Sentieon DNA-Seq call SNP

> 本流程由华中农业大学信息学院杨庆勇课题组提供和维护。

1.介绍

- 如对下面的参数有什么不懂的地方，可以参考Sentieon手册。

2.运行示例

1. 方式一

```
# 若有VCF database,可参考该流程
bsub -J sen-test -n 16 -R span[hosts=1] -o %J.out -e %J.err -q normal "bash sentieon_quickstart0.sh"

# 利用50x拟南芥NGS数据
bsub -J sen-test -n 16 -R span[hosts=1] -o %J.out -e %J.err -q normal "bash sentieon_ath50x.sh"
```

2. 方式二
    - 该脚本mapping结果文件全部保存为CRAM文件，对于CRAM格式的介绍请参考：cram的介绍和cram和bam的对比。CRAM能极度的节约空间，适合长期保存比对文件，samtools的作者Li Heng对CRAM格式的的评价。
    - 中间生成的mapq过滤的比对文件、dedup的比对文件，本流程都进行了删除处理，如有特殊需要，请更改相应的脚本。

```
# 利用50x拟南芥NGS数据
bsub < sentieon_ath50x.lsf

# 若有VCF database,可参考该流程
bsub < sentieon_Test_People
```

3.脚本内容

1. sentieon_ath50x.sh

```
#!/bin/sh
# *******************************************
*******************************************

# Update with the fullpath location of your sample fastq
set -x
data_dir="$( cd -P "$( dirname "$0" )" && pwd )"  #workdir
fastq_1=/public/exercise/sentieon/tair10_1.fastq.gz
fastq_2=/public/exercise/sentieon/tair10_2.fastq.gz

# Update with the location of the reference data files
fasta=/public/exercise/sentieon/reference_Tair10/Arabidopsis_thaliana.TAIR10.dna.toplevel.modified.fa

# Set SENTIEON_LICENSE if it is not set in the environment
module load SAMtools/1.9
```

```
#module load sentieon/201808.07
 export SENTIEON_LICENSE=mn01:9000

# Update with the location of the Sentieon software package
SENTIEON_INSTALL_DIR=/public/home/software/opt/bio/software/Sentieon/201808.07

# It is important to assign meaningful names in actual cases.
# It is particularly important to assign different read group names.
sample="tair10"
group="G"
platform="ILLUMINA"

# Other settings
nt=16 #number of threads to use in computation

# ******************************************
# 0. Setup
# ******************************************
workdir=$data_dir/result-tair10
mkdir -p $workdir
logfile=$workdir/run.log
exec >$logfile 2>&1
cd $workdir

#Sentieon proprietary compression
bam_option="--bam_compression 1"

# ******************************************
# 1. Mapping reads with BWA-MEM, sorting
# ******************************************
#The results of this call are dependent on the number of threads used. To have number of threads independent results, add chunk size option -K 10000000

# speed up memory allocation malloc in bwa
export LD_PRELOAD=$SENTIEON_INSTALL_DIR/lib/libjemalloc.so
export MALLOC_CONF=lg_dirty_mult:-1

( $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -M -R "@RG\tID:$group\tSM:$sample\tPL:$platform" -t $nt -K 10000000 $fasta $fastq_1 $fastq_2 || echo -n 'error' ) | $SENTIEON_INSTALL_DIR/bin/sentieon util sort $bam_option -r $fasta -o sorted.bam -t $nt --sam2bam -i -

# ******************************************
# 2. Metrics
# ******************************************
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i sorted.bam --algo MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat --adapter_seq '' aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot GCBias -o gc-report.pdf gc_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot QualDistribution -o qd-report.pdf qd_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot MeanQualityByCycle -o mq-report.pdf mq_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot InsertSizeMetricAlgo -o is-report.pdf is_metrics.txt

# ******************************************
# 3. Remove Duplicate Reads
# To mark duplicate reads only without removing them, remove "--rmdup" in the second command
# ******************************************
$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 --rmdup --score_info score.txt --metrics dedup_metrics.txt $bam_option deduped.bam

# ******************************************

# ******************************************
# 5. Base recalibration
# ******************************************

# Perform recalibration
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam --algo QualCal recal_data.table

# Perform post-calibration check (optional)
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo QualCal recal_data.table.post
$SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt --algo QualCal --plot --before recal_data.table --after recal_data.table.post recal.csv
$SENTIEON_INSTALL_DIR/bin/sentieon plot QualCal -o recal_plots.pdf recal.csv


# ******************************************
# 6. HC Variant caller
# Note: Sentieon default setting matches versions before GATK 3.7.
# Starting GATK v3.7, the default settings have been updated multiple times.
# Below shows commands to match GATK v3.7 - 4.1
# Please change according to your desired behavior.
# ******************************************

# Matching GATK 3.7, 3.8, 4.0
#$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo Haplotyper  --emit_conf=10 --call_conf=10 output-hc.vcf.gz

# Matching GATK 4.1
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo Haplotyper  --genotype_model multinomial --emit_conf 30 --call_conf 30 output-hc.vcf.gz

```

2. sentieon_quickstart0.sh

```
#!/bin/sh
# *******************************************
# Script to perform DNA seq variant calling
# using a single sample with fastq files
# named 1.fastq.gz and 2.fastq.gz
# *******************************************

# Update with the fullpath location of your sample fastq
set -x
data_dir="$( cd -P "$( dirname "$0" )" && pwd )"  #workdir
data_dir2=/public/exercise/sentieon                      #datadir
fastq_1=$data_dir2/1.fastq.gz
fastq_2=$data_dir2/2.fastq.gz #If using Illumina paired data

# Update with the location of the reference data files
fasta=$data_dir2/reference/ucsc.hg19_chr22.fasta
dbsnp=$data_dir2/reference/dbsnp_135.hg19_chr22.vcf
known_1000G_indels=$data_dir2/reference/1000G_phase1.snps.high_confidence.hg19_chr22.sites.vcf
known_Mills_indels=$data_dir2/reference/Mills_and_1000G_gold_standard.indels.hg19_chr22.sites.vcf

```

```
# Set SENTIEON_LICENSE if it is not set in the environment
module load SAMtools/1.9
#module load sentieon/201808.07
 export SENTIEON_LICENSE=mn01:9000

# Update with the location of the Sentieon software package
SENTIEON_INSTALL_DIR=/public/home/software/opt/bio/software/Sentieon/201808.07

# It is important to assign meaningful names in actual cases.
# It is particularly important to assign different read group names.
sample="sample_name"
group="read_group_name"
platform="ILLUMINA"

# Other settings
nt=16 #number of threads to use in computation

# ******************************************
# 0. Setup
# ******************************************
workdir=$data_dir/result
mkdir -p $workdir
logfile=$workdir/run.log
exec >$logfile 2>&1
cd $workdir

#Sentieon proprietary compression
bam_option="--bam_compression 1"

# ******************************************
# 1. Mapping reads with BWA-MEM, sorting
# ******************************************
#The results of this call are dependent on the number of threads used. To have number of threads independent results, add chunk size option -K 10000000

# speed up memory allocation malloc in bwa
export LD_PRELOAD=$SENTIEON_INSTALL_DIR/lib/libjemalloc.so
export MALLOC_CONF=lg_dirty_mult:-1

( $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -M -R "@RG\tID:$group\tSM:$sample\tPL:$platform" -t $nt -K 10000000 $fasta $fastq_1 $fastq_2 || echo -n 'error' ) | $SENTIEON_INSTALL_DIR/bin/sentieon util sort $bam_option -r $fasta -o sorted.bam -t $nt --sam2bam -i -

# ******************************************
# 2. Metrics
# ******************************************
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i sorted.bam --algo MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat --adapter_seq '' aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot GCBias -o gc-report.pdf gc_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot QualDistribution -o qd-report.pdf qd_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot MeanQualityByCycle -o mq-report.pdf mq_metrics.txt
$SENTIEON_INSTALL_DIR/bin/sentieon plot InsertSizeMetricAlgo -o is-report.pdf is_metrics.txt

# ******************************************
# 3. Remove Duplicate Reads
# To mark duplicate reads only without removing them, remove "--rmdup" in the second command
# ******************************************
```

```
$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 --rmdup --score_info score.txt --metrics dedup_metrics.txt $bam_option deduped.bam

# ******************************************

# ******************************************
# 5. Base recalibration
# ******************************************

# Perform recalibration
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam --algo QualCal -k $dbsnp -k $known_Mills_indels -k $known_1000G_indels recal_data.table

# Perform post-calibration check (optional)
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo QualCal -k $dbsnp -k $known_Mills_indels -k $known_1000G_indels recal_data.table.post
$SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt --algo QualCal --plot --before recal_data.table --after recal_data.table.post recal.csv
$SENTIEON_INSTALL_DIR/bin/sentieon plot QualCal -o recal_plots.pdf recal.csv


# ******************************************
# 6. HC Variant caller
# Note: Sentieon default setting matches versions before GATK 3.7.
# Starting GATK v3.7, the default settings have been updated multiple times.
# Below shows commands to match GATK v3.7 - 4.1
# Please change according to your desired behavior.
# ******************************************

# Matching GATK 3.7, 3.8, 4.0
$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo Haplotyper -d $dbsnp --emit_conf=10 --call_conf=10 output-hc.vcf.gz

# Matching GATK 4.1
#$SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q recal_data.table --algo Haplotyper -d $dbsnp --genotype_model multinomial --emit_conf 30 --call_conf 30 output-hc.vcf.gz

```

3. sentieon_ath50x.lsf

```
#BSUB -J Sentieon
#BSUB -n 16
#BSUB -R span[hosts=1]
#BSUB -o %J.out
#BSUB -e %J.err
#BSUB -q normal

# 加载所需软件
# module load sentieon/201808.07
export SENTIEON_LICENSE=mn01:9000
module load SAMtools/1.9
release_dir=/public/home/software/opt/bio/software/Sentieon/201808.07

# 样本名称
i="tair10"
# fastq文件路径
fq1=/public/exercise/sentieon/tair10_1.fastq.gz
fq2=/public/exercise/sentieon/tair10_2.fastq.gz
```

```
# 参考基因组文件
fasta=/public/exercise/sentieon/reference_Tair10/Arabidopsis_thaliana.TAIR10.dna.toplevel.modified.fa
# 输出文件路径
workdir=`pwd`/ath50x_result
# 需要使用的核心数
nt=16
# 比对信息
group_prefix="read_group_name"
platform="ILLUMINA"
mq=30

[ ! -d $workdir ] && mkdir -p $workdir
cd $workdir
# 输出文件
rawCram=$i.cram
sortedCram=$i.q$mq.sorted.cram
depCram=$i.deduped.cram
realnCram=$i.realn.cram
outvcf=$i.vcf
exec > $workdir/$i.callVCF.log 2>&1 # call vcf的日志文件

# ******************************************
# 1. 利用 BWA-MEM 进行比对并排序
# ******************************************
( $release_dir/bin/sentieon bwa mem -M -R "@RG\tID:${i}\tSM:${i}\tPL:$platform" \
-t $nt -K 10000000 $fasta $fq1 $fq2 || echo -n 'error' ) | samtools sort -@ $nt  --output-fmt CRAM \
--reference $fasta -o $rawCram - && samtools index -@ $nt $rawCram
samtools view -hCS -T $fasta -q $mq -o $sortedCram $rawCram && \
samtools index -@ $nt $sortedCram
samtools flagstat $rawCram > $i.stat.raw.txt && \
samtools flagstat $sortedCram > $i.stat.q$mq.txt &

# ******************************************
# 2. Calculate data metrics
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo MeanQualityByCycle ${i}_mq_metrics.txt \
--algo QualDistribution ${i}_qd_metrics.txt --algo GCBias --summary ${i}_gc_summary.txt ${i}_gc_metrics.txt \
--algo AlignmentStat --adapter_seq '' ${i}_aln_metrics.txt --algo InsertSizeMetricAlgo ${i}_is_metrics.txt
$release_dir/bin/sentieon plot metrics -o ${i}_metrics-report.pdf gc=${i}_gc_metrics.txt \
qd=${i}_qd_metrics.txt mq=${i}_mq_metrics.txt isize=${i}_is_metrics.txt
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo LocusCollector --fun score_info ${i}_score.txt

# ******************************************
# 3. 去除 Duplicate Reads
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo Dedup --rmdup --cram_write_options version=3.0 \
--score_info ${i}_score.txt --metrics ${i}_dedup_metrics.txt $depCram && rm -f $sortedCram

# ******************************************
# 4. Indel 重排序 (可选)
# 如果只需要最终的比对结果文件，到这里就可以了，这条命令下面的命令都可以注释掉
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $depCram --algo Realigner --cram_write_options version=3.0 \
$realnCram && rm -f $depCram

```

```
# ******************************************
# 5. Variant calling
# ******************************************
$release_dir/bin/sentieon driver -t $nt -r $fasta -i $realnCram --algo Genotyper $outvcf
```

4. sentieon_Test_People

```
#BSUB -J Sentieon
#BSUB -n 16
#BSUB -R span[hosts=1]
#BSUB -o %J.out
#BSUB -e %J.err
#BSUB -q normal

# 加载所需软件
# module load sentieon/201808.07
export SENTIEON_LICENSE=mn01:9000
module load SAMtools/1.9
release_dir=/public/home/software/opt/bio/software/Sentieon/201808.07

# 样本名称
i="Test"
# fastq文件路径
fq1=/public/exercise/sentieon/1.fastq.gz
fq2=/public/exercise/sentieon/2.fastq.gz
# 参考基因组文件
fasta=/public/exercise/sentieon/reference/ucsc.hg19_chr22.fasta
# 输出文件路径
workdir=`pwd`/People_result
# 需要使用的核心数
nt=16
# 相应数据库
dbsnp=/public/exercise/sentieon/reference/dbsnp_135.hg19_chr22.vcf
known_1000G_indels=/public/exercise/sentieon/reference/1000G_phase1.snps.high_confidence.hg19_chr22.sites.vcf
known_Mills_indels=/public/exercise/sentieon/reference/Mills_and_1000G_gold_standard.indels.hg19_chr22.sites.vcf
# 比对信息
group_prefix="read_group_name"
platform="ILLUMINA"
mq=30

[ ! -d $workdir ] && mkdir -p $workdir
cd $workdir
# 输出文件
rawCram=$i.cram
sortedCram=$i.q$mq.sorted.cram
depCram=$i.deduped.cram
realnCram=$i.realn.cram
outvcf=$i.vcf
exec > $workdir/$i.callVCF.log 2>&1 # call vcf的日志文件

# ******************************************
# 1. 利用 BWA-MEM 进行比对并排序
# ******************************************
( $release_dir/bin/sentieon bwa mem -M -R "@RG\tID:${i}\tSM:${i}\tPL:$platform" \
-t $nt -K 10000000 $fasta $fq1 $fq2 || echo -n 'error' ) | samtools sort -@ $nt  --output-fmt CRAM \
--reference $fasta -o $rawCram - && samtools index -@ $nt $rawCram
samtools view -hCS -T $fasta -q $mq -o $sortedCram $rawCram && \
samtools index -@ $nt $sortedCram
samtools flagstat $rawCram > $i.stat.raw.txt && \
samtools flagstat $sortedCram > $i.stat.q$mq.txt &

# ******************************************
# 2. Calculate data metrics
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo MeanQualityByCycle ${i}_mq_metrics.txt \
--algo QualDistribution ${i}_qd_metrics.txt --algo GCBias --summary ${i}_gc_summary.txt ${i}_gc_metrics.txt \
--algo AlignmentStat --adapter_seq ’’ ${i}_aln_metrics.txt --algo InsertSizeMetricAlgo ${i}_is_metrics.txt
$release_dir/bin/sentieon plot metrics -o ${i}_metrics-report.pdf gc=${i}_gc_metrics.txt \
qd=${i}_qd_metrics.txt mq=${i}_mq_metrics.txt isize=${i}_is_metrics.txt
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo LocusCollector --fun score_info ${i}_score.txt

# ******************************************
# 3. 去除 Duplicate Reads
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo Dedup --rmdup --cram_write_options version=3.0 \
--score_info ${i}_score.txt --metrics ${i}_dedup_metrics.txt $depCram && rm -f $sortedCram

# ******************************************
# 4. Indel 重排序 (可选)
# 如果只需要最终的比对结果文件，到这里就可以了，这条命令下面的命令都可以注释掉
# ******************************************
$release_dir/bin/sentieon driver -r $fasta -t $nt -i $depCram --algo Realigner -k ${known_1000G_indels} --cram_write_options version=3.0 \
$realnCram && rm -f $depCram

# ******************************************
# 5. Variant calling
# ******************************************
$release_dir/bin/sentieon driver -t $nt -r $fasta -i $realnCram --algo Genotyper -d ${dbsnp} ${outvcf}


```