Title: "SNP一般流程"
author: "xsc"
date: "2021/6/25"
output:
word_document: default
html_document: default

本文为日常操作所写笔记，错误之处望各位指教。内容可能与其他创作者存在冲突，但是目前流程就是如此，不存在抄袭。

1 比对

1.1 构建参考基因组索引

```
cd 00ref
bwa index IRGSP-1.0_genome.fasta
```

1.2 比对

```
#!/bin/bash
List=./list.txt
Picard=/home/xushichang/tools/picard.jar
Refence=/home/xushichang/000/00ref/IRGSP-1.0_genome.fasta
cat $List | while read lind
do
  bwa mem -t 12 \
    ../01ref/IRGSP-1.0_genome.fasta \
    ../${lind}.1_1_output.fastq \
    ../${lind}.1_2_output.fastq > ./${lind}.sam   #比对
  samtools view -bS ${lind}.sam -o ${lind}.bam     #sam转换为bam
  java -jar ${Picard} SortSam
    INPUT=${lind}.bam \
    OUTPUT=${lind}_sort.bam \
    SORT_ORDER=coordinate  #bam文件按染色体组型排序
done
samtools merge -f SRR363063_all.bam SRR3630632_sort.bam SRR3630633_sort.bam
SRR3630634_sort.bam SRR3630635_sort.bam SRR3630636_sort.bam SRR3630637_sort.bam    #
合并
samtools index SRR363063_all.bam     #合并后索引
```

使用smtools软件的flagstat工具生成bam文件的统计比对信息：

```
samtools flagstat SRR363063_all.bam

```

```
36522266 + 0 in total (QC-passed reads + QC-failed reads)
#通过QC的reads的数量是36522266，未通过QC的reads的数量为0，以为着一共有36522266条reads
0 + 0 secondary
93974 + 0 supplementary
0 + 0 duplicates
35780157 + 0 mapped (97.97% : N/A)
# 总体上reads的匹配率
36428292 + 0 paired in sequencing
# 双端reads数
18214146 + 0 read1
#read1中reads数
18214146 + 0 read2
#read2中reads数
33347066 + 0 properly paired (91.54% : N/A)
#完美匹配的reads数：比对到同一条参考序列，并且两条reads之间的距离符合设置的阈值
35461496 + 0 with itself and mate mapped
#paired reads中两条都比对到参考序列上的reads数
224687 + 0 singletons (0.62% : N/A)
#单独一条匹配到参考序列上的reads数，和上一个相加，则是总的匹配上的reads数。
1131220 + 0 with mate mapped to a different chr
#paired reads中两条分别比对到两条不同的参考序列的reads数
552776 + 0 with mate mapped to a different chr (mapQ>=5)
#同上一个，只是其中比对质量>=5的reads的数量
```

2 SNP

Work list:/home/xushichang/000/2.SNP

2.1 添加头文件

```
java -jar /home/xushichang/tools/picard.jar AddOrReplaceReadGroups \
    CN=BGI \
    CREATE_INDEX=TRUE \
    RGPL=illumina \
    SM=rice \
    SO=coordinate \
    RGLB=SRR363063 \
    RGID=SRR363063 \
    RGPU=SRR363063 \
    VALIDATION_STRINGENCY=LENIENT \
    I=../1.mapping/SRR363063_all.bam \
    O=/home/xushichang/000/2.SNP/SRR363063_all_arrg.bam
```

2.2 分析bam文件的碱基质量。

```
samtools faidx IRGSP-1.0_genome.fasta    #建立faidx索引
gatk CreateSequenceDictionary -R IRGSP-1.0_genome.fasta -O IRGSP-1.0_genome.dict  #生成参考基因组的dict文件
gatk BaseRecalibrator \
    -R ../00ref/IRGSP-1.0_genome.fasta \
    -I SRR363063_all_arrg.bam \
    --known-sites /home/xushichang/000/00ref/all_snps.vcf \
    -O SRR363063_all_date.table
```

2.3 bam文件质量校准(Base Quality Score Recalibration (BQSR) #2)

```
gatk ApplyBQSR \
    -R ../00ref/IRGSP-1.0_genome.fasta \
    -I SRR363063_all_arrg.bam \
    --bqsr-recal-file SRR363063_all_date.table \
    -O SRR363063_all_recal.bam

```

2.4 调用变体(Call Variants)

第一轮变异调用。在此步骤中识别的变体将被过滤并作为基础质量得分重新校准 (BQSR) 的输入提供

```
gatk HaplotypeCaller \
        -R ../00ref/IRGSP-1.0_genome.fasta \
        -I SRR363063_all_recal.bam \
        -O raw_variants.vcf
```

2.5 提取 SNP 和插入缺失(Extract SNPs & Indels)

```
    gatk SelectVariants \
        -R ../00ref/IRGSP-1.0_genome.fasta \
        -V raw_variants.vcf \
        -select-type SNP \
        -O raw_snps.vcf
```

```
gatk SelectVariants \
        -R ../00ref/IRGSP-1.0_genome.fasta \
        -V raw_variants.vcf \
        -select-type INDEL \
        -O raw_indels.vcf
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  rice
chr01   1337    .       G       A       27.94   .       AC=1;AF=0.500;AN=2;BaseQRankSum=0.967;DP=5;ExcessHet=3.0103;FS=0.
000;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=9.31;ReadPosRankSum=0.967;SOR=1.179       GT:AD:DP:GQ:PL  0/1:1,2:3:
14:56,0,14
chr01   1708    .       C       T       34.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=-0.719;DP=9;ExcessHet=3.0103;FS=1
0.792;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=3.86;ReadPosRankSum=1.383;SOR=2.206     GT:AD:DP:GQ:PL  0/1:7,2:9:
63:63,0,711
chr01   1729    .       A       G       28.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=0.875;DP=11;ExcessHet=3.0103;FS=1
2.632;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=2.62;ReadPosRankSum=0.980;SOR=2.203     GT:AD:DP:GQ:PL  0/1:9,2:11
:57:57,0,896
chr01   1733    .       C       A       28.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=0.431;DP=9;ExcessHet=3.0103;FS=15
.563;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=3.20;ReadPosRankSum=0.000;SOR=2.199      GT:AD:DP:GQ:PL  0/1:7,2:9:
57:57,0,896
chr01   1741    .       G       A       34.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=-1.025;DP=9;ExcessHet=3.0103;FS=1
5.563;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=3.86;ReadPosRankSum=-0.812;SOR=2.199    GT:AD:DP:GQ:PL  0/1:7,2:9:
63:63,0,789
chr01   2282    .       T       C       68.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=1.150;DP=8;ExcessHet=3.0103;FS=7.
068;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=8.60;ReadPosRankSum=-0.956;SOR=2.807      GT:AD:DP:GQ:PL  0/1:5,3:8:
97:97,0,269
chr01   2284    .       G       A       68.77   .       AC=1;AF=0.500;AN=2;BaseQRankSum=0.431;DP=7;ExcessHet=3.0103;FS=3.
680;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=9.82;ReadPosRankSum=-0.366;SOR=2.258      GT:AD:DP:GQ:PL  0/1:5,2:7:
:
```

前五列信息为

\1. 染色体(Chromosome)

\2. 起始位置(Start)

\3. 结束位置(End)

\4. 参考等位基因(Reference Allele)

\5. 替代等位基因(Alternative Allele)

ANNOVAR注释时主要也是利用前五列信息对数据库进行比对，注释变异。Info和Format信息同样重要，比如DP代表测序深度，这些内容的含义再vcf文件的开头都有介绍，请仔细阅读并理解相应内容的意义。

2.6 过滤SNP (Filter SNPs)

```
gatk VariantFiltration \
        -R ../00ref/IRGSP-1.0_genome.fasta \
        -V raw_snps.vcf \
        -O filtered_snps.vcf \  #SNP结果
        -filter-name "QD_filter" -filter "QD < 2.0" \
        -filter-name "FS_filter" -filter "FS > 60.0" \
        -filter-name "MQ_filter" -filter "MQ < 40.0" \
        -filter-name "SOR_filter" -filter "SOR > 4.0" \
        -filter-name "MQRankSum_filter" -filter "MQRankSum < -12.5" \
        -filter-name "ReadPosRankSum_filter" -filter "ReadPosRankSum < -8.0"
```

QD,描述单位深度的变异值，越大可信度越高。一般过滤掉<2的值。

FS，描述正负链特异性，差异性较大，说明测序或组装的过程中不够随机。FS越小越好。一般过掉掉>40（严格）或60（普通）

MQ 使用bwa-mem的话，正常值应该是60，描述某个位点测序reads的质量值的离散程度。

MQ< 40.0

MQRankSum < -12.5

SOR，也是表示正负链特异性，正常值在0-3，过滤掉>3的值。

2.7 过滤插入缺失 (Filter Indels)

```
gatk VariantFiltration \
        -R ../00ref/IRGSP-1.0_genome.fasta \
        -V raw_indels.vcf \
        -O filtered_indels.vcf \
        -filter-name "QD_filter" -filter "QD < 2.0" \
        -filter-name "FS_filter" -filter "FS > 200.0" \
        -filter-name "SOR_filter" -filter "SOR > 10.0"
```

2.8 注释SNP并预测（Annotate SNPs and Predict Effects）

```
java -Xmx8g -jar /home/samuel/tools/snpEff/snpEff.jar Oryza_sativa filtered_snps.vcf
> filtered_snps.eff.vcf
```