***一、Preparation:***

1. 下载安装：<https://software.broadinstitute.org/gatk/download/>
2. 参考基因组获取：NCBI/ensembl/UCSC…..

***二、Data Pre-processing:***

对应链接：<https://github.com/gatk-workflows/gatk4-data-processing>

1. **Map to reference（将原始序列和参考基因组进行比对）：**

①Indexing reference:

$bwa index-a bwtsw ref.fa  
\*参数-a：  
1）bwtsw ：>10M  
2）is ：<2G

是否指定-a参数是可选的，bwa索引会根据基因组大小自动选择适当的索引方法

②Indexing GATK:

$ samtools faidx ref.fa

\*使用Samtools为参考序列创建一个索引，为GATK快速检索fasta上的任何序列做好准备

③Generate the.dict file:

$ gatk CreateSequenceDictionary-R reference.fa -O reference.dict

④Sequence alignment:

$ bwa mem ref.fa sample\_1.fq sample\_2.fq -R'@RG\tID:sample\tLB:sample\tSM:sample\tPL:ILLUMINA'\  
2>sample\_map.log | samtools sort -@ 20 -O bam -o sample.sorted.bam 1>sample\_sort.log 2>&1

1)-R 选项为必须选项，用于定义头文件中的SAM/BAM文件中的read group和sample信息  
\*ID：输入reads集的ID号； LB： reads集的文库名； SM：样本名称； PL：测序平台

2)查看SAM/BAM文件中的read group和sample信息：  
$ samtools view -H /path/to/my.bam | grep '^@RG'

3)read group信息不仅会加在头信息部分，也会在比对结果的每条记录里添加一个 RG:Z:\*标签:$ samtools view /path/to/my.bam | grep '^@RG'

4)若原始SAM/BAM文件没有read group和sample信息，可以通过AddOrReplaceReadGroups添加这部分信息:$ java -jar picard.jar AddOrReplaceReadGroups

1. **Mask duplicates：**

①SortSam:  
$ gatk SortSam -I mapping/T.chr17.sam -O preprocess/T.chr17.sort.bam -R database/chr17.fa -SO coordinate --CREATE\_INDEX  
$ java -jar picard.jar SortSam \  
I=input.bam \  
O=sorted.bam \  
SORT\_ORDER=coordinate

②Markduplicates:  
gatk MarkDuplicates -I preprocess/T.chr17.sort.bam –O preprocess/T.chr17. markdup.bam -M preprocess/T.chr17.metrics --CREATE\_INDEX

**3、Base (Quality Score) Recalibration:**

①Establish a model：  
# Download the VCF file of known-site on Ensembl  
$ wget -c -P Ref/mouse/mm10/vcf ftp://ftp.ensembl.org/pub/release-93/variation/vcf/mus\_musculus/mus\_musculus.vcf.gz >download.log &  
$ cd Ref/mouse/mm10/vcf && gunzip mus\_musculus.vcf.gz && mv mus\_musculus.vcf dbsnp\_150.mm10.vcf  
# Index VCF files，need IndexFeatureFile  
$ gatk IndexFeatureFile -F dbsnp\_150.mm10.vcf  
# Establish the model  
$ gatk BaseRecalibrator -R Ref/mouse/mm10/bwa/mm10.fa -I PharmacogenomicsDB/mouse/SAM/ERR118300.enriched.markdup.bam -O \  
PharmacogenomicsDB/mouse/SAM/ERR118300.recal.table --known-sites Ref/mouse/mm10/vcf/dbsnp\_150.mm10.vcf

1. Quality Score Recalibration:

$ gatk ApplyBQSR -R Ref/mouse/mm10/bwa/mm10.fa -I PharmacogenomicsDB/mouse /SAM /ERR118300.enriched.markdup.bam -bqsr \

PharmacogenomicsDB/mouse/SAM/ ERR118300. recal.table -O PharmacogenomicsDB/ mouse /SAM/ERR118300.recal.bam

***三、Joint calling:***

**1、Call variants per-sample:Separately generate the intermediate file -- gVCF file -- required for subsequent analysis for each sample**

$ gatk HaplotypeCaller -R Ref/chr17.fa -I sam/T.chr17.recal.bam -ERC GVCF --dbsnp ../Ref/VCF/dbsnp\_138.hg19.vcf \

-O calling/T.chr17.raw.snps.indels.vcf -L chr17:7400000-7800000

**2、Consolidate GVCFs:Merge multiple GVCF files to get GenomicsDB for joint genotyping**

$ gatk --java-options "-Xmx4g -Xms4g" GenomicsDBImport \

-V data/gvcfs/mother.g.vcf.gz \

-V data/gvcfs/father.g.vcf.gz \

-V data/gvcfs/son.g.vcf.gz \

--genomicsdb-workspace-path my\_database \

-L 20

**3、Joint-Call Cohort**

$ gatk --java-options "-Xmx4g" GenotypeGVCFs \  
 -R Homo\_sapiens\_assembly38.fasta \  
 -V gendb://my\_database \  
 -O output.vcf.gz

***四、Filter***

**1、Extract the SNPs from the call set**

$ gatk SelectVariants -R Ref/chr17.fa -V calling/T.chr17.raw.snps.indels.genotype. vcf \

--select-type-to-include SNP -O filter/T.chr17.raw.snps.genotype.vcf

1. **Apply the filter to the SNP call set**

$ gatk VariantFiltration -R Ref/chr17.fa -V filter/T.chr17.raw.snps.genotype.vcf --filter- expression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || SOR > 3.0 || MQRankSum < -12.5 || \  
ReadPosRankSum < -8.0" --filter-name "SNP\_FILTER" -O filter/T.chr17.filter.snps.genotype.vcf

1. **Extract the Indels from the call set**

$ gatk SelectVariants -R Ref/chr17.fa -V calling/T.chr17.raw.snps.indels.genotype.vcf \  
--select-type-to-include INDEL -O filter/T.chr17.raw.indels.genotype.vcf

1. **Apply the filter to the Indel call set**

$ gatk VariantFiltration -R Ref/chr17.fa -V filter/T.chr17.raw.indels.genotype.vcf –filter -expression "QD < 2.0 || FS > 200.0 || SOR > 10.0 || MQRankSum < -12.5 || ReadPosRankSum < \

-20.0" --filter-name "INDEL\_FILTER" -O filter/T.chr17.filter.indels.genotype.vcf

1. **Combine SNP and indel call set**

$ gatk MergeVcfs -I filter/T.chr17.filter.snps.genotype.vcf -I \

filter/T.chr17.filter.indels.genotype.vcf -O filter/T.chr17.filter.snps.indels.genotype.vcf

1. **Get passed call set**

$ gatk SelectVariants -R Ref/chr17.fa -V filter/T.chr17.filter.snps.indels.genotype.vcf -O  
T.chr17.pass.snps.indels.genotype.vcf -select "vc.isNotFiltered()"