

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

```
1 # 若有VCF database,可参考该流程
2 bsub -J sen-test -n 16 -R span[hosts=1] -o %J.out -e %J.err -q normal "bash
  sentieon_quickstart0.sh"
3
4 # 利用50x拟南芥NGS数据
5 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。

```
1 # 利用50x拟南芥NGS数据
2 bsub < sentieon_ath50x.lsf
3
4 # 若有VCF database,可参考该流程
5 bsub < sentieon_Test_People
```

3.脚本内容

1. sentieon_ath50x.sh

```
1 #!/bin/sh
2 # *****
3 *****
4
5 # Update with the fullpath location of your sample fastq
6 set -x
7 data_dir="$( cd -P "$( dirname "$0" )" && pwd )" #workdir
8 fastq_1=/public/exercise/sentieon/tair10_1.fastq.gz
9 fastq_2=/public/exercise/sentieon/tair10_2.fastq.gz
10
11 # Update with the location of the reference data files
12 fasta=/public/exercise/sentieon/reference_Tair10/Arabidopsis_thaliana.TAIR10.dna.top
  level.modified.fa
13
14 # Set SENTIEON_LICENSE if it is not set in the environment
15 module load SAMtools/1.9
```

```

16 #module load sentieon/201808.07
17   export SENTIEON_LICENSE=mn01:9000
18
19 # Update with the location of the Sentieon software package
20 SENTIEON_INSTALL_DIR=/public/home/software/opt/bio/software/Sentieon/201808.07
21
22 # It is important to assign meaningful names in actual cases.
23 # It is particularly important to assign different read group names.
24 sample="tair10"
25 group="G"
26 platform="ILLUMINA"
27
28 # Other settings
29 nt=16 #number of threads to use in computation
30
31 # *****
32 # 0. Setup
33 # *****
34 workdir=$data_dir/result-tair10
35 mkdir -p $workdir
36 logfile=$workdir/run.log
37 exec >$logfile 2>&1
38 cd $workdir
39
40 #Sentieon proprietary compression
41 bam_option="--bam_compression 1"
42
43 # *****
44 # 1. Mapping reads with BWA-MEM, sorting
45 # *****
46 #The results of this call are dependent on the number of threads used. To have
47 #number of threads independent results, add chunk size option -K 1000000
48
49 # speed up memory allocation malloc in bwa
50 export LD_PRELOAD=$SENTIEON_INSTALL_DIR/lib/libjemalloc.so
51 export MALLOC_CONF=lg_dirty_mult:-1
52
53 ( $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -M -R
54 "$RG\tID:$group\tSM:$sample\tPL:$platform" -t $nt -K 1000000 $fasta $fastq_1
55 $fastq_2 || echo -n 'error' ) | $SENTIEON_INSTALL_DIR/bin/sentieon util sort
56 $bam_option -r $fasta -o sorted.bam -t $nt --sam2bam -i -
57
58 # *****
59 # 2. Metrics
60 # *****
61 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i sorted.bam --algo
62 MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo
63 GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat --adapter_seq ''
64 aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt
65 $SENTIEON_INSTALL_DIR/bin/sentieon plot GCBias -o gc-report.pdf gc_metrics.txt
66 $SENTIEON_INSTALL_DIR/bin/sentieon plot QualDistribution -o qd-report.pdf
67 qd_metrics.txt
68 $SENTIEON_INSTALL_DIR/bin/sentieon plot MeanQualityByCycle -o mq-report.pdf
69 mq_metrics.txt
70 $SENTIEON_INSTALL_DIR/bin/sentieon plot InsertSizeMetricAlgo -o is-report.pdf
71 is_metrics.txt
72
73 # *****
74 # 3. Remove Duplicate Reads
75 # To mark duplicate reads only without removing them, remove "--rmdup" in the second
76 #command
77 # *****
78 $SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt -i sorted.bam --algo LocusCollector
79 --fun score_info score.txt

```

```

68 $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
69
70 # *****
71
72 # *****
73 # 5. Base recalibration
74 # *****
75
76 # Perform recalibration
77 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam --algo
QualCal recal_data.table
78
79 # Perform post-calibration check (optional)
80 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q
recal_data.table --algo QualCal recal_data.table.post
81 $SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt --algo QualCal --plot --before
recal_data.table --after recal_data.table.post recal.csv
82 $SENTIEON_INSTALL_DIR/bin/sentieon plot QualCal -o recal_plots.pdf recal.csv
83
84
85 # *****
86 # 6. HC Variant caller
87 # Note: Sentieon default setting matches versions before GATK 3.7.
88 # Starting GATK v3.7, the default settings have been updated multiple times.
89 # Below shows commands to match GATK v3.7 - 4.1
90 # Please change according to your desired behavior.
91 # *****
92
93 # Matching GATK 3.7, 3.8, 4.0
94 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q
recal_data.table --algo Haplotype --emit_conf=10 --call_conf=10 output-hc.vcf.gz
95
96 # Matching GATK 4.1
97 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q
recal_data.table --algo Haplotype --genotype_model multinomial --emit_conf 30 --
call_conf 30 output-hc.vcf.gz
98

```

2. sentieon_quickstart0.sh

```

1 #!/bin/sh
2 # *****
3 # Script to perform DNA seq variant calling
4 # using a single sample with fastq files
5 # named 1.fastq.gz and 2.fastq.gz
6 # *****
7
8 # Update with the fullpath location of your sample fastq
9 set -x
10 data_dir="$( cd -P "$( dirname "$0" )" && pwd )" #workdir
11 data_dir2=/public/exercise/sentieon #datadir
12 fastq_1=$data_dir2/1.fastq.gz
13 fastq_2=$data_dir2/2.fastq.gz #If using Illumina paired data
14
15 # Update with the location of the reference data files
16 fasta=$data_dir2/reference/ucsc.hg19_chr22.fasta
17 dbsnp=$data_dir2/reference/dbsnp_135.hg19_chr22.vcf
18 known_1000G_indels=$data_dir2/reference/1000G_phase1.snps.high_confidence.hg19_chr2
2. sites.vcf
19 known_Mills_indels=$data_dir2/reference/Mills_and_1000G_gold_standard.indels.hg19_c
hr22.sites.vcf
20

```

```

21 # Set SENTIEON_LICENSE if it is not set in the environment
22 module load SAMtools/1.9
23 #module load sentieon/201808.07
24 export SENTIEON_LICENSE=mn01:9000
25
26 # Update with the location of the Sentieon software package
27 SENTIEON_INSTALL_DIR=/public/home/software/opt/bio/software/Sentieon/201808.07
28
29 # It is important to assign meaningful names in actual cases.
30 # It is particularly important to assign different read group names.
31 sample="sample_name"
32 group="read_group_name"
33 platform="ILLUMINA"
34
35 # Other settings
36 nt=16 #number of threads to use in computation
37
38 # *****
39 # 0. Setup
40 # *****
41 workdir=$data_dir/result
42 mkdir -p $workdir
43 logfile=$workdir/run.log
44 exec >$logfile 2>&1
45 cd $workdir
46
47 #Sentieon proprietary compression
48 bam_option="--bam_compression 1"
49
50 # *****
51 # 1. Mapping reads with BWA-MEM, sorting
52 # *****
53 #The results of this call are dependent on the number of threads used. To have
54 #number of threads independent results, add chunk size option -K 10000000
55
56 # speed up memory allocation malloc in bwa
57 export LD_PRELOAD=$SENTIEON_INSTALL_DIR/lib/libjemalloc.so
58 export MALLOC_CONF=lg_dirty_mult:-1
59
60 ( $SENTIEON_INSTALL_DIR/bin/sentieon bwa mem -M -R
61 "@RG\tID:$group\tSM:$sample\tPL:$platform" -t $nt -K 10000000 $fasta $fastq_1
62 $fastq_2 || echo -n 'error' ) | $SENTIEON_INSTALL_DIR/bin/sentieon util sort
63 $bam_option -r $fasta -o sorted.bam -t $nt --sam2bam -i -
64
65 # *****
66 # 2. Metrics
67 # *****
68 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i sorted.bam --algo
69 MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo
70 GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat --adapter_seq
71 ' ' aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt
72
73 $SENTIEON_INSTALL_DIR/bin/sentieon plot GCBias -o gc-report.pdf gc_metrics.txt
74 $SENTIEON_INSTALL_DIR/bin/sentieon plot QualDistribution -o qd-report.pdf
75 qd_metrics.txt
76 $SENTIEON_INSTALL_DIR/bin/sentieon plot MeanQualityByCycle -o mq-report.pdf
77 mq_metrics.txt
78 $SENTIEON_INSTALL_DIR/bin/sentieon plot InsertSizeMetricAlgo -o is-report.pdf
79 is_metrics.txt
80
81 # *****
82 # 3. Remove Duplicate Reads
83 # To mark duplicate reads only without removing them, remove "--rmdup" in the
84 #second command
85 # *****

```

```

74 $SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt -i sorted.bam --algo
LocusCollector --fun score_info score.txt
75 $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
76
77 # *****
78
79 # *****
80 # 5. Base recalibration
81 # *****
82
83 # Perform recalibration
84 $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
85
86 # Perform post-calibration check (optional)
87 $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
88 $SENTIEON_INSTALL_DIR/bin/sentieon driver -t $nt --algo QualCal --plot --before
recal_data.table --after recal_data.table.post recal.csv
89 $SENTIEON_INSTALL_DIR/bin/sentieon plot QualCal -o recal_plots.pdf recal.csv
90
91
92 # *****
93 # 6. HC Variant caller
94 # Note: Sentieon default setting matches versions before GATK 3.7.
95 # Starting GATK v3.7, the default settings have been updated multiple times.
96 # Below shows commands to match GATK v3.7 - 4.1
97 # Please change according to your desired behavior.
98 # *****
99
100 # Matching GATK 3.7, 3.8, 4.0
101 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q
recal_data.table --algo Haplotype -d $dbsnp --emit_conf=10 --call_conf=10 output-
hc.vcf.gz
102
103 # Matching GATK 4.1
104 $SENTIEON_INSTALL_DIR/bin/sentieon driver -r $fasta -t $nt -i deduped.bam -q
recal_data.table --algo Haplotype -d $dbsnp --genotype_model multinomial --
emit_conf 30 --call_conf 30 output-hc.vcf.gz
105

```

3. sentieon_ath50x.lsf

```

1 #BSUB -J Sentieon
2 #BSUB -n 16
3 #BSUB -R span[hosts=1]
4 #BSUB -o %J.out
5 #BSUB -e %J.err
6 #BSUB -q normal
7
8 # 加载所需软件
9 # module load sentieon/201808.07
10 export SENTIEON_LICENSE=mn01:9000
11 module load SAMtools/1.9
12 release_dir=/public/home/software/opt/bio/software/Sentieon/201808.07
13
14 # 样本名称
15 i="tair10"
16 # fastq文件路径
17 fq1=/public/exercise/sentieon/tair10_1.fastq.gz
18 fq2=/public/exercise/sentieon/tair10_2.fastq.gz

```

```

19 # 参考基因组文件
20 fasta=/public/exercise/sentieon/reference_Tair10/Arabidopsis_thaliana.TAIR10.dna.top
   level.modified.fa
21 # 输出文件路径
22 workdir=`pwd`/ath50x_result
23 # 需要使用的核心数
24 nt=16
25 # 比对信息
26 group_prefix="read_group_name"
27 platform="ILLUMINA"
28 mq=30
29
30 [ ! -d $workdir ] && mkdir -p $workdir
31 cd $workdir
32 # 输出文件
33 rawCram=$i.cram
34 sortedCram=$i.q$mq.sorted.cram
35 depCram=$i.deduped.cram
36 realnCram=$i.realn.cram
37 outvcf=$i.vcf
38 exec > $workdir/$i.callVCF.log 2>&1 # call vcf的日志文件
39
40 # *****
41 # 1. 利用 BWA-MEM 进行比对并排序
42 # *****
43 ( $release_dir/bin/sentieon bwa mem -M -R "@RG\tID:${i}\tSM:${i}\tPL:$platform" \
44 -t $nt -K 1000000 $fasta $fq1 $fq2 || echo -n 'error' ) | samtools sort -@ $nt --
   output-fmt CRAM \
45 --reference $fasta -o $rawCram - && samtools index -@ $nt $rawCram
46 samtools view -hCS -T $fasta -q $mq -o $sortedCram $rawCram && \
47 samtools index -@ $nt $sortedCram
48 samtools flagstat $rawCram > $i.stat.raw.txt && \
49 samtools flagstat $sortedCram > $i.stat.q$mq.txt &
50
51 # *****
52 # 2. Calculate data metrics
53 # *****
54 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo
   MeanQualityByCycle ${i}_mq_metrics.txt \
55 --algo QualDistribution ${i}_qd_metrics.txt --algo GCBias --summary
   ${i}_gc_summary.txt ${i}_gc_metrics.txt \
56 --algo AlignmentStat --adapter_seq '' ${i}_aln_metrics.txt --algo
   InsertSizeMetricAlgo ${i}_is_metrics.txt
57 $release_dir/bin/sentieon plot metrics -o ${i}_metrics-report.pdf
   gc=${i}_gc_metrics.txt \
58 qd=${i}_qd_metrics.txt mq=${i}_mq_metrics.txt isize=${i}_is_metrics.txt
59 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo
   LocusCollector --fun score_info ${i}_score.txt
60
61 # *****
62 # 3. 去除 Duplicate Reads
63 # *****
64 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo Dedup --
   rmdup --cram_write_options version=3.0 \
65 --score_info ${i}_score.txt --metrics ${i}_dedup_metrics.txt $depCram && rm -f
   $sortedCram
66
67 # *****
68 # 4. Indel 重排序 (可选)
69 # 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
70 # *****
71 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $depCram --algo Realigner --
   cram_write_options version=3.0 \
72 $realnCram && rm -f $depCram
73

```

```

74 # *****
75 # 5. Variant calling
76 # *****
77 $release_dir/bin/sentieon driver -t $nt -r $fasta -i $realnCram --algo Genotyper
$outvcf

```

4. sentieon_Test_People

```

1 #BSUB -J Sentieon
2 #BSUB -n 16
3 #BSUB -R span[hosts=1]
4 #BSUB -o %J.out
5 #BSUB -e %J.err
6 #BSUB -q normal
7
8 # 加载所需软件
9 # module load sentieon/201808.07
10 export SENTIEON_LICENSE=mn01:9000
11 module load SAMtools/1.9
12 release_dir=/public/home/software/opt/bio/software/Sentieon/201808.07
13
14 # 样本名称
15 i="Test"
16 # fastq文件路径
17 fq1=/public/exercise/sentieon/1.fastq.gz
18 fq2=/public/exercise/sentieon/2.fastq.gz
19 # 参考基因组文件
20 fasta=/public/exercise/sentieon/reference/ucsc.hg19_chr22.fasta
21 # 输出文件路径
22 workdir=`pwd`/People_result
23 # 需要使用的核心数
24 nt=16
25 # 相应数据库
26 dbsnp=/public/exercise/sentieon/reference/dbsnp_135.hg19_chr22.vcf
27 known_1000G_indels=/public/exercise/sentieon/reference/1000G_phase1.snps.high_confid
ence.hg19_chr22.sites.vcf
28 known_Mills_indels=/public/exercise/sentieon/reference/Mills_and_1000G_gold_standard
.indels.hg19_chr22.sites.vcf
29 # 比对信息
30 group_prefix="read_group_name"
31 platform="ILLUMINA"
32 mq=30
33
34 [ ! -d $workdir ] && mkdir -p $workdir
35 cd $workdir
36 # 输出文件
37 rawCram=$i.cram
38 sortedCram=$i.q$mq.sorted.cram
39 depCram=$i.deduped.cram
40 realnCram=$i.realn.cram
41 outvcf=$i.vcf
42 exec > $workdir/$i.callVCF.log 2>&1 # call vcf的日志文件
43
44 # *****
45 # 1. 利用 BWA-MEM 进行比对并排序
46 # *****
47 ( $release_dir/bin/sentieon bwa mem -M -R "@RG\tID:${i}\tSM:${i}\tPL:$platform" \
48 -t $nt -K 1000000 $fasta $fq1 $fq2 || echo -n 'error' ) | samtools sort -@ $nt --
output-fmt CRAM \
49 --reference $fasta -o $rawCram - && samtools index -@ $nt $rawCram
50 samtools view -hCS -T $fasta -q $mq -o $sortedCram $rawCram && \
51 samtools index -@ $nt $sortedCram
52 samtools flagstat $rawCram > $i.stat.raw.txt && \

```

```

53 samtools flagstat $sortedCram > $i.stat.q$mq.txt &
54
55 # *****
56 # 2. Calculate data metrics
57 # *****
58 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo
MeanQualityByCycle ${i}_mq_metrics.txt \
59 --algo QualDistribution ${i}_qd_metrics.txt --algo GCbias --summary
${i}_gc_summary.txt ${i}_gc_metrics.txt \
60 --algo AlignmentStat --adapter_seq '' ${i}_aln_metrics.txt --algo
InsertSizeMetricAlgo ${i}_is_metrics.txt
61 $release_dir/bin/sentieon plot metrics -o ${i}_metrics-report.pdf
gc=${i}_gc_metrics.txt \
62 qd=${i}_qd_metrics.txt mq=${i}_mq_metrics.txt isize=${i}_is_metrics.txt
63 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo
LocusCollector --fun score_info ${i}_score.txt
64
65 # *****
66 # 3. 去除 Duplicate Reads
67 # *****
68 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $sortedCram --algo Dedup --
rmdup --cram_write_options version=3.0 \
69 --score_info ${i}_score.txt --metrics ${i}_dedup_metrics.txt $depCram && rm -f
$sortedCram
70
71 # *****
72 # 4. Indel 重排序 (可选)
73 # 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
74 # *****
75 $release_dir/bin/sentieon driver -r $fasta -t $nt -i $depCram --algo Realigner -k
${known_1000G_indels} --cram_write_options version=3.0 \
76 $realnCram && rm -f $depCram
77
78 # *****
79 # 5. Variant calling
80 # *****
81 $release_dir/bin/sentieon driver -t $nt -r $fasta -i $realnCram --algo Genotyper -d
${dbsnp} ${outvcf}
82
83

```