(BWA-PICARD-GATK variant calling pipeline)

\*\*This protocol is optimized for influenza A virus genome data

|  |  |
| --- | --- |
| pre STEP 1 | Indexing Reference file |
| TOOL | bwa/intel/0.7.15 |
| INPUT | reference.fasta (reference genome file) |
| OUTPUT | reference.fasta(or fa).ann; reference.fasta.pac; reference.fasta.amb; reference.fasta.bwt; reference.fasta.sa |
| COMMAND | ① load module  module load bwa/intel/0.7.15  ② Indexing  bwa index reference.fasta |

|  |  |
| --- | --- |
| pre STEP 2 | Creating dictionary file of the reference |
| TOOL | picard/2.8.2 |
| INPUT | reference.fasta (reference genome file) |
| OUTPUT | reference.fasta(or fa).dict |
| COMMAND | ① load module  module load picard/2.8.2  ② creating dictionary  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar CreateSequenceDictionary R=reference.fasta O=reference.fasta.dict |
| NOTES | ① Documentation link  https://software.broadinstitute.org/gatk/documentation/article?id=1601  ② “O =reference.fasta.dict” (O - uppercase) |

|  |  |
| --- | --- |
| STEP 1 | Alignment – Map to Reference |
| TOOL | bwa/intel/0.7.15 |
| INPUT | $i.trimmed.r1.fastq; $i.trimmed.r2.fastq; reference.fasta(or fa) (reference genome file) |
| OUTPUT | $i.aligned\_reads.sam |
| COMMAND | ① load module  module load bwa/intel/0.7.15  ② alignment  bwa mem -M reference.fasta $i.trimmed.r1.fastq $i.trimmed.r2.fastq > $i.aligned\_reads.sam |
| Loop command | ① creating a shell script file  vim GATK-STEP1.sh  ② writing commands (press ‘I’ and input below)  \*\* This part is only valid on NYU HPC server (Prince)  \*\* Replace this part with the job-submission commands for your own HPC server  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP1  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP1.err  cd /path/to/the/files  module load bwa/intel/0.7.15  for i in {start#..end#};do  bwa mem -M reference.fasta $i.trimmed.r1.fastq $i.trimmed.r2.fastq > $i.aligned\_reads.sam  done |
| NOTES | $i - sample name ( or number ) |

|  |  |
| --- | --- |
| STEP 2 | Sort SAM file by coordinate + convert to BAM |
| TOOL | picard/2.8.2 |
| INPUT | .aligned\_reads.sam |
| OUTPUT | .sorted\_reads.bam |
| COMMAND | ① load module  module load picard/2.8.2  ② alignment  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar SortSam INPUT=$i.aligned\_reads.sam OUTPUT=$i.sorted\_reads.bam\  SORT\_ORDER=coordinate |
| Loop command | ① creating a shell script file  vim GATK-STEP2.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP2  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP2.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar SortSam INPUT=$i.aligned\_reads.sam OUTPUT=$i.sorted\_reads.bam\  SORT\_ORDER=coordinate  done |
| NOTE |  |

|  |  |
| --- | --- |
| STEP 3 | Collect Alignment & Insert Size Metrics |
| TOOL | ① picard/2.8.2  ② r/intel/3.3.2  ③ samtools/intel/1.3.1 |
| INPUT | $i.sorted\_reads.bam; reference.fasta(or fa) (reference genome file) |
| OUTPUT | ① $i.alignment\_metrics.txt  ② $i.insert\_metrics.txt  ③ $i.insert\_size\_histogram.pdf |
| COMMAND | ① load module  module load picard/2.8.2  module load r/intel/3.3.2  module load samtools/intel/1.3.1  ② Collecting Alignment Summary Metrics  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar CollectInsertSizeMetrics I=$i.sorted\_reads.bam\  O=$i.alignment\_metrics.txt HISTOGRAM\_FILE =$i.insert\_size\_histogram\_1.pdf M=0.5  ③ Collect Insert Size Metrics  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar CollectInsertSizeMetrics INPUT=$i.sorted\_reads.bam\  OUTPUT=$i.insert\_metrics.txt HISTOGRAM\_FILE =$i.insert\_size\_histogram\_2.pdf M=0.5  ④ Depth size  samtools depth -a $i.sorted\_reads.bam > $i.depth\_out.txt |
| Loop command | ① creating a shell script file  vim GATK-STEP3.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP3  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP3.err  cd /path/to/the/files  module load picard/2.8.2  module load r/intel/3.3.2  module load samtools/intel/1.3.1  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar CollectInsertSizeMetrics I=$i.sorted\_reads.bam\  O=$i.alignment\_metrics.txt HISTOGRAM\_FILE =1.insert\_size\_histogram.pdf M=0.5;  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar CollectInsertSizeMetrics INPUT=$i.sorted\_reads.bam\  OUTPUT=$i.insert\_metrics.txt HISTOGRAM\_FILE =$i.insert\_size\_histogram\_2.pdf M=0.5;  samtools depth -a $i.sorted\_reads.bam > $i.depth\_out.txt;  done |
| NOTES | load 3 modules |

|  |  |
| --- | --- |
| STEP 4 | Mark Duplicates |
| TOOL | picard/2.8.2 |
| INPUT | $i.sorted\_reads.bam |
| OUTPUT | ① $i.dedup\_reads.bam  ② $i.metrics.txt |
| COMMAND | ① load module  module load picard/2.8.2  ② Mark duplicate  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar MarkDuplicates\  INPUT=$i.sorted\_reads.bam OUTPUT=$i.dedup\_reads.bam\  METRICS\_FILE=$i.metrics.txt |
| Loop command | ① creating a shell script file  vim GATK-STEP4.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP4  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP4.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar MarkDuplicates\  INPUT=$i.sorted\_reads.bam OUTPUT=$i.dedup\_reads.bam\  METRICS\_FILE=$i.metrics.txt  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 5 | Build BAM Index |
| TOOL | picard/2.8.2 |
| INPUT | $i.dedup\_reads.bam |
| OUTPUT | ① $i.dedup\_reads.bai |
| COMMAND | ① load module  module load picard/2.8.2  ② Build BAM Index  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=$i.dedup\_reads.bam |
| Loop command | ① creating a shell script file  vim GATK-STEP5.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP5  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP5.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=.dedup\_reads.bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 6 | Add Read Group |
| TOOL | picard/2.8.2 |
| INPUT | .dedup\_reads.bam |
| OUTPUT | ① .RDGR.bam |
| COMMAND | ① load module  module load picard/2.8.2  ② Add or Replace Read Group  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar AddOrReplaceReadGroups\  I=$i.dedup\_reads.bam O=$i.RDGR.bam RGID=4 RGLB=lib1 RGPL=illumine RGPU=unit1 RGSM=20 |
| Loop command | ① creating a shell script file  vim GATK-STEP5.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=STEP6  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=STEP6.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar AddOrReplaceReadGroups\  I=.dedup\_reads.bam O=.RDGR.bam RGID=4 RGLB=lib1 RGPL=illumine RGPU=unit1 RGSM=20  done |
| NOTES | ①Documentation link  <http://broadinstitute.github.io/picard/command-line-overview.html#AddOrReplaceReadGroups>  ②“O=$i.RDGR.bam” (O - uppercase) |

|  |  |
| --- | --- |
| STEP 7 | Build Read Group BAM Index |
| TOOL | picard/2.8.2 |
| INPUT | $i.RDGR.bam |
| OUTPUT | $i.RDGR.bai |
| COMMAND | ① load module  module load picard/2.8.2  ② Build Read Group BAM Index  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=$i.RDGR.bam |
| Loop  command | ① creating a shell script file  vim GATK-STEP5.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP7  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP7.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=$i.RDGR.bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 8 | Creat Realignment Targets |
| TOOL | gatk/3.7-0 |
| INPUT | $i.RDGR.bam; reference.fasta (reference genome file) |
| OUTPUT | ① $i.realignment\_targets.list |
| COMMAND | ① load module  module load gatk/3.7-0  ② Creat Realignment Tatgets  java -jar /share/apps/gatk/3.7-0/GenomeAanlysisTK.jar -T RealignerTargetCreator -R reference.fasta -I $i.RDGR.bam\  -o $i.realignment\_targets.list |
| Loop  command | ① creating a shell script file  vim GATK-STEP8.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP8  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP8.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAanlysisTK.jar -T RealignerTargetCreator -R reference.fasta -I $i.RDGR.bam\  -o $i.realignment\_targets.list  done |
| NOTES | ① Input should be RDGRed BAM files  ②“-o .realignment\_targets.list ” (o: lowercase)  ③ dictionary file of the reference genome is needed (\*\*pre-STEP2 should precede!!) |

|  |  |
| --- | --- |
| STEP 9 | Realign Indels |
| TOOL | gatk/3.7-0 |
| INPUT | $i.realignment\_target.list; $i.RDGR.bam; reference.fasta(or fa) (reference genome file) |
| OUTPUT | ① $i.realigned\_reads.bam |
| COMMAND | ① load module  module load gatk/3.7-0  ② Realign Indels  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T IndelRealigner -R h3n2.fa -I $i.RDGR.bam\  -targetIntervals $i.realignment\_targets.list -o $i.realigned\_reads.bam |
| Loop  command | ① creating a shell script file  vim GATK-STEP9.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP9  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP9.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T IndelRealigner -R h3n2.fa -I $i.RDGR.bam\  -targetIntervals $i.realignment\_targets.list -o $i.realigned\_reads.bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 10 | Add Read Group to $i.realigned\_reads.bam |
| TOOL | picard/2.8.2 |
| INPUT | $i.realigned\_reads.bam |
| OUTPUT | $i.RDGR\_realinged\_reads.bam |
| COMMAND | ① load module  module load picard/2.8.2  ② Add or Replace Read Group  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar AddOrReplaceReadGroups\  I=$i.realigned\_reads.bam O=$i.RDGR\_realigned\_reads.bam RGID=4 RGLB=lib1 RGPL=illumine RGPU=unit1 RGSM=20 |
| Loop command | ① creating a shell script file  vim GATK-STEP10.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP10  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP10.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar AddOrReplaceReadGroups\  I=$i.realigned\_reads.bam O=$i.RDGR\_realigned\_reads.bam RGID=4 RGLB=lib1 RGPL=illumine RGPU=unit1 RGSM=20  done |
| NOTES | ①Documentation link  <http://broadinstitute.github.io/picard/command-line-overview.html#AddOrReplaceReadGroups>  ②“O=$i.RDGR\_realigned\_reads.bam” (O - uppercase) |

|  |  |
| --- | --- |
| STEP 11 | Build Read Group BAM Index |
| TOOL | picard/2.8.2 |
| INPUT | $i.RDGR\_realigned\_reads.bam |
| OUTPUT | $i. RDGR\_realigned\_reads.bai |
| COMMAND | ① load module  module load picard/2.8.2  ② Build Read Group BAM Index  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=$i.RDGR\_realigned\_reads.bam |
| Loop  command | ① creating a shell script file  vim GATK-STEP11.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP11  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP11.err  cd /path/to/the/files  module load picard/2.8.2  for i in {start#..end#};do  java -jar /share/apps/picard/2.8.2/picard-2.8.2.jar BuildBamIndex INPUT=$i.RDGR\_realigned\_reads.bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 12A | Call Variants (HaplotypeCaller) |
| TOOL | gatk/3.7-0 |
| INPUT | $i.RDGR\_realigned\_reads.bam; reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.raw.vcf |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T HaplotypeCaller -R reference.fasta\  -I $i.RDGR\_realigned\_reads.bam -o $i.GATK.raw.vcf |
| Loop  Command | ① creating a shell script file  vim GATK-STEP12.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP12  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP12.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {1..107};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T HaplotypeCaller -R reference.fasta\  -I $i.RDGR\_realigned\_reads.bam -o $i.GATK.raw.vcf bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 13 | Extract SNVs & Indels |
| TOOL | gatk/3.7-0 |
| INPUT | $i.GATK.raw.vcf; reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.raw.indels.vcf; $i.GATK.raw.snvs.vcf |
| COMMAND | ① java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T SelectVariants -R reference.fasta\  -V $i.GATK.raw.vcf -selectType SNP -o $i.GATK.raw.snvs.vcf  ② java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T SelectVariants -R reference.fasta\  -V $i.GATK.raw.vcf -selectType INDEL -o $i.GATK.raw.indels.vcf |
| Loop  Command | ① creating a shell script file  vim GATK-STEP13.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP13  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP13.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T SelectVariants -R reference.fasta\  -V $i.GATK.raw.vcf -selectType SNP -o $i.raw.snvs.vcf;  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T SelectVariants -R reference.fasta\  -V $i.GATK.raw.vcf -selectType INDEL -o $i.raw.indels.vcf  done |
| NOTES |  |

<Filtering VCF>

|  |  |
| --- | --- |
| STEP 14-1 | Filter SNVs (or SNPs) (both SNVs and Indels) |
| TOOL | gatk/3.7-0 |
| INPUT | $i.GATK.raw.vcf; reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.flt.vcf |
| COMMAND | ① java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.vcf  ② Mirella sample Example  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.vcf \  --filterExpression ‘DP > 200 || QUAL >20’ --filterName “DP200\_QUAL20” -o $i.GATK.flt.vcf |
| Loop  Command | ① creating a shell script file  vim GATK-STEP14-1.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP14-1  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP14-1.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.vcf  done |
| NOTES | targetting both Indels and SNVs |

|  |  |
| --- | --- |
| STEP 14-2 | Filter SNVs (or SNPs) and Indels |
| TOOL | gatk/3.7-0 |
| INPUT | $i.GATK.raw.indels.vcf; $i.GATK.raw.snvs.vcf; reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.flt.indels.vcf; $i.GATK.flt.snvs.vcf |
| COMMAND | ① java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.indels.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.indels.vcf  ② java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.snvs.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.snvs.vcf |
| Loop  Command | ① creating a shell script file  vim GATK-STEP14-2.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP14-2  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP14-2.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.indels.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.indels.vcf;  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T VariantFiltration -R reference.fasts -V $i.GATK.raw.snvs.vcf \  --filterExpression ‘QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0’ \  --filterName “filter\_name” -o $i.GATK.flt.snvs.vcf  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 15 | Base Quality Score Recalibration (BQSR) #1 |
| TOOL | gatk/3.7-0 |
| INPUT | ① $i.RDGR\_realigned\_reads.bam  ② $i.GATK.flt.indels.vcf  ③ $i.GATK.flt.snvs.vcf  ④ reference.fasta(reference genome file) |
| OUTPUT | $i.GATK\_recal\_data.table |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fasta\  -I $i.RDGR\_realigned\_reads.bam -knownSites $i.flt.snvs.vcf -knownSites $i.flt.indels.vcf -o $i.recal\_data.table |
| Loop  Command | ① creating a shell script file  vim GATK-STEP15.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP15  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error= GATK-STEP15.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fasta\  -I $i.RDGR\_realigned\_reads.bam -knownSites $i.flt.snvs.vcf -knownSites $i.flt.indels.vcf -o $i.recal\_data.table  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 16 | Base Quality Score Recalibration (BQSR) #2 |
| TOOL | gatk/3.7-0 |
| INPUT | ① $i.GATK\_recal\_data.table  ② $i.RDGR\_realigned\_reads.bam  ③ $i.GATK.flt.indels.vcf  ④ $i.GATK.flt.snvs.vcf  ⑤ reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.post\_recal\_data.table |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fasta -I $i.RDGR\_realigned\_reads.bam -knownSites $i.flt.snvs.vcf -knownSites $i.flt.indels.vcf -BQSR $i.GATK\_recal\_data.table\  -o $i.recal\_data.table |
| Loop  Command | ① creating a shell script file  vim GATK-STEP16.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP16  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP16.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fasta\  -I $i.RDGR\_realigned\_reads.bam -knownSites $i.flt.snvs.vcf -knownSites $i.flt.indels.vcf -o $i.recal\_data.table  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 17 | Analyze Covariates |
| TOOL | gatk/3.7-0 |
| INPUT | ① $i.GATK\_recal\_data.table  ② $i.GATK.post\_recal\_data.table  ③ reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.recalibration\_plots.pdf |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T AnalyzeCovariates -R reference.fasta\  -before $i.GATK\_recal\_data.table -after $i.GATK.post\_recal\_data.table -plots $i.GATK.recalibration\_plots.pdf |
| Loop  Command | ① creating a shell script file  vim GATK-STEP17.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP17  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP17.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T AnalyzeCovariates -R reference.fasta\  -before $i.GATK\_recal\_data.table -after $i.GATK.post\_recal\_data.table -plots $i.GATK.recalibration\_plots.pdf  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 18 | Apply BQSR |
| TOOL | gatk/3.7-0 |
| INPUT | ① $i.GATK\_recal\_data.table  ② $i.RDGR\_realigned\_reads.bam  ③ reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.recal\_reads.bam |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T PrintReads -R reference.fasta -I $i.RDGR\_realigned\_reads.bam\  -BQSR $i.GATK.recal\_data.table -o $i.RDGR.recal\_reads.bam |
| Loop  Command | ① creating a shell script file  vim GATK-STEP18.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP17  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP18.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T PrintReads -R reference.fasta -I $i.RDGR\_realigned\_reads.bam\  -BQSR $i.GATK.recal\_data.table -o $i.GATK.recal\_reads.bam  done |
| NOTES |  |

|  |  |
| --- | --- |
| STEP 19 | Call Variants |
| TOOL | gatk/3.7-0 |
| INPUT | ① $i.GATK\_recal\_data.table  ② $i.RDGR\_realigned\_reads.bam  ③ reference.fasta(reference genome file) |
| OUTPUT | $i.GATK.recal\_reads.bam |
| COMMAND | java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T HaplotypeCaller -R reference.fasta(reference genome file) -I |
| Loop  Command | ① creating a shell script file  vim GATK-STEP19.sh  ② writing commands (press ‘I’)  #!/bin/bash  #  ##SBATCH --nodes=1  #SBATCH --nodes=1  #SBATCH --ntasks-per-node=1  #SBATCH --cpus-per-task=2  #SBATCH --time=24:00:00  #SBATCH --mem=20GB  #SBATCH --job-name=GATK-STEP19  #SBATCH --mail-type=END  #SBATCH --mail-user=mk6134@nyu.edu  #SBATCH --output=slurm\_%j.out  #SBATCH --error=GATK-STEP19.err  cd /path/to/the/files  module load gatk/3.7-0  for i in {start#..end#};do  java -jar /share/apps/gatk/3.7-0/GenomeAnalysisTK.jar -T PrintReads -R reference.fasta -I $i.RDGR\_realigned\_reads.bam\  -BQSR $i.GATK.recal\_data.table -o $i.RDGR.recal\_reads.bam  done |
| NOTES |  |

Reference

1. Khalfan M., Variant Calling Pipeline: FastQ to Annotated SNPs in Hours (The Genomics Core Facility @ NYU CGSBSkip to content -https://gencore.bio.nyu.edu/variant-calling-pipeline/)

2. Genome Analysis ToolKit documentations(Broad Institute: https://software.broadinstitute.org/gatk/documentation/)