Arguments Correspondence

Release 201911

Sentieon, Inc

Jun 12, 2020

Contents

1 Introduction

This documents describes how to execute the Broad institute GATK Best Practices described in https://www.broadinstitute.org/gatk/guide/best-practices using the Sentieon Genomics software. The document also described the correspondence between arguments of the different tools used.

This document should help you determine how to convert your existing pipelines to using Sentieon and allow you to provide feedback to the Sentieon team on what arguments are required for your work but are unavailable in the Sentieon Genomics software.

1.1 Correspondence of tools

The table below shows the Sentieon tool that implement functionality consistent with existing GATK pipeline tools.

Table 1.1: Broad/GATK matching tools

Sentieon tool	GATK pipeline tool	Version correspondence for 201911
Sentieon BWA	BWA	BWA 0.7.17
Dedup and LocusCollector	Picard MarkDuplicates	Picard 2.9.0
Realigner	RealignerTargetCreator and IndelRealigner	GATK 3.7/GATK3.8
QualCal	BaseRecalibrator	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
ReadWriter	PrintReads	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
QualCal	AnalyzeCovariates	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
Genotyper	UnifiedGenotyper	GATK 3.7/GATK3.8
Haplotyper	HaplotypeCaller	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
GVCFtyper	GenotypeGVCFs	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
VarCal	VariantRecalibrator	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
ApplyVarCal	ApplyRecalibration/ApplyVQSR	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
TNsnv	MuTect	MuTect 1.1.5
TNhaplotyper	MuTect2	GATK 3.7/GATK3.8
TNhaplotyper2	GATK4 Mutect2	GATK 4.0.2.1 Mutect2
RNASplitReadsAtJunction	SplitNCigarReads	GATK 3.7/GATK3.8/GATK 4.0/GATK 4.1
AlignmentStat	Picard CollectAlignmentSummaryMetrics	Picard 2.9.0
BaseDistributionByCycle	Picard CollectBaseDistributionByCycle	Picard 2.9.0
CollectVCMetrics	Picard CollectVariantCallingMetrics	Picard 2.9.0
ContaminationAssessment	ContEst	GATK 3.7/GATK3.8
CoverageMetrics	DepthOfCoverage	GATK 3.7/GATK3.8
GCBias	Picard CollectGcBiasMetrics	Picard 2.9.0
HsMetricAlgo	Picard CollectHsMetrics	Picard 2.9.0
InsertSizeMetricAlgo	Picard CollectInsertSizeMetrics	Picard 2.9.0
MeanQualityByCycle	Picard MeanQualityByCycle	Picard 2.9.0
QualDistribution	Picard QualityScoreDistribution	Picard 2.9.0
QualityYield	Picard CollectQualityYieldMetrics	Picard 2.9.0
SequenceArtifactMetricsAlgo	Picard CollectSequencingArtifactMetrics	Picard 2.9.0
WgsMetricsAlgo	Picard CollectWgsMetrics	Picard 2.9.0

2 Detailed description per stage

2.1 Map to Reference - Alignment

GATK Best Practices command line

```
bwa mem -M -R '@RG\tID:GROUP_NAME \tSM:SAMPLE_NAME \tPL:PLATFORM' -p \
   -t NUMBER_THREADS REFERENCE.FASTA SAMPLE.FQ > ALIGNED.SAM
java -jar picard.jar SortSam INPUT=ALIGNED.SAM \
   OUTPUT=SORTED.SAM SORT_ORDER=coordinate
```

(continues on next page)

(continued from previous page)

```
samtools view -bS SORTED.SAM > SORTED.BAM
samtools index SORTED.BAM
```

Sentieon command line

```
sentieon bwa mem -M -R '@RG\tID:GROUP_NAME \tSM:SAMPLE_NAME \tPL:PLATFORM' -p \
   -t NUMBER_THREADS REFERENCE.FASTA SAMPLE.FQ | sentieon util sort \
   -o SORTED.BAM -t NUMBER_THREADS --sam2bam -i -
```

The BWA alignment command is identical, except that we recommend that the results from BWA be piped to the sorting stage in Sentieon, instead of outputting to a SAM file.

The sorting using Sentieon can only be ordered by coordinate.

Sentieon will automatically create an index file for the sorted bam file.

2.2 Mark Duplicates - Dedup

GATK Best Practices command line

```
java -jar picard.jar MarkDuplicates INPUT=SORTED.BAM \
   OUTPUT=DEDUP.BAM METRICS_FILE=DEDUP_METRICS.TXT \
   REMOVE_DUPLICATES=true
java -jar picard.jar BuildBamIndex INPUT=DEDUP.BAM
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -i SORTED.BAM --algo LocusCollector\
   --fun score_info SCORE.TXT.GZ
sentieon driver -t NUMBER_THREADS -i SORTED.BAM --algo Dedup --rmdup \
   --score_info SCORE.TXT.GZ --metrics DEDUP_METRICS.TXT DEDUP.BAM
```

The last argument of the Sentieon command line is the output bam file. Sentieon will automatically create an index file for the deduped bam file.

Table 2.1: Argument correspondence for Dedup

Picard option	Sentieon option	Meaning
INPUT=SORTED.BAM	-i SORTED.BAM	Input the bam file
OUTPUT=DEDUP.BAM	N/A	Output bam file
METRICS_FILE=METRICS.TXT	–metrics METRICS.TXT	Output metrics
REMOVE_DUPLICATES=true	–rmdup	Remove duplicates from bam
OPTICAL_DUPLICATE_PIXEL_DISTANCE=DISTANCE	–optical_dup_pix_dist DISTANCE	Optical duplicate distance

2.3 Realign Indels - Realignment

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
   -R REFERENCE.FASTA -I DEDUP.BAM -L INTERVAL \
   -known KNOWN_SITES.VCF -o REALIGNEMNT_TARGETS.LIST
```

(continues on next page)

(continued from previous page)

```
java -jar GenomeAnalysisTK.jar -T IndelRealigner \
   -R REFERENCE.FASTA -I DEDUP.BAM \
   -targetIntervals REALIGNEMNT_TARGETS.LIST \
   -known KNOWN_SITES.VCF -o REALIGNED.BAM
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i DEDUP.BAM \
   --algo Realigner -k KNOWN_SITES.VCF --interval_list INTERVAL \
   REALIGNED.BAM
```

The last argument of the Sentieon command line is the output bam file.

Table 2.2: Argument correspondence for Realign

GATK option	Sentieon option	Meaning
-I DEDUP.BAM	-i DEDUP.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-o REALIGNED.BAM	N/A	Output bam file
-known KNOWN_SITES.VCF	-k KNOWN_SITES.VCF	Known sites
-L INTERVAL	–interval_list INTERVAL	Interval to restrict calculation

2.4 Recalibrate Bases - BQSR

BQSR - calculate recalibration

GATK Best Practices command line to generate the recalibration table

```
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator \
   -R REFERENCE.FASTA -I REALIGNED.BAM -L INTERVAL \
   -knownSites KNOWN_SITES.VCF -o RECAL_DATA.TABLE
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i REALIGNED.BAM\
   --interval INTERVAL \
   --algo QualCal -k KNOWN_SITES.VCF RECAL_DATA.TABLE
```

The last argument of the Sentieon command line is the recalibrated data table.

Table 2.3: Argument correspondence - calculate BQSR - GATK3

GATK3 option	Sentieon option	Meaning
-I REALIGNED.BAM	-i REALIGNED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-o RECAL_DATA.TABLE	N/A	Output file
-knownSites KNOWN_SITES.VCF	-k KNOWN_SITES.VCF	Known sites
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation

Table 2.4: Argument correspondence - calculate BQSR - GATK4

GATK4 option	Sentieon option	Meaning
-I REALIGNED.BAM	-i REALIGNED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-O RECAL_DATA.TABLE	N/A	Output file
–known-sites KNOWN_SITES.VCF	-k KNOWN_SITES.VCF	Known sites
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation

BQSR - apply recalibration

GATK Best Practices command line to generate the recalibration table

```
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator \
   -R REFERENCE.FASTA -I REALIGNED.BAM -L INTERVAL \
   -knownSites KNOWN_SITES.VCF -BQSR RECAL_DATA.TABLE \
   -o RECAL_DATA.TABLE.POST
java -jar GenomeAnalysisTK.jar -T PrintReads \
   -R REFERENCE.FASTA -I REALIGNED.BAM -L INTERVAL \
   -BQSR RECAL_DATA.TABLE -o RECALED.BAM
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i REALIGNED.BAM\
   -q RECAL_DATA.TABLE --interval INTERVAL \
   --algo QualCal -k KNOWN_SITES RECAL_DATA.TABLE.POST \
   --algo ReadWriter RECALED.BAM
```

The last argument of the Sentieon command line is the output bam file.

The Sentieon ReadWriter command can be run together either with the step generating the **RECAL_DATA.TABLE.POST** above, or with the variant calling step to speed up the pipeline.

Table 2.5: Argument correspondence - apply BQSR - GATK3

GATK3 option	Sentieon option	Meaning
-I REALIGNED.BAM	-i REALIGNED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-o RECALED.BAM	N/A	Output file
-knownSites KNOWN_SITES.VCF	-k KNOWN_SITES.VCF	Known sites
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
-BQSR RECAL_DATA.TABLE	-q RECAL_DATA.TABLE	Recalibration table

Table 2.6: Argument correspondence - apply BQSR - GATK4

GATK4 option	Sentieon option	Meaning
-I REALIGNED.BAM	-i REALIGNED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-O RECALED.BAM	N/A	Output file
–known-sites KNOWN_SITES.VCF	-k KNOWN_SITES.VCF	Known sites
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
-bqsr RECAL_DATA.TABLE	-q RECAL_DATA.TABLE	Recalibration table

BQSR - plot recalibration

GATK Best Practices command line to generate the recalibration table

```
java -jar GenomeAnalysisTK.jar -T AnalyzeCovariates \
   -R REFERENCE.FASTA -before RECAL_DATA.TABLE \
   -after RECAL_DATA.TABLE.POST -csv RECAL_RESULT.CSV -plots BQSR.PDF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS --algo QualCal --plot --before \
   RECAL_DATA.TABLE --after RECAL_DATA.TABLE.POST RECAL_RESULT.CSV
sentieon plot QualCal -o BQSR.PDF RECAL_RESULT.CSV
```

Table 2.7: Argument correspondence - plot BQSR - GATK3

GATK3 option	Sentieon option	Meaning
-R REFERENCE.FASTA	N/A	Reference file
-before RECAL_DATA.TABLE	–before RECAL_DATA.TABLE	Recalibration table
-after RECAL_DATA.TABLE	–after RECAL_DATA.TABLE	After-recalibration table
-plots BQSR.PDF	-o BQSR.PDF	Report file
-csv RECAL_RESULT.CSV	N/A	Output csv file

Table 2.8: Argument correspondence - plot BQSR - GATK4

GATK4 option	Sentieon option	Meaning
-before RECAL_DATA.TABLE	–before RECAL_DATA.TABLE	Recalibration table
-after RECAL_DATA.TABLE	–after RECAL_DATA.TABLE	After-recalibration table
-plots BQSR.PDF	-o BQSR.PDF	Report file
-csv RECAL_RESULT.CSV	N/A	Output csv file

2.5 Unified Genotyper - Genotyper

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper \
   -R REFERENCE.FASTA -I RECALED.BAM -L INTERVAL \
   -D DBSNP.VCF --glm [SNP/INDEL/BOTH] -mbq QUALITY \
   -stand_emit_conf CONFIDENCE -stand_call_conf CONFIDENCE \
   --output_mode [EMIT_VARIANTS_ONLY/EMIT_ALL_CONFIDENT_SITES/EMIT_ALL_SITES] \
   -o OUTPUT.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i RECALED.BAM\
   --interval INTERVAL \
   --algo Genotyper \
   -d DBSNP.VCF --var_type [SNP/INDEL/BOTH] --min_base_qual QUALITY \
   --emit_conf CONFIDENCE --call_conf CONFIDENCE \
   --emit_mode [VARIANT/CONFIDENT/ALL] \
   OUTPUT.VCF
```

The last argument of the Sentieon command line is the variant vcf file. The tool will output a compressed VCF file when using .gz extension.

Bear in mind that since GATK 3.7, the `stand_emit_conf` is no longer supported, and the default value for `stand_call_conf` has been changed from 30 to 10, while the default in Sentieon `call_conf` stayed at 30.

Table 2.9: Argument correspondence - UnifiedGenotyper

GATK option	Sentieon option	Meaning
-I RECALED.BAM	-i RECALED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-D DBSNP.VCF	-d DBSNP.VCF	dbSNP file
–glm [SNP/INDEL/BOTH]	–var_type [SNP/INDEL/BOTH]	Variant output type
-mbq QUALITY	–min_base_qual QUALITY	Minimum base quality
-stand_emit_conf CONFIDENCE	–emit_conf CONFIDENCE	Emit confidence threshold
-stand_call_conf CONFIDENCE	–call_conf CONFIDENCE	Call confidence threshold
–output_mode MODE	–emit_mode MODE	Emit mode
-ploidy PLOIDY	–ploidy PLOIDY	Ploidy of the sample
-o OUTPUT.VCF	N/A	Output variant file
-alleles GIVEN.VCF -gt_mode GENOTYPE_GIVEN_ALLELES	–given GIVEN.VCF	Perform variant calling using only the variants provided in the GIVEN_VCF

2.6 HaplotypeCaller - Halotyper

GATK Best Practices command line - VCF output

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller \
   -R REFERENCE.FASTA -I RECALED.BAM -L INTERVAL \
   -D DBSNP.VCF -mbq QUALITY --minPruning FACTOR \
   -stand_emit_conf CONFIDENCE -stand_call_conf CONFIDENCE \
   -pcrModel [HOSTILE/AGGRESSIVE/CONSERVATIVE/NONE] \
   --output_mode [EMIT_VARIANTS_ONLY/EMIT_ALL_CONFIDENT_SITES/EMIT_ALL_SITES] \
   -o OUTPUT.VCF
```

Sentieon command line - VCF output

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i RECALED.BAM\
   --interval INTERVAL \
   --algo Haplotyper -d DBSNP.VCF \
   --min_base_qual QUALITY --prune_factor FACTOR \
   --emit_conf CONFIDENCE --call_conf CONFIDENCE \
   --emit_mode [VARIANT/CONFIDENT/ALL]
   --pcr_indel_model [HOSTILE/AGGRESSIVE/CONSERVATIVE/NONE] \
   OUTPUT.VCF
```

GATK Best Practices command line - gVCF output

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller \
   -R REFERENCE.FASTA -I RECALED.BAM -L INTERVAL \
   -D DBSNP.VCF -mbq QUALITY --minPruning FACTOR \
   -stand_emit_conf CONFIDENCE -stand_call_conf CONFIDENCE \
   -pcrModel [HOSTILE/AGGRESSIVE/CONSERVATIVE/NONE] \
   --emitRefConfidence GVCF \
   -o OUTPUT.VCF
```

Sentieon command line - gVCF output

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i RECALED.BAM \
   --interval INTERVAL \
   --algo Haplotyper -d DBSNP.VCF \
   --min_base_qual QUALITY --prune_factor FACTOR \
   --emit_conf CONFIDENCE --call_conf CONFIDENCE \
   --emit_mode GVCF
   --pcr_indel_model [HOSTILE/AGGRESSIVE/CONSERVATIVE/NONE] \
   OUTPUT.VCF
```

The last argument of the Sentieon command line is the output vcf file. The tool will output a compressed VCF file when using .gz extension.

Bear in mind that since GATK 3.7, the `stand_emit_conf` is no longer supported. Also, the default value for `stand_call_conf` was changed from 30 to 10 in the GATK 3.7 to GATK 4.0 and was reverted to 30 in the GATK 4.1, while the default in Sentieon `call_conf` has remained at 30.

Since the GATK 4.1 `-newQual` is default genotyping model.

Table 2.10: Argument correspondence - HaplotypeCaller - GATK3

GATK3 option	Sentieon option	Meaning
-I RECALED.BAM	-i RECALED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-D DBSNP.VCF	-d DBSNP.VCF	dbSNP file
-mbq QUALITY	–min_base_qual QUALITY	Minimum base quality
-stand_emit_conf CONFIDENCE	–emit_conf CONFIDENCE	Emit confidence threshold
-stand_call_conf CONFIDENCE	–call_conf CONFIDENCE	Call confidence threshold
–output_mode MODE	–emit_mode MODE	Emit mode
–emitRefConfidence GVCF	–emit_mode gvcf	Produce a g.vcf output
-ploidy PLOIDY	–ploidy PLOIDY	Ploidy of the sample
-o OUTPUT.VCF	N/A	Output variant file
-alleles GIVEN.VCF -gt_mode GENOTYPE_GIVEN_ALLELES	–given GIVEN.VCF	Perform variant calling using only the variants provided in the GIVEN_VCF
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
-mmq QUALITY	–min_map_qual QUALITY	Minimum mapping quality
–minPruning FACTOR	–prune_factor FACTOR	Pruning factor
-pcrModel MODEL	–pcr_indel_model MODEL	PCR model
–dontUseSoftClippedBases	–trim_soft_clip	Trim off soft-clipped bases
–annotation ANNOTATION	–annotation ANNOTATION	Annotations to apply to the variant calls
–excludeAnnotation ANNOTATION	–annotation !ANNOTATION	Annotations to exclude in the variant calls by using the '!' prefix
-newQual	–genotype_model multinomial	Use the new simplified allele count model

Table 2.11: Argument correspondence - HaplotypeCaller - GATK4

GATK4 option	Sentieon option	Meaning
-I RECALED.BAM	-i RECALED.BAM	Input the bam file
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-D DBSNP.VCF	-d DBSNP.VCF	dbSNP file
-mbq QUALITY	–min_base_qual QUALITY	Minimum base quality
N/A	–emit_conf CONFIDENCE	Emit confidence threshold
-stand-call-conf CONFIDENCE	–call_conf CONFIDENCE	Call confidence threshold
–output-mode MODE	–emit_mode MODE	Emit mode
-ERC GVCF	–emit_mode gvcf	Produce a g.vcf output
-ploidy PLOIDY	–ploidy PLOIDY	Ploidy of the sample
-O OUTPUT.VCF	N/A	Output variant file
–alleles GIVEN.VCF –genotyping-mode GENOTYPE_GIVEN_ALLELES	–given GIVEN.VCF	Perform variant calling using only the variants provided in the GIVEN_VCF
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
–minimum-mapping-quality QUALITY	–min_map_qual QUALITY	Minimum mapping quality
–min-pruning FACTOR	–prune_factor FACTOR	Pruning factor
–pcr-indel-model MODEL	–pcr_indel_model MODEL	PCR model
–dont-use-soft-clipped-bases	–trim_soft_clip	Trim off soft-clipped bases
–annotation ANNOTATION	–annotation ANNOTATION	Annotations to apply to the variant calls
–annotations-to-exclude ANNOTATION	–annotation !ANNOTATION	Annotations to exclude in the variant calls by using the '!' prefix
-new-qual	–genotype_model multinomial	Use the new simplified allele count model

2.7 Joint Genotype - GVCFtyper

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs \
   -R REFERENCE.FASTA -L INTERVAL \
   -D DBSNP.VCF \
   -stand_emit_conf CONFIDENCE -stand_call_conf CONFIDENCE \
   -V INPUT_GVCF_1 -V INPUT_GVCF_2 -V INPUT_GVCF_3 \
   -o OUTPUT.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   --interval INTERVAL \
   --algo GVCFtyper \
   -d DBSNP.VCF \
   --emit_conf CONFIDENCE --call_conf CONFIDENCE \
   --emit_mode [VARIANT/CONFIDENT/ALL] \
   -v INPUT_GVCF_1 -v INPUT_GVCF_2 -v INPUT_GVCF_3 \
   OUTPUT.VCF
```

The last argument of the Sentieon command line is the output vcf file. The tool will output a compressed VCF file when using .gz extension.

Bear in mind that since GATK 3.7, the `stand_emit_conf` is no longer supported. Also, the default value for `stand_call_conf` was changed from 30 to 10 in the GATK 3.7 to GATK 4.0 and was reverted to 30 in the GATK 4.1, while the default in Sentieon `call_conf` has remained at 30.

Since the GATK 4.1 `-newQual` is default genotyping model.

Table 2.12: Argument correspondence - GenotypeGVCF - GATK3

GATK3 option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
-D DBSNP.VCF	-d DBSNP.VCF	dbSNP file
-stand_emit_conf CONFIDENCE	–emit_conf CONFIDENCE	Emit confidence threshold
-stand_call_conf CONFIDENCE	–call_conf CONFIDENCE	Call confidence threshold
N/A	–emit_mode MODE	Emit mode
-V INPUT_GVCF_X	-v INPUT_GVCF_X	g.vcf input files
-o OUTPUT.VCF	N/A	Output variant file
-newQual	–genotype_model multinomial	Use the new simplified allele count model

Table 2.13: Argument correspondence - GenotypeGVCF - GATK4

GATK4 option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-L INTERVAL	–interval INTERVAL	Interval to restrict calculation
-D DBSNP.VCF	-d DBSNP.VCF	dbSNP file
N/A	–emit_conf CONFIDENCE	Emit confidence threshold
-stand-call-conf CONFIDENCE	–call_conf CONFIDENCE	Call confidence threshold
N/A	–emit_mode MODE	Emit mode
-V INPUT_GVCF_X	-v INPUT_GVCF_X	g.vcf input files
-O OUTPUT.VCF	N/A	Output variant file
-new-qual	–genotype_model multinomial	Use the new simplified allele count model

2.8 Filter Variants - VQSR

VQSR - calculate recalibration

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T VariantRecalibrator\
   -R REFERENCE.FASTA -input INPUT.VCF \
   -an ANNOTATION_1 -an ANNOTATION_2 . . .  \
   -mode [SNP/INDEL] \
   --resource:RESOURCE_PARAM RESOURCE.VCF . . . \
   -tranche TRANCH_THRES -tranche TRANCH_THRES . . . \
   --maxGaussians MAX_GAUSS --maxNegativeGaussians MAX_GAUSS \
   --maxIterations MAX_ITERATIONS \
   --aggregate AGREGATE_VCF \
   -tranchesFile TRANCHES_FILE \
   -rscriptFile R_PLOT_FILE \
   -recalFile RECAL_FILE
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   --algo VarCal -v INPUT.VCF \
   --annotation ANNOTATION_1 --annotation ANNOTATION_2 . . . \
   --var_type [SNP/INDEL]  \
   --resource RESOURCE.VCF --resource_param RESOURCE_PARAM . . . \
   --tranche TRANCH_THRES --tranche TRANCH_THRES . . . \
   --max_gaussian MAX_GAUSS --max_neg_gaussian MAX_GAUSS \
   --max_iter MAX_ITERATIONS \
   --nthr NUMBER_THREADS_EM --srand RANDOM_SEED \
   --aggregate_data AGREGATE_VCF \
   --tranches_file TRANCHES_FILE \
   --plot_file PLOT_FILE \
   RECAL_FILE
```

The last argument of the Sentieon command line is the output recal file.

The resource argument in Sentieon is split into 2 consecutive arguments, one with the resource file and one with the resource parameters.

Table 2.14: Argument correspondence - calculate VQSR - GATK3

GATK3 option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-input INPUT.VCF	-v INPUT.VCF	vcf input file
-an ANNONTATION	–annotation ANNOTATION	Annotation to use
-mode [SNP/INDEL]	–var_type [SNP/INDEL]	Mode to use
–resource	–resource/–resource_param	Resources to use
-tranche TRANCH_THRES	–tranche TRANCH_THRES	Thresholds for tranches
–maxGaussians MAX_GAUSS	–max_gaussians MAX_GAUSS	Max number of Gaussians used for positive model
–maxNegativeGaussians MAX_GAUSS	–max_neg_gaussians MAX_GAUSS	Max number of Gaussians used for negative model
–maxIterations MAX_ITERATIONS	–max_iter MAX_ITERATIONS	Max number of iterations
N/A	–srand RANDOM_SEED	Random seed for the EM calculation
–aggregate AGREGATE_VCF	–aggregate_data AGREGATE_VCF	Input aggregate data
-tranchesFile TRANCHES_FILE	–tranches_file TRANCHES_FILE	Output tranches file
-rscriptFile R_PLOT_FILE	–plot_file PLOT_FILE	Output file for plotting
-recalFile RECAL_FILE	N/A	Output recalibration file
-MQCap NUMBER	–max_mq NUMBER	Maximum MQ in the data

Table 2.15: Argument correspondence - calculate VQSR - GATK4

GATK4 option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
–variant INPUT.VCF	-v INPUT.VCF	vcf input file
-an ANNONTATION	–annotation ANNOTATION	Annotation to use
-mode [SNP/INDEL]	–var_type [SNP/INDEL]	Mode to use
–resource	–resource/–resource_param	Resources to use
-tranche TRANCH_THRES	–tranche TRANCH_THRES	Thresholds for tranches
–max-gaussians MAX_GAUSS	–max_gaussians MAX_GAUSS	Max number of Gaussians used for positive model
–max-negative-gaussians MAX_GAUSS	–max_neg_gaussians MAX_GAUSS	Max number of Gaussians used for negative model
–max-iterations MAX_ITERATIONS	–max_iter MAX_ITERATIONS	Max number of iterations
N/A	–srand RANDOM_SEED	Random seed for the EM calculation
–aggregate AGREGATE_VCF	–aggregate_data AGREGATE_VCF	Input aggregate data
–tranches-file TRANCHES_FILE	–tranches_file TRANCHES_FILE	Output tranches file
–rscript-file R_PLOT_FILE	–plot_file PLOT_FILE	Output file for plotting
–recal-file RECAL_FILE	N/A	Output recalibration file
-mq-cap NUMBER	–max_mq NUMBER	Maximum MQ in the data

VQSR - apply recalibration

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T ApplyRecalibration\
   -R REFERENCE.FASTA -input INPUT.VCF \
   -mode [SNP/INDEL] --ts_filter_level SENSITIVITY \
   -tranchesFile TRANCHES_FILE -recalFile RECAL_FILE \
   -o OUTPUT.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   --algo ApplyVarCal -v INPUT.VCF \
   --var_type [SNP/INDEL] --sensitivity SENSITIVITY \
   --tranches_file TRANCHES_FILE --recal RECAL_FILE \
   OUTPUT.VCF
```

The last argument of the Sentieon command line is the output vcf file. The tool will output a compressed VCF file when using .gz extension.

Table 2.16: Argument correspondence - apply VQSR - GATK3

GATK option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-input INPUT.VCF	-v INPUT.VCF	vcf input file
–ts_filter_level SENSITIVITY	–sensitivity SENSITIVITY	Sensitivity
-mode [SNP/INDEL]	–var_type [SNP/INDEL]	Mode to use
-tranchesFile TRANCHES_FILE	–tranches_file TRANCHES_FILE	Input tranches file
-recalFile RECAL_FILE	–recal RECAL_FILE	Input recalibration file
-o OUTPUT.VCF	N/A	Output variant file

Table 2.17: Argument correspondence - apply VQSR - GATK4

GATK option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-V INPUT.VCF	-v INPUT.VCF	vcf input file
–ts_filter_level SENSITIVITY	–sensitivity SENSITIVITY	Sensitivity
-mode [SNP/INDEL]	–var_type [SNP/INDEL]	Mode to use
–tranches-file TRANCHES_FILE	–tranches_file TRANCHES_FILE	Input tranches file
–recal-file RECAL_FILE	–recal RECAL_FILE	Input recalibration file
-O OUTPUT.VCF	N/A	Output variant file

2.9 MuTect - TNsnv

MuTect Best Practices command line

```
java -jar mutect.jar -T MuTect \
   -R REFERENCE.FASTA -L INTERVAL \
   -I:normal NORMAL_RECALED.BAM -I:tumor TUMOR_RECALED.BAM \
   --dbsnp DBSNP.VCF -o CALL_STATS_OUTPUT.TXT -vcf OUTPUT.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i NORMAL_RECALED.BAM -i TUMOR_RECALED.BAM \
   --interval INTERVAL \
   --algo TNsnv --dbsnp DBSNP.VCF \
   --tumor_sample TUMOR_SM --normal_sample NORMAL_SM \
   -call_stats_out CALL_STATS_OUTPUT.TXT OUTPUT.VCF
```

The last argument of the Sentieon command line is the output vcf file. The tool will output a compressed VCF file when using .gz extension.

The `normal_sample` and `tumor_sample` arguments are required in Sentieon as the BAM files are not explicitly differentiated, and could be a single co-realigned BAM file.

Table 2.18: Argument correspondence - TNsnv

GATK option	Sentieon option	Meaning
N/A	-i COREALIGNED.BAM	Input the corealigned bam file
-I:normal NORMAL_RECALED.BAM	-i NORMAL_RECALED.BAM	Input the bam files
-I:tumor TUMOR_RECALED.BAM	-i TUMOR_RECALED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
–normal_sample_name NORMAL_SAMPLE	–normal_sample NORMAL_SAMPLE	Input normal sample name
–tumor_sample_name TUMOR_SAMPLE	–tumor_sample TUMOR_SAMPLE	Input tumor sample name
–dbsnp DBSNP.VCF	–dbsnp DBSNP.VCF	dbSNP file
–cosmic COSMIC.VCF	–cosmic COSMIC.VCF	Input cosmic VCF file
–normal_panel PON.VCF	–pon PON.VCF	Input panel-of-normal VCF file
–artifact_detection_mode	–detect_pon	Turn on mode to detect artifacts in normal sample, used to generate the panel-of-normal
-vcf OUTPUT.VCF	N/A	Output tumor variants file

Continued on next page

Table 2.18 – continued from previous page

GATK option	Sentieon option	Meaning
-o CALL_STATS.OUT	–call_stats_out CALL_STATS.OUT	Output call statistics file
–coverage_file COVERAGE_FILE	–stdcov_out COVERAGE_FILE	Output standard coverage wiggle file
–tumor_depth_file FILE	–tumor_depth_out FILE	Output wiggle file of depth of tumor reads
–normal_depth_file FILE	–normal_depth_out FILE	Output wiggle file of depth of normal reads
–power_file FILE	–power_out FILE	Output power file
–min_qscore QUALITY	–min_base_qual QUALITY	Filtering quality of the bases used in variant calling
–initial_tumor_lod NUMBER	–min_init_tumor_lod NUMBER	Minimum tumor log odds in the initial pass calling variants
–tumor_lod NUMBER	–min_tumor_lod NUMBER	Minimum tumor log odds in the final call of variants
–normal_lod NUMBER	–min_normal_lod NUMBER	Minimum normal log odds used to check that the tumor variant is not a normal variant
–fraction_contamination NUMBER	–contamination_frac NUMBER	Estimation of the contamination fraction from other samples
–minimum_mutation_cell_fraction NUMBER	–min_cell_mutation_frac NUMBER	Minimum fraction of cells which have mutation
–strand_artifact_lod NUMBER	–min_strand_bias_lod NUMBER	Minimum log odds for calling strand bias
–strand_artifact_power_threshold NUMBER	–min_strand_bias_power NUMBER	Minimum power for calling strand bias
–dbsnp_normal_lod NUMBER	–min_dbsnp_normal_lod NUMBER	Minimum log odds for calling normal non-variant at dbsnp sites
–minimum_normal_allele_fraction NUMBER	–min_normal_allele_frac NUMBER	Minimum allele fraction to be considered in normal
–tumor_f_pretest NUMBER	–min_tumor_allele_frac NUMBER	Minimum allelic fraction in tumor sample
–gap_events_threshold NUMBER	–max_indel NUMBER	Maximum of nearby indel events that are allowed
–heavily_clipped_read_fraction NUMBER	–max_read_clip_frac NUMBER	Maximum fraction of soft/hard clipped bases in a read
–fraction_mapq0_threshold NUMBER	–max_mapq0_frac NUMBER	Maximum ratio of reads whose mapq are 0 used to determine poor mapped area
–pir_median_threshold NUMBER	–min_pir_median NUMBER	Minimum read position median
–pir_mad_threshold NUMBER	–min_pir_mad NUMBER	Minimum read position median absolute deviation
–required_maximum_alt_allele_mapping_quality_score NUMBER	–max_alt_mapq NUMBER	Required maximum value of alt allele mapping quality score
–max_alt_alleles_in_normal_count NUMBER	–max_normal_alt_cnt NUMBER	Maximum alt alleles count in normal pileup
–max_alt_alleles_in_normal_qscore_sum NUMBER	–max_normal_alt_qsum NUMBER	Maximum quality score sum of alt allele in normal pileup
–max_alt_allele_in_normal_fraction NUMBER	–max_normal_alt_frac NUMBER	Maximum fraction of alt allele in normal pileup

Continued on next page

Table 2.18 – continued from previous page

GATK option	Sentieon option	Meaning
–power_constant_af NUMBER	–power_allele_frac NUMBER	Allele fraction used in power calculations

2.10 MuTect2 - TNhaplotyper

MuTect2 Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T MuTect2 \
   -R REFERENCE.FASTA -L INTERVAL \
   -I:normal NORMAL_RECALED.BAM -I:tumor TUMOR_RECALED.BAM \
   -D DBSNP.VCF -o OUTPUT.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i NORMAL_RECALED.BAM -i TUMOR_RECALED.BAM \
   --interval INTERVAL \
   --algo TNhaplotyper --dbsnp DBSNP.VCF \
   --tumor_sample TUMOR_SM --normal_sample NORMAL_SM \
   OUTPUT.VCF
```

The last argument of the Sentieon command line is the output vcf file. The tool will output a compressed VCF file when using .gz extension.

The `normal_sample` and `tumor_sample` arguments are required in Sentieon as the BAM files are not explicitly differentiated, and could be a single co-realigned BAM file.

Table 2.19: Argument correspondence - TNhaplotyper

GATK option	Sentieon option	Meaning
N/A	-i COREALIGNED.BAM	Input the corealigned bam file
-I:normal NORMAL_RECALED.BAM	-i NORMAL_RECALED.BAM	Input the bam files
-I:tumor TUMOR_RECALED.BAM	-i TUMOR_RECALED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
N/A	–normal_sample NORMAL_SAMPLE	Input normal sample name
N/A	–tumor_sample TUMOR_SAMPLE	Input tumor sample name
-D DBSNP.VCF	–dbsnp DBSNP.VCF	dbSNP file
–cosmic COSMIC.VCF	–cosmic COSMIC.VCF	Input cosmic VCF file
–normal_panel PON.VCF	–pon PON.VCF	Input panel-of-normal VCF file
–artifact_detection_mode	–detect_pon	Turn on mode to detect artifacts in normal sample. It is used to generate the panel-of-normals
-o OUTPUT.VCF	N/A	Output tumor variants file.
-mbq QUALITY	–min_base_qual QUALITY	Filtering quality of the bases used in variant calling
–minPruning FACTOR	–prune_factor FACTOR	Pruning factor
-pcrModel MODEL	–pcr_indel_model MODEL	PCR model
–initial_tumor_lod NUMBER	–min_init_tumor_lod NUMBER	Minimum tumor log odds in the initial pass calling variants
–initial_normal_lod NUMBER	–min_init_normal_lod NUMBER	Minimum normal log odds in the initial pass calling variants
–tumor_lod NUMBER	–min_tumor_lod NUMBER	Minimum tumor log odds in the final call of variants
–normal_lod NUMBER	–min_normal_lod NUMBER	Minimum normal log odds used to check that the tumor variant is not a normal variant
–max_alt_alleles_in_normal_count NUMBER	–max_normal_alt_cnt NUMBER	Maximum alt alleles count in normal pileup
–max_alt_alleles_in_normal_qscore_sum NUMBER	–max_normal_alt_qsum NUMBER	Maximum quality score sum of alt allele in normal pileup
–max_alt_allele_in_normal_fraction NUMBER	–max_normal_alt_frac NUMBER	Maximum fraction of alt allele in normal pileup
-contaminationFile TAB_FILE	–tumor_contamination_frac NUMBER	Estimation of the contamination fraction from other samples on the tumor sample
	–normal_contamination_frac NUMBER	Estimation of the contamination fraction from other samples on the normal sample

2.11 GATK4 Mutect2 - TNhaplotyper2 and tnhapfilter

GATK4 Mutect2 Best Practies command line

```
java -jar gatk-package.jar Mutect2 -R REFERENCE.FASTA \
   -O TMP.VCF -tumor TUMOR_SM -normal NORMAL_SM \
   -I TUMOR_RECALED.BAM -I NORMAL_RECALED.BAM \
   -L INTERVAL
java -jar gatk-package.jar FilterMutectCalls -L INTERVAL \
   -O OUTPUT.VCF -V TMP.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i NORMAL_RECALED.BAM -i TUMOR_RECALED.BAM \
   --interval INTERVAL \
   --algo TNhaplotyper2 --tumor_sample TUMOR_SM \
   --normal_sample NORMAL_SM TMP.VCF
sentieon tnhapfilter --tumor_sample TUMOR_SM \
   --normal_sample NORMAL_SM \
   -v TMP.VCF OUTPUT.VCF
```

Table 2.20: Argument correspondence - TNhaplotyper2

GATK4 option	Sentieon option	Meaning
-I TUMOR_RECALED.BAM	-i TUMOR_RECALED.BAM	Input the bam files
-I NORMAL_RECALED.BAM	-i NORMAL_RECALED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference fasta
-tumor TUMOR_SM	–tumor_sample TUMOR_SM	Input tumor sample name
-normal NORMAL_SM	–normal_sample NORMAL_SM	Input normal sample name
-pon PON_FILE	–pon PON_FILE	A panel-of-normal file
-mbq MIN_BQ	–min_base_qual MIN_BQ	Minimum base quality
–min-pruning PRUNE	–prune_factor PRUNE	Pruning factor in local assembly
–pcr-indel-model INDEL_MODEL	–pcr_indel_model INDEL_MODEL	PCR indel error model
-init-lod INIT_T_LOD	–min_init_tumor_lod INIT_T_LOD	Minimum tumorLOD for candidate selection
-emit-lod T_LOD	–min_tumor_lod T_LOD	Minimum tumorLOD for called variants
–normal-lod N_LOD	–min_normal_lod N_LOD	Minimum normalLOD for called variants
-contamination T_FRAC	–tumor_contamination_frac T_FRAC	Estimated tumor contamination fraction
-contamination N_FRAC	–normal_contamination_frac N_FRAC	Estimated normal contamination fraction
–germline-resource GERMLINE.VCF	–germline_vcf GERMLINE.VCF	A germline VCF containing allele frequencies
–af-of-alleles-not-in-resource AF	–default_af AF	Allele frequency for variants not found in the germline VCF
–max-population-af MAX_AF	–max_germline_af MAX_AF	Maximum germline allele frequency in tumor-only mode
–genotype-pon-sites true	–call_pon_sites	Call candidate variants in the PoN

Table 2.21: Argument correspondence - tnhapfilter

GATK4 option	Sentieon option	Meaning
N/A	–tumor_sample TUMOR_SM	Input tumor sample name
N/A	–normal_sample NORMAL_SM	Input normal sample name
–max-strand-artifact-probability NUMBER	–max_strand_prob NUMBER	Maximum strand artifact probability
–normal-artifact-lod NUMBER	–max_normal_art_lod NUMBER	Maximum log odds of a normal artifact
–min-median-base-quality NUMBER	–min_median_base_qual NUMBER	Minimum median base quality
–contamination-fraction-to-filter NUMBER	–contamination NUMBER	Contamination fraction to filter
–min-strand-artifact-allele-fraction NUMBER	–min_strand_af NUMBER	Minimum strand artifact allele fraction
–max-median-fragment-length-difference NUMBER	–max_diff_fraglen NUMBER	Maximum median difference in fragment length
–min-median-mapping-quality NUMBER	–min_median_mapq NUMBER	Minimum median mapping quality
–tumor-lod NUMBER	–min_tumor_lod NUMBER	Minimum tumor log odds
–max-events-in-region NUMBER	–max_event_cnt NUMBER	Maximum events in region
–max-germline-posterior NUMBER	–max_germline_prob NUMBER	Maximum germline probability
–max-alt-allele-count NUMBER	–max_alt_cnt NUMBER	Maximum alt allele count
–min-median-read-position NUMBER	–min_pir_median NUMBER	Minimum median read position

2.12 SplitNCigarReads - RNASplitReadsAtJunction

SplitNCigarReads Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T SplitNCigarReads \
   -R REFERENCE.FASTA -I DEDUPED.BAM -o SPLIT.BAM\
   -rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 \
   -U ALLOW_N_CIGAR_READS
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo RNASplitReadsAtJunction \
   --reassign_mapq 255:60 SPLIT.BAM
```

The last argument of the Sentieon command line is the output bam file.

Table 2.22: Argument correspondence - SplitNCigarReads and RNASplitReadsAtJunction - GATK3

GATK3 option	Sentieon option	Meaning
-I DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference fasta
-rf ReassignOneMappingQuality -RMQF IN_QUAL -RMQT OUT_QUAL	–reassign_mapq IN_QUAL:OUT_QUAL	Reassign Mapping Quality from STAR
–doNotFixOverhangs	–ignore_overhang	Whether to ignore overhang
–maxBasesInOverhang NUMBER	–overhang_max_bases NUMBER	Max number of bases allowed in a hard-clipped overhang. Overhang will not be clipped if there are more than this value of bases
–maxMismatchesInOverhang NUMBER	–overhang_max_mismatches NUMBER	Max number of mismatches allowed in a non-hard-clipped overhang. Complete overhang will be hard-clipped if # of mismatches is above this value

Table 2.23: Argument correspondence - SplitNCigarReads and RNASplitReadsAtJunction - GATK4

GATK4 option	Sentieon option	Meaning
-I DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference fasta
N/A	–reassign_mapq IN_QUAL:OUT_QUAL	Reassign Mapping Quality from STAR
–do-not-fix-overhangs	–ignore_overhang	Whether to ignore overhang
–max-bases-in-overhang NUMBER	–overhang_max_bases NUMBER	Max number of bases allowed in a hard-clipped overhang. Overhang will not be clipped if there are more than this value of bases
–max-mismatches-in-overhang NUMBER	–overhang_max_mismatches NUMBER	Max number of mismatches allowed in a non-hard-clipped overhang. Complete overhang will be hard-clipped if # of mismatches is above this value

2.13 CollectAlignmentSummaryMetrics - AlignmentStat

Picard CollectAlignmentSummaryMetrics command line

```
java -jar picard.jar CollectAlignmentSummaryMetrics \
   I=ALIGNED.BAM O=ALN_METRICS.TXT \
   R=REFERENCE.FASTA \
   ADAPTER_SEQUENCE=ADAPTERS_SEQ
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i ALIGNED.BAM --algo AlignmentStat \
   --adapter_seq ADAPTERS_SEQ ALN_METRICS.TXT
```

Table 2.24: Argument correspondence - CollectAlignmentSummaryMetrics and AlignmentStat

Picard option	Sentieon option	Meaning
I=ALIGNED.BAM	-i ALIGNED.BAM	Input the bam files
O=ALN_METRICS.TXT	N/A	Output metrics
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
ADAPTER_SEQUENCE=ADAPTERS_SEQ	–adapter_seq ADAPTERS_SEQ	A string of adapters

2.14 CollectBaseDistributionByCycle - BaseDistributionByCycle

Picard CollectBaseDistributionByCycle command line

```
java -jar picard.jar CollectBaseDistributionByCycle \
   I=ALIGNED.BAM O=BASE_DISTRIBUTION_METRICS.TXT \
   CHART_OUTPUT=BASE_DISTRIBUTION.PDF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i ALIGNED.BAM --algo BaseDistributionByCycle \
   BASE_DISTRIBUTION_METRICS.TXT
```

Table 2.25: Argument correspondence - CollectBaseDistributionByCycle and BaseDistributionByCycle

Picard option	Sentieon option	Meaning
I=ALIGNED.BAM	-i ALIGNED.BAM	Input the bam files
O=BASE_DISTRIBUTION_METRICS.TXT	N/A	Output metrics
CHART_OUTPUT=BASE_DISTRIBUTION.PDF	N/A	Output chart
ALIGNED_READS_ONLY=true	–aligned_reads_only true	Calculate the base distribution over aligned reads only
PF_READS_ONLY=true	–pf_reads_only true	Calculate the base distribution over PF reads only

2.15 CollectVariantCallingMetrics - CollectVCMetrics

Picard CollectVariantCallingMetrics command line

```
java -jar picard.jar CollectVariantCallingMetrics \
   I=CALLS.VCF O=VC_METRICS_OUT DBSNP=DBSNP.VCF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   --algo CollectVCMetrics -d DBSNP.VCF -v CALLS.VCF \
   VC_METRICS_OUT
```

Table 2.26: Argument correspondence - CollectVCMetrics and CollectVariantCallingMetrics

Picard option	Sentieon option	Meaning
I=CALLS.VCF	-v CALLS.VCF	vcf input file
O=VC_METRICS_OUT	N/A	Output basename
DBSNP=DBSNP.VCF	-d DBSNP.VCF	dbSNP file

2.16 ContEst - ContaminationAssessment

GATK Best Practices command line

```
java -jar GenomeAnalysisTK.jar -T ContEst -I TUMOR_RECALED.BAM \
   -R REFERENCE.FASTA -pf POPULATION.VCF --genotypes GENOTYPES.VCF \
   -o OUTPUT.TXT
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA -i TUMOR_RECALED.BAM \
   --algo ContaminationAssessment --pop_vcf POPULATION.VCF \
   --genotype_vcf GENOTYPES.VCF OUTPUT.TXT
```

Table 2.27: Argument correspondence - ContaminationAssessment and ContEst

GATK option	Sentieon option	Meaning
-I TUMOR_RECALED.BAM	-i TUMOR_RECALED.BAM	Input the bam files
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference fasta
-pf POPULATION.VCF	–pop_vcf POPULATION.VCF	The VCF file containing allele frequency information for the population
–genotypes GENOTYPES.VCF	–genotype_vcf GENOTYPES.VCF	The VCF file containing variants reported for the inidividual
-llc [META/SAMPLE/READGROUP]	–type [META/SAMPLE/READGROUP]	Assess contamination by sample, lane or all reads
–min_qscore MIN_BQ	–min_base_qual MIN_BQ	Any bases with a quality less than MIN_BQ will be ignored
–min_mapq MIN_MAPQ	–min_map_qual MIN_MAPQ	Any reads with a mapping quality less than MIN_MAPQ will be ignored
-mbc MINIMUM_BASE_COUNT	–min_basecount MINIMUM_BASE_COUNT	The minimum number of bases present at a locus for contamination to be assessed
–beta_threshold TRIM	–trim_thresh TRIM	Theshold that will be used to trim sites
–trim_fraction TRIM_FRACTION	–trim_frac TRIM_FRACTION	Maximum fraction of sites that may be trimmed
-pc PRECISION	–precision PRECISION	The precision on the output percent number
-br BASE_REPORT	–base_report BASE_REPORT	The output file that will contain an extended report on the processed data
-population POPULATION	–population POPULATION	A population for the baseline allele frequency of the sample
-o OUTPUT.TXT	N/A	The output file

2.17 DepthOfCoverage - CoverageMetrics

GATK Best Practices command

```
java -jar GenomeAnalysisTK.jar -T DepthOfCoverage \
   -R REFERENCE.FASTA -I DEDUPED.BAM \
   -geneList GENE_LIST.REFSEQ -ct THRESHOLD \
   -o OUTPUT_BASE
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo CoverageMetrics \
   --gene_list GENE_LIST.REFSEQ --cov_thresh THRESHOLD \
   OUTPUT_BASE
```

Table 2.28: Argument correspondence - CoverageMetrics and DepthOfCoverage

GATK option	Sentieon option	Meaning
-R REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
-I DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
-geneList GENE_LIST.REFSEQ	–gene_list GENE_LIST.REFSEQ	RefSeq file used to aggregate results to the gene level
–countType COUNT_TYPE	–count_type COUNT_TYPE	Determines how to deal with overlapping paried ends
-pt PARTITION	–partition PARTITION	Determines how to partition the data
-ct THRESHOLD	–cov_thresh THRESHOLD	Add aggregation metrics for the percentage of bases with coverage greater than THRESHOLD
–start MIN_DEPTH	–histogram_low MIN_DEPTH	The smallest histogram bin
–stop MAX_DEPTH	–histogram_high MAX_DEPTH	The largest histogram bin
–nBins NUM_BINS	–histogram_bin_count NUM_BINS	The number of histogram bins
-mmq MIN_MAPQ	–min_map_qual MIN_MAPQ	Minimum mapping quality of reads used
–maxMappingQuality MAX_MAPQ	–max_map_qual MAX_MAPQ	Maximum mapping quality of reads used
-mbq MIN_BASEQ	–min_base_qual MIN_BASEQ	Minimum base quality of bases used
–maxBaseQuality MAX_BASEQ	–max_base_qual MAX_BASEQ	Maximum base quality of bases used
-omitBaseOutput	–omit_base_output	Omit output of the per locus coverage
-omitSampleSummary	–omit_sample_stat	Omit output of the summary results
-omitLocusTable	–omit_locus_stat	Omit output of histogram files
-omitIntervals	–omit_interval_stat	Omit output of interval statistics
-baseConts	–print_base_counts	Include the number of "ACGTND" in the output per locus coverage
–includeRefNSites	–include_ref_N	Include coverage data in loci where the reference genome is set to N
–ignoreDeletionSites	–ignore_del_sites	Ignore coverage data in loci where there are deletions
-dels	–include_del	Include deletions and add deletion counts
-o OUTPUT_BASE	N/A	Output file basename

2.18 CollectGcBiasMetrics - GCBias

Picard CollectGcBiasMetrics command line

```
java -jar picard.jar CollectGcBiasMetrics \
   I=DEDUPED.BAM O=GC_METRICS.TXT CHART=GC_BIAS.PDF \
   S=SUMMARY.TXT R=REFERENCE.FASTA ASSUME_SORTED=true
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo GCBias --summary SUMMARY.TXT \
   GC_METRICS.TXT
sentieon plot GCBias -o GC_BIAS.PDF GC_METRICS.TXT
```

Table 2.29: Argument correspondence - GCBias and CollectGcBiasMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
O=GC_METRICS.TXT	N/A	GC bias metrics results
CHART=GC_BIAS.PDF	-o GC_BIAS.PDF	GC bias metrics report
S=SUMMARY.TXT	–summary SUMMARY.TXT	GC bias metrics summary results
LEVEL=LEVEL	–accum_level LEVEL	The accumulation level

2.19 CollectHsMetrics - HsMetricAlgo

Picard CollectHsMetrics command line

```
java -jar picard.jar CollectHsMetrics \
   I=DEDUPED.BAM O=HS_METRICS.TXT R=REFERENCE.FASTA \
   BAIT_INTERVALS=BAITS TARGET_INTERVALS=TARGETS
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo HsMetricAlgo --targets_list TARGETS \
   --baits_list BAITS HS_METRICS.TXT
```

Table 2.30: Argument correspondence - HsMetricAlgo and CollectHsMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
O=HS_METRICS.TXT	N/A	HS metrics results
BAIT_INTERVALS=BAITS	–baits_list BAITS	Interval list input file of baits
TARGET_INTERVALS=TARGETS	–targets_list TARGETS	Interval list input file of targets
CLIP_OVERLAPPING_READS	–clip_overlapping_reads	Clip overlapping reads
MINIMUM_MAPPING_QUALITY=MIN_MAPQ	–min_map_qual MIN_MAPQ	Minimum read mapping quality
MINIMUM_BASE_QUALITY=MIN_BASEQ	–min_base_qual MIN_BASEQ	Minimum base quality
COVERAGE_CAP=COVERAGE	–coverage_cap COVERAGE	Maximum coverage limit in the histogram

2.20 CollectInsertSizeMetrics - InsertSizeMetricAlgo

Picard CollectInsertSizeMetrics command line

```
java -jar picard.jar CollectInsertSizeMetrics \
   I=DEDUPED.BAM O=IS_METRICS.TXT R=REFERENCE.FASTA \
   H=IS_METRICS.PDF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo InsertSizeMetricAlgo \
   IS_METRICS.TXT
sentieon plot InsertSizeMetricAlgo -o IS_METRICS.PDF IS_METRICS.TXT
```

Table 2.31: Argument correspondence - InsertSizeMetricAlgo and CollectInsertSizeMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=IS_METRICS.TXT	N/A	IS metrics results
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
H=IS_METRICS.PDF	-o IS_METRICS.PDF	Insert size metrics report

2.21 MeanQualityByCycle - MeanQualityByCycle

Picard MeanQualityByCycle command line

```
java -jar picard.jar MeanQualityByCycle \
   I=DEDUPED.BAM O=MQ_METRICS.TXT R=REFERENCE.FASTA \
   CHART=MQ_METRICS.PDF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo MeanQualityByCycle \
   MQ_METRICS.TXT
sentieon plot MeanQualityByCycle -o MQ_METRICS.PDF MQ_METRICS.TXT
```

Table 2.32: Argument correspondence - MeanQualityByCycle and MeanQualityByCycle

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=MQ_METRICS.TXT	N/A	MQ metrics results
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
CHART=MQ_METRICS.PDF	-o MQ_METRICS.PDF	Mean quality metrics report

2.22 QualityScoreDistribution - QualDistribution

Picard QualityScoreDistribution command line

```
java -jar picard.jar QualityScoreDistribution \
   I=DEDUPED.BAM O=QD_METRICS.TXT \
   CHART=QD_METRICS.PDF
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo QualDistribution \
   QD_METRICS.TXT
sentieon plot QualDistribution -o QD_METRICS.PDF QD_METRICS.TXT
```

Table 2.33: Argument correspondence - QualDistribution and QualityScoreDistribution

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=QD_METRICS.TXT	N/A	QD metrics results
N/A	-r REFERENCE.FASTA	Reference file
CHART=QD_METRICS.PDF	-o QD_METRICS.PDF	Quality distribution metrics report

2.23 CollectQualityYieldMetrics - QualityYield

Picard CollectQualityYieldMetrics command line

```
java -jar picard.jar CollectQualityYieldMetrics \
   I=DEDUPED.BAM O=YIELD_METRICS.TXT
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo QualityYield \
   YIELD_METRICS.TXT
```

Table 2.34: Argument correspondence - QualityYield and CollectQualityYieldMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=YIELD_METRICS.TXT	N/A	Quality yield metrics results
INCLUDE_SECONDARY_ALIGNMENTS=true	--include_supplementary	Include supplementary alignments in the calculation
INCLUDE_SUPPLEMENTAL_ALIGNMENTS=true	--include_secondary	Include secondary alignments in the calculation

2.24 CollectSequencingArtifactMetrics - SequenceArtifactMetricsAlgo

Picard CollectSequencingArtifactMetrics command line

```
java -jar picard.jar CollectSequencingArtifactMetrics \
   I=DEDUPED.BAM O=ARTIFACT_METRICS_BASE R=REFERENCE.FASTA \
   DB_SNP=DBSNP.VCF
```

(continues on next page)

(continued from previous page)

```
java -jar picard.jar ConvertSequencingArtifactToOxoG \
   I=DEDUPED.BAM O=ARTIFACT_METRICS_BASE R=REFERENCE.FASTA \
   OUTPUT_BASE=oxog_metrics
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo SequenceArtifactMetricsAlgo \
   --dbsnp DBSNP.VCF ARTIFACT_METRICS_BASE
```

Table 2.35: Argument correspondence - SequenceArtifactMetricsAlgo and CollectSequencingArtifactMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=ARTIFACT_METRICS_BASE	N/A	Artifact metrics output base
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
DB_SNP=DBSNP.VCF	–dbsnp DBSNP.VCF	A dbSNP file to exclude known polymorphisms
Q=MIN_BQ	–min_base_qual MIN_BQ	Minimum base quality for a base to be included
MQ=MIN_MAPQ	–min_map_qual MIN_MAPQ	Minimum mapping quality for a read to be included
MIN_INS=MIN_INSERT	–min_insert_size MIN_INSERT	Minimum insert size to include a read
MAX_INS=MAX_INSERT	–max_insert_size MAX_INSERT	Maximum insert size to include a read
UNPAIRED=true	–include_unpaired	Include unpaired reads
TANDEM=true	–tandem_reads	Include tandem reads
INCLUDE_DUPLICATES=true	–include_duplicates	Include duplicate reads
INCLUDE_NON_PF_READS=true	–include_non_pf_reads	Include non-PF reads
CONTEXT_SIZE=CONTEXT	–context_size CONTEXT	The number of context bases to include on each size

2.25 CollectWgsMetrics - WgsMetricsAlgo

Picard CollectWgsMetrics command line

```
java -jar picard.jar CollectWgsMetrics \
   I=DEDUPED.BAM O=WGS_METRICS.TXT R=REFERENCE.FASTA
```

Sentieon command line

```
sentieon driver -t NUMBER_THREADS -r REFERENCE.FASTA \
   -i DEDUPED.BAM --algo WgsMetricsAlgo \
   WGS_METRICS.TXT
```

Table 2.36: Argument correspondence - WgsMetricsAlgo and CollectWgsMetrics

Picard option	Sentieon option	Meaning
I=DEDUPED.BAM	-i DEDUPED.BAM	Input the bam files
O=WGS_METRICS.TXT	N/A	WGS metrics results
R=REFERENCE.FASTA	-r REFERENCE.FASTA	Reference file
MQ=MIN_MAPQ	–min_map_qual MIN_MAPQ	Minimum mapping quality for a read to be included
Q=MIN_BQ	–min_base_qual MIN_BQ	Minimum base quality for a base to be included
CAP=COVERAGE_CAP	–coverage_cap COVERAGE_CAP	Maximum coverage limit for the histogram
COUNT_UNPAIRED=true	–include_unpaired true	Count unpaired reads and paired reads with one end unmapped
INCLUDE_BQ_HISTOGRAM=true	–base_qual_histogram true	Report a base quality histogram
SAMPLE_SIZE=SAMPLE_SIZE	–sample_size SAMPLE_SIZE	Sample size used for theoretical het sensitivity sampling

3 Other differences in usage

Sentieon refers to tools as algorithms, so the option `-T` in GATK3 corresponds to the option `--algo` in Sentieon.

Sentieon produces log files directly to stdout and stderr, so the option `-log` is not available.

Sentieon tries to use as many threads as the system has available, while GATK uses 1 thread by default. As such omitting option `-nt` in GATK, is not the same as omitting the option `-t` in Sentieon.

Sentieon does not do any down-sampling, so the following options are not available: `--downsample_to_coverage`, `--downsample_to_fraction`, `--downsampling_type`, `...`

Other general level arguments that are currently supported by Sentieon are:

- `--bam_compression`: for algorithms that output a bam

©Sentieon Inc.
465 Fairchild Drive, Suite 135, Mountain View CA 94043
www.sentieon.com