quals 20 --static_quantized_quals 30 --disable_indel_quals && \
samtools view -C -T $fasta -@ 2 -o recaled_37.cram recaled_37.bam && \
samtools index -c recaled_37.cram recaled_37.cram.index
```

To run the equivalent command with Sentieon®:

```
INTERVAL_ARG="--interval chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,chr11,\
chr12,chr13,chr14,chr15,chr16,chr17,chr18,chr19,chr20,chr21,chr22"
sentieon driver -t $NT -r $FASTA --interval $INTERVAL_ARG -i markduped.bam --algo QualCal -k $KNOWN_MILLS_INDELS \
   -k $KNOWN_1000G_INDELS -k $KNOWN_DBSNP recal_data_Sentieon.table && \
sentieon driver -t $NT -r $FASTA -i markduped.bam \
   --read_filter QualCalFilter,table=recal_data_Sentieon.table,prior=-1.0,indel=false,levels=10/20/30,min_qual=6 \
   --algo ReadWriter recaled_RW.cram
```

Bear in mind that Sentieon® does not do any downsampling, as the Sentieon® tools are efficient enough that they are able to handle all the depth in your sequencing. In addition, this flow is differernt from the normal best practices flow to implement the special binning required in the CCDG functional equivalent pipeline.

3 Pipeline script using Sentieon®

The following script will perform the CCDG functional equivalent pipeline on you input FASTQs using Sentieon®:

```
#!/bin/sh
# ********************************************************************************
# Script to perform DNA seq variant calling using Sentieon following
# the functional equivalent pipeline described in
# https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md
# ********************************************************************************

# Update with the fullpath location of your sample fastq
SM="sample" #sample name
RGID="rg_$SM" #read group ID
PL="ILLUMINA" #or other sequencing platform
FASTQ_1="${SAMPLE}_r1.fastq.gz"
FASTQ_2="${SAMPLE}_r2.fastq.gz" #if using 2 FASTQ inputs

# Update with the location of the reference data files
FASTA_DIR="/home/regression/references/hg38bundle"
FASTA="$FASTA_DIR/Homo_sapiens_assembly38.fasta"
KNOWN_DBSNP="$FASTA_DIR/Homo_sapiens_assembly38.dbsnp138.vcf.gz"
KNOWN_INDELS="$FASTA_DIR/Homo_sapiens_assembly38.known_indels.vcf.gz"
KNOWN_MILLS="$FASTA_DIR/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz"

# Update with the location of the Sentieon software package and license file
SENTIEON_INSTALL_DIR=/home/release/sentieon-genomics-|release_version|
export SENTIEON_LICENSE=/home/Licenses/Sentieon.lic #or using licsrvr: c1n11.sentieon.com:5443

# Other settings
NT=$(nproc) #number of threads to use in computation
SAMBLASTER=/home/release/other_tools/samblaster-0.1.23/samblaster
START_DIR=$PWD

# ****************************************
# 0. Setup
# ****************************************
workdir="$START_DIR/${SM}" #Determines where the output files will be stored
mkdir -p $workdir
logfile=$workdir/run.log
exec >>$logfile 2>&1
cd $workdir
```

(continues on next page)

(continued from previous page)

```
# ****************************************
# main pipeline with Sentieon
# ****************************************
# 1. Mapping BWA-MEM 0.7.15 util sort
SENTIEON_VERSION=$($SENTIEON_INSTALL_DIR/bin/sentieon driver --version)
if (( $(echo "${SENTIEON_VERSION##*-} < 201911" |bc -l) )); then
        $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -R "@RG\tID:$RGID\tSM:$SM\tPL:$PL" -t $NT \
           -K 100000000 -Y $FASTA $FASTQ_1 $FASTQ_2 | \
        $SAMBLASTER --addMateTags -a | \
        $SENTIEON_INSTALL_DIR/bin/sentieon util sort -r $FASTA -o sorted.bam -t $NT --sam2bam -i -
else
        #Sentieon 201911 and higher use BWA 0.7.17, which already produce MC tags in the output
        $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -R "@RG\tID:$RGID\tSM:$SM\tPL:$PL" -t $NT \
           -K 100000000 -Y $FASTA $FASTQ_1 $FASTQ_2 | \
        $SENTIEON_INSTALL_DIR/bin/sentieon util sort -r $FASTA -o sorted.bam -t $NT --sam2bam -i -
fi

# 2. Mark Duplicates with Sentieon
$SENTIEON_INSTALL_DIR/bin/sentieon driver -t $NT -i sorted.bam --algo LocusCollector --fun score_info score.txt
$SENTIEON_INSTALL_DIR/bin/sentieon driver -t $NT -i sorted.bam --algo Dedup --score_info score.txt \
   --metrics mark_dup_metrics.txt --output_dup_read_name tmp_dup_qname.txt
$SENTIEON_INSTALL_DIR/bin/sentieon driver -t $NT -i sorted.bam --algo Dedup \
   --dup_read_name tmp_dup_qname.txt markduped.bam

# 3. Base Quality Score Recalibration with Sentieon
interval_arg="--interval chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,chr11,\
chr12,chr13,chr14,chr15,chr16,chr17,chr18,chr19,chr20,chr21,chr22"
$SENTIEON_INSTALL_DIR/bin/sentieon driver $interval_arg -r $FASTA -t $NT -i markduped.bam \
   --algo QualCal -k $KNOWN_MILLS -k $KNOWN_INDELS -k $KNOWN_DBSNP recal_data.table
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $FASTA -t $NT -i markduped.bam \
   --read_filter QualCalFilter,table=recal_data.table,prior=-1.0,indel=false,levels=10/20/30,min_qual=6 \
   --algo ReadWriter recaled_RW.cram

# 4. Haplotyper with Sentieon
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $FASTA -t $NT -i recaled_RW.cram --algo Haplotyper Haplotyper.vcf.gz
```