SNP INDEL calling

Haoxiang Lin

Abstract

Citation: Haoxiang Lin SNP INDEL calling. **protocols.io**
https://www.protocols.io/view/snp-indel-calling-grkbv4w
Published: 22 Jun 2017

Protocol

BWA ALN

Step 1.

BWA alignment

SOFTWARE PACKAGE (Linux)
BWA, 0.6.1
DATASET
Raw reads
cmd COMMAND

```
bwa aln  -n 3 -o 1 -e 50 -t 4 -I $REF $READ -f $SAI
```

EXPECTED RESULTS

SAI

BWA SAMPE

Step 2.

BWA SAMPE

SOFTWARE PACKAGE (Linux)
BWA, 0.6.1
DATASET
SAI and Raw reads
cmd COMMAND

```
bwa sampe -a $INSERTSIZE -r $READ_GROUP_INFO $SAI1 $SAI2 $READ1 $READ2 | samtools view -S -
b - -o $BAM
```

EXPECTED RESULTS

BAM

BAM sort

Step 3.

BAM sort

SOFTWARE PACKAGE (LINUX)
SAMtools, 0.1.18
DATASET
Sorted BAM
cmd COMMAND

```
samtools sort -m 3000000000 $BAM $BAM_SORT_PREFIX
```

EXPECTED RESULTS

Sorted BAM

BAM remove duplication

Step 4.

BAM remove duplication

SOFTWARE PACKAGE (LINUX)
SAMtools, 0.1.18
DATASET
Sorted BAM
cmd COMMAND

```
samtools rmdup $SORTED_BAM $SORTE_RMDUP_BAM
```

EXPECTED RESULTS

Sorted and duplication removed BAM

BAM markduplicate

Step 5.

BAM markduplicate

SOFTWARE PACKAGE (LINUX)
Picard, 1.61
DATASET
remove duplicate BAM
cmd COMMAND

```
java -jar picard-
tools-1.61/MarkDuplicates.jar I=$BAM O=$DEDUP_BAM M=$METRICS CREATE_INDEX=true VALIDATION_S
TRINGENCY=SILENT
```

EXPECTED RESULTS

Duplication marked BAM

Read realign

Step 6.

Read realign

SOFTWARE PACKAGE (LINUX)

GATK, 2.7

DATASET

Sorted.markdup.BAM

COMMAND

```
java  -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -nt 4 -R $HG19 -I $BAM -
o $intervals --known Mills_and_1000G_gold_standard.indels.hg19.sites.vcf --
known 1000G_phase1.indels.hg19.vcf

java -jar GenomeAnalysisTK.jar -T IndelRealigner -model USE_SW -LOD 0.4 -
known Mills_and_1000G_gold_standard.indels.hg19.sites.vcf -
known 1000G_phase1.indels.hg19.vcf -R $HG19 --targetIntervals $INTERVALS -I $BAM -
o $REALN_BAM
```

EXPECTED RESULTS

Realign BAM

BQSR

Step 7.

BQSR

SOFTWARE PACKAGE (LINUX)

GATK, 2.7

DATASET

Duplication marked BAM

COMMAND

```
java -jar GenomeAnalysisTK.jar   -T BaseRecalibrator --
knownSites Mills_and_1000G_gold_standard.indels.hg19.sites.vcf --
knownSites 1000G_phase1.indels.hg19.vcf --knownSites dbsnp_135.hg19.vcf -R $HG19 -I $BAM -
o $RECAL_FILE  -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 -L chrX -L chrY -L chrM -rf BadCigar

java -jar GenomeAnalysisTK.jar -T PrintReads -nct 6 -BQSR $RECAL_FILE  -R $HG19  -I $BAM -
o $RECAL_BAM
```

EXPECTED RESULTS

Recalibrate BAM

BAM merge

Step 8.

BAM merge

SOFTWARE PACKAGE (LINUX)

SAMtools, 0.1.18

DATASET

Recalibrate BAM

COMMAND

```
samtools merge -R $CHR -rh $READ_GROUP_INFO $MERGE_BAM $INPUT_BAM_LIST
```

BAM reduce

Step 9.

BAM reduce

SOFTWARE PACKAGE (LINUX)
GATK, 2.7
DATASET
Merged BAM and site VCF
COMMAND
```
java  -jar $GATK -R $HG19 -T ReduceReads -I $MERGE_BAM -o $REDUCE_BAM     -L $CHR
```

EXPECTED RESULTS

Redcude BAM

UnifiedGenotyper
Step 10.

UnifiedGenotyper

SOFTWARE PACKAGE (LINUX)
GATK, 2.7
COMMAND
```
java -jar GenomeAnalysisTK.jar  -R $HG19  -T UnifiedGenotyper -I $BAM_LIST --
dbsnp dbsnp_135.hg19.vcf -o $VCF -stand_call_conf 50.0 -stand_emit_conf 10.0  -dcov 40000 -
nt 1 -glm BOTH -A AlleleBalance -A HomopolymerRun -A InbreedingCoeff -A Coverage -
A HaplotypeScore -l INFO --max_alternate_alleles 4 -baqGOP 30 -L $REGION
```

EXPECTED RESULTS

VCF

VQSR
Step 11.

VQSR

COMMAND
```
java -jar $GATK  -l INFO -R $HG19 -T VariantRecalibrator -input $SNP_VCF -
resource:hapmap,known=false,training=true,truth=true,prior=15.0 hapmap_3.3.hg19.sites.vcf -
resource:omni,known=false,training=true,truth=false,prior=12.0 1000G_omni2.5.hg19.sites.vcf
 -resource:dbsnp,known=true,training=false,truth=false,prior=8.0 dbsnp_135.hg19.vcf -
an HaplotypeScore -an ReadPosRankSum -an FS -recalFile $RECAL_FILE -
tranchesFile $TRANCHES_FILE -rscriptFile ./GATK.SNP.plot.R --TStranche 90.0 --
TStranche 93.0 --TStranche 95.0 --TStranche 97.0 --TStranche 99.0 --TStranche 100.0 -
mode SNP

java -jar $GATK  -l INFO -R $HG19 -T VariantRecalibrator -input $INDEL_VCF -
resource:mills,VCF,known=true,training=true,truth=true,prior=12.0 Mills_and_1000G_gold_stan
dard.indels.hg19.sites.vcf -
resource:mills,VCF,known=true,training=true,truth=true,prior=12.0 1000G_phase1.indels.hg19.
vcf -an FS -an HaplotypeScore -an ReadPosRankSum -an MQRankSum --maxGaussians 4 -std 10.0 -
percentBad 0.12 -recalFile $RECAL_FILE -tranchesFile $TRANCHES_FILE -
rscriptFile $bin/GATK.INDEL.plot.R --TStranche 90.0 --TStranche 93.0 --TStranche 95.0 --
TStranche 97.0 --TStranche 99.0 --TStranche 100.0 -mode INDEL
```

```
java -jar $GATK  -R $HG19 -T ApplyRecalibration -input $SNP_VCF --ts_filter_level 99.0 -
recalFile $RECAL_FILE -tranchesFile $TRANCHES_FILE -mode SNP -o $SNP_VQSR_VCF

java -jar $GATK  -R $HG19 -T ApplyRecalibration -input $INDEL_VCF --ts_filter_level 95.0 -
recalFile $RECAL_FILE -tranchesFile $TRANCHES_FILE -mode INDEL -o $INDEL_VQSR_VCF

# Following, use in-
house script to keep variant which pass VQSR filter and remove variant with HRUN > 6(for SN
P) or HRUN > 10 (for INDEL)
```

Impuation

Step 12.

Impuation

SOFTWARE PACKAGE (LINUX)

BEAGLE, v3

DATASET

VCF

COMMAND

```
java -Xmx2g -jar beagle.jar nthreads=4  window=3000 overlap=600 impute-
its=10  gl=$VCF out=$OUT_VCF
```

EXPECTED RESULTS

Imputation VCF