HiSeq Human DNA Resequencing Data Analysis Protocols

Call Variants

文档作者	叶永鑫
提交日期	2012-03-14
更新说明	
文档备注	未完善，需验证

目的：call variants
输入：sorted mapping 结果 bam 文件
库文件：ref.fastn, dbsnp.vcf,

假设已经得到 mapping 结果 bam 文件

（一）用 samtools、bcftools、vcfutils call variants：
1. filter bam 文件：
留下 unique mapping（对于 bwa 的结果，grep XT:A:U）；
滤去 duplicate（Picard-MarkDuplicates）；
（PE）留下 proper paired；
（PE）留下 insertion size 合理的；（与上面那个有重复？bwa –a 选项？）
（测通的）合并或截短（依赖于建库 fragment 长度，exome seq 有不少测通的 PE reads）；
（RNA）留下链方向正确的

1.5. realign? recalibration?

2. samtools、bcftools call variants：（http://samtools.sourceforge.net/mpileup.shtml）

```
samtools mpileup -uf ref.fa aln1.bam aln2.bam | bcftools view -bvcg - > var.raw.bcf
bcftools view var.raw.bcf | vcfutils.pl varFilter –D500 –d 50 > var.flt.vcf
#  如果平均深度为 50X
```

samtools call genotype 不准，仅作辅助作用

（二）用 GATK call variants：
1. MarkDuplicates, Realign, Recalibration（顺序？分 lane 分 sample？）
1.1. 先筛出 unique mapping；
1.2. Picard-MarkDuplicates

```
run_picard.sh    MarkDuplicates    I=read.add_rg.bam    O=read.rmdup.bam
M=read.rmdup.matrix REMOVE_DUPLICATES=true    CREATE_INDEX=true
rm read.rmdup.matrix
```

1.3. GATK-Realign

```
run_gatk.sh   -T RealignerTargetCreator   -R ref.fa   -I read.rmdup.bam   -o
read.rmdup.realign.intervals   -known dbsnp.vcf
run_gatk.sh   -T IndelRealigner   -R ref.fa   -I read.rmdup.bam   -targetIntervals
read.rmdup.realign.intervals   -o read.rmdup.realign.bam
rm read.rmdup.realign.intervals
```

Picard - fix mate information?

1.4. GATK-Recalibration

```
run_gatk.sh   -T CountCovariates   -R ref.fa   -I read.rmdup.realign.bam   -cov
ReadGroupCovariate   -cov QualityScoreCovariate   -cov CycleCovariate   -cov DinucCovariate
-knownSites dbsnp_132.vcf   -recalFile read.rmdup.realign.recal.csv   [--standard_covs]
```

（--standard_covs 与几个-cov 等价）

```
run_gatk.sh   -T TableRecalibration   -R ref.fa   -I read.rmdup.realign.bam    -recalFile
read.rmdup.realign.recal.csv   -o read.rmdup.realign.recal.bam
rm read.rmdup.realign.recal.csv
rm read.rmdup.bam read.rmdup.realign.bam
```

2. UnifiedGenotyper

（http://www.broadinstitute.org/gsa/gatkdocs/release/org_broadinstitute_sting_gatk_walkers_genotyper_UnifiedGenotyper.html）

```
run_gatk.sh   -T UnifiedGenotyper   -R ref.fa   -I read.prepared.bam    -glm BOTH   --dbsnp
dbsnp.vcf   -stand_call_conf 30   -stand_emit_conf 30   -o raw.vcf   [-alleles dbsnp.vcf]
```

或

```
    run_gatk.sh   -T UnifiedGenotyper   -nt 2   -R ref.fa   -I read.prepared.bam   -D[--dbsnp]
dbsnp.vcf   -glm SNP –mbq[--min_base_quality_score] 20   -hets[--heterozygosity] 0.001   -l
INFO   -A[--annotation] AlleleBalance   -A DepthOfCoverage   -stand_call_conf 30
-stand_emit_conf 10   -dcov 200   -o raw.SNP.vcf
run_gatk.sh   -T UnifiedGenotyper   -nt 2   -R ref.fa   -I read.prepared.bam   -D dbsnp.vcf   -glm
INDEL -mbq 20   -indelHeterozygosity 0.000125   -l INFO   -A AlleleBalance   -A
DepthOfCoverage   -stand_call_conf 30   -stand_emit_conf 10   -dcov 200   -o raw.indel.vcf
```

（一起 call 较快）

或

```
run_gatk.sh   -T UnifiedGenotyper   -R ref.fa   -I read.prepared.bam   -glm BOTH
[-B:alleles,VCF dbsnp.vcf –BTI alleles] -B:dbsnp,VCF dbsnp.vcf   -stand_call_conf 50
-stand_emit_conf 10   -dcov 1000   --min_base_quality_score 30   -A DepthOfCoverage   -A
AlleleBalance   -o raw.vcf   -metrics raw.metrics
```

3. Variant Select

```
run_gatk.sh   -T SelectVariants   -R ref.fa    --variant raw.vcf   -selectType SNP   -selectType
MNP   -o raw.snp.vcf
run_gatk.sh   -T SelectVariants   -R ref.fa    --variant raw.vcf   -selectType INDEL   -o
raw.indel.vcf
```

4. Variant Filtration

```
run_gatk.sh    -T VariantFiltration -R ref.fa    --variant recal.SNP.vcf    -o flt.SNP.vcf
--clusterWindowSize 10    --filterExpression "MQ0>=4&&((MQ0/(1.0*DP))>0.1)"    --filterName
"HARD_TO_VALIDATE"
run_gatk.sh    -T VariantFiltration    -R ref.fa    --variant raw.indel.vcf    -o flt.indel.vcf
--filterExpression "MQ0>=4&&((MQ0/(1.0*DP))>0.1)"    --filterName "HARD_TO_VALIDATE"
--filterExpression "QUAL<10"    --filterName "QualFilter"
```

或

```
run_gatk.sh    -T VariantFiltration    --clusterWindowSize 10    --filterExpression
"MQ0>=4&&((MQ0/(1.0*DP))>0.1)"    --filterName "HARD_TO_VALIDATE"    --filterExpression
"DP<10" --filterName "LowCoverage"    --filterExpression "QUAL<30.0"    --filterName
"VeryLowQual" --filterExpression "QUAL>30.0&&QUAL<50.0"    --filterName "LowQual"
--filterExpression "QD<5.0" --filterName "LowQD"    --filterExpression "SB>-0.10"    --filterName
"StrandBias"    -B:mask,VCF raw.indel.vcf    --maskExtension 0    --maskName Indel    -R ref.fa
-B:variant,VCF recal.snp.vcf    -o flt.snp.vcf
```

（对 exome seq 不推荐 Qual 过滤）

或

```
run_gatk.sh    -T VariantFiltration    --filterExpression "QD < 2.0 || ReadPosRankSum < -20.0 ||
FS > 200.0"    --filterName GATKStandard    -R ref.fa    --variant raw.indel.vcf    -o flt.indel.vcf
```

（三）Report

Ti、Tv

VariantEvalu

QC countLoci countPairs ...

GATK-Phasing，对于家系重要

GATK-somaticIndelDetector   dIndel？